Towards Crowdsourcing Clickbait Labels for YouTube Videos

Jiani Qu **Anny Marleen Hißbach** **Tim Gollub** **Martin Potthast**

Bauhaus-Universität Weimar and Leipzig University

<first>.<last>@uni-weimar.de and martin.potthast@uni-leipzig.de

Abstract

Clickbait is increasingly used by publishers on social media platforms to spark their users' natural curiosity and to elicit clicks on their content. Every click earns them display advertisement revenue. Social media users who are tricked into clicking may experience a sense of disappointment or agitation, and social media operators have been observing growing amounts of clickbait on their platforms. As largest video-sharing platform on the web, YouTube, too, suffers from clickbait. Many users and YouTubers alike have complained about this development. In this paper, we lay the foundation for crowdsourcing the first YouTube clickbait corpus by (1) augmenting the YouTube 8M dataset with meta data to obtain a large-scale base population of videos, and by (2) studying the task design suitable to manual clickbait identification.

Introduction

Clickbait is a marketing instrument employed by many publishers on social media that entices and manipulates users to click on a certain link by using eye-catching teaser content, exaggerated descriptions, by omitting key information, or even via outright deception—irrespective of whether users are actually interested in the content's topic or not. This usually serves the purpose of maximizing the revenue generated through display advertisement on the content's page. At the same time, it induces a frustrating user experience both on the social media platform as well as on the publisher's page. In recent years, clickbait has been on the rise, threatening to clog up the social media channel just as spam almost did for email, and causing quality content to be buried. News publishers are considered a primary source of clickbait, which is usually in direct violation of journalistic codes of ethics (Potthast et al. 2016). However, also on entertainment platforms, such as YouTube, this problem is increasingly observed due to the considerable amount of advertisement revenue earned by YouTubers (i.e., professional video uploaders) through views on their videos. Many well-known YouTubers have expressed their concerns about this situation: such a market environment is basically a race to the bottom, where people are more or less forced to employ clickbait to avoid their content from being lost among all the catchy titles.

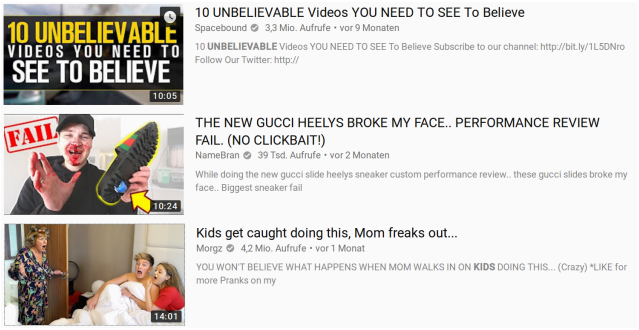

Figure 1: Videos advertised using clickbait teasers.

Clickbait became a focus of computer science research only recently, two of the earliest contributions having been made by Chakraborty et al. (2016) and Potthast et al. (2016). Both rely on a rich set of hand-crafted features to detect clickbait headlines and clickbait tweets, respectively. Wei and Wan (2017) and Biyani, Tsioutsiouliklis, and Blackmer (2016) attempted to categorize clickbait news headlines, where the former distinguishes only ambiguous and misleading ones, and the latter distinguishes eight categories. Agrawal (2016) and Anand, Chakraborty, and Park (2016) employed deep neural networks for clickbait detection, reporting higher precision than the aforementioned studies. Moreover, for the Clickbait Challenge 2017, Potthast et al. (2018) introduced a new large-scale benchmark corpus on which twelve new approaches have been evaluated, which almost exclusively employ deep learning to various degrees of effectiveness.[1] All of the aforementioned studies focused on the news domain.

While news clickbait appears to be rather well-understood, this is not the case for entertainment clickbait. Figure 1 shows examples we encountered in our annotation study. While closely resembling the most extreme forms of clickbait encountered in the news domain, it is still unclear how entertainment clickbait differs from news clickbait. To render clickbait detection on YouTube feasible, however, we have to first understand how it can be reliably identified by a human, so that a valid clickbait dataset can be constructed for training. An important prerequisite for this purpose is an extensive population of YouTube videos containing both clickbait and non-clickbait videos from which to sample. This paper con-

[1] http://www.clickbait-challenge.org/

tributes by laying the foundation for the construction of a YouTube clickbait corpus: (1) the YouTube 8M dataset is augmented with meta information not originally found in it, yielding a base population of YouTube videos, and (2) we conduct a preliminary study on annotating YouTube videos with regard to clickbait as a first step toward crowdsourcing clickbait annotations. In what follows, we detail the corpus construction and report on the results of our preliminary annotation study.

Corpus Construction: Base Population

An important prerequisite for the construction of a valid corpus is to draw a representative sample of documents from the underlying population. YouTube offers little to no help in this regard, since neither its web front end nor its APIs allow for enumerating all videos available, nor tapping into the stream of videos uploaded every day. If not for the recently released YouTube 8M dataset (Abu-El-Haija et al. 2016), which has been constructed by researchers working at YouTube, we would be left with no choice but to crawl YouTube ourselves. Below, we briefly review the construction and original purpose of the YouTube 8M dataset, describe our efforts to augment the dataset with the meta data necessary for clickbait detection (rendering the dataset also useful for tackling other research questions), and give a brief overview of the corpus statistics. Altogether, the resulting corpus compiles (if available) the meta data, comments, thumbnails, and subtitles ("captions") of 6,192,353 videos in a unified format, which we make available to other researchers on request.[2] Table 1 gives an overview of the corpus.

YouTube 8M is a large benchmark for multi-label video classification; it compiles about 500,000 hours of videos annotated with visually recognizable entities. It is available for download in the form of precomputed frame-based feature representations, allowing for the development of classification technology, but comes without any other meta data about the videos. Given that there are currently no other datasets of significant size for YouTube which have been drawn in a sensible manner from the population of all YouTube videos, this corpus provides for the best alternative available to study YouTube-related tasks. As a consequence, its sampling criteria will also apply to our corpus: a video has a length between 120 and 500 seconds, it corresponds to one of 10,000 visual entities of the dataset, and it has more than 1,000 views. The videos we use in our corpus are extracted from the publicly available partitions "Train" and "Validate", which contain 90% of the dataset's videos.

Augmented YouTube 8M. We used the YouTube Data API[3] to crawl a variety of meta data for the videos of YouTube 8M. First point of interest was the "video resource," which comprises data about the video, such as the video's title, description, uploader name, tags, view count, and more. Also included in the meta data is whether comments have been left for the video. If so, we downloaded them as well, including information about their authors, likes, dislikes, and responses. There is no property which specifies a video's language, since this information is not mandatory when uploading a video. Also, the API provides only information about the available captions, but not the captions themselves. Only the uploader of a video is given access to its captions via the API; we extracted them using youtube-dl.[4] For each video, all manually created captions were downloaded, and auto-generated captions in the "default" language and English. The "default" auto-generated caption gives perhaps the only hint at a video's original language. Finally, we downloaded all thumbnails used to advertise a video, which are not available via the API, but only via a canonical URL.

Table 1: Overview of the augmented YouTube 8M corpus. Text lengths are measured in words, counts are numbers of videos; for tags also the sum totals of unique tags is given.

Data item	Count	Distribution			
		min	mean	max	stdev
Videos	6,192,353				
length (s)		120	229.6	500	107.8
views		1,000	60,552.4	>2 billion	803.7
Thumbnails	6,192,353				
Titles	6,192,353				
length		0	7.1	44	0.0
Descriptions	5,917,215				
length		0	45.9	2355	0.0
Tags	5,738,782				
count	72,700,304	0	12.7	146	0.0
Comments	4,732,577	0	41.7	946,863	0.4
length		0	13.9	190,080	0.0
Captions	1,662,459	1	1.4	48	0.0
length		0	472.4	499,650	0.3

Our corpus provides the possibility to recreate the way a video is presented on YouTube (meta data and thumbnail), what the actual content is ((sub)titles and descriptions), and how its viewers reacted (comments), forming the basis to studying the requirements of clickbait annotation.

Annotation Study

Answers to the questions "What are characteristics of clickbait on YouTube?" and "Can it be systematically, yet manually identified?" are important prerequisites to crowdsourcing clickbait annotations. We conducted a controlled annotation study by manually examining a total of 109 YouTube videos. The study was carried out in three stages by two reviewers.

Video review procedure. The review of each video was done according to a rigorous plan: Each video was independently reviewed twice to reduce bias with regard to clickbait classification. The reviewed video properties are title, description, thumbnail, the video itself, and its comments. Also, likes and dislikes were noted and their ratio was calculated. Observations and a clickbait classification were written down in a structured lab notebook after reviewing the video teaser (title, thumbnail, and the first 123 characters of the description) as displayed on YouTube's web front end, and after watching the video (when views, likes, and comments are first visible). Finally, as an exercise to better understand clickbait on YouTube, each video judged as clickbait was given

[2] A public download link is not available for licensing reasons.

[3] https://developers.google.com/youtube/v3/docs/

[4] https://rg3.github.io/youtube-dl/

a non-clickbait title and a short, teaser-fitting description hinting its actual content, and vice versa. In Stages 2 and 3, clickbait was not judged on a binary scale, anymore, but on a scale from 0 (no clickbait) to 5 (strong clickbait).

Stage 1: random sampling. 24 random videos from the corpus were reviewed. The classifications of both reviewers agreed with each other, albeit only 1 video was judged to be clickbait for using excessive exclamation marks in the title and having a poor match between title and content. The graded clickbait score for each video was similar (i.e., always within 1 point of each other). Given the small sample size, we can only say that the prevalence of clickbait on YouTube is low, which led us to adjust our sampling procedure.

Stage 2: stratified sampling. Dividing the view count distribution of the videos into 4 parts (0-2180, 2181-4720, 4721-14936, 14937-max, where the first three intervals' upper bounds are medians), we randomly chose 10 videos per part, presuming a correlation between the number of clicks a video receives and its clickbaitiness. Again, the results showed no differences between reviewers, and the determination of clickbaitiness was mostly the same. Among the 40 videos, only one was considered as debatable clickbait in the group with fewest views (0-2180): the title was misleading and its correspondence with the video's content was weak. However, the title and description seemed to be autogenerated from YouTube itself, which, to the best of our knowledge, happens for direct uploads from video cameras. Finally, we decided to hunt down clickbait.

Stage 3: targeted selection. Obviously, targeted selection causes reviewer bias. Still, since unbiased sampling failed with respect to our goal of determining the difficulty of manual clickbait classification on YouTube, we were left with no alternative. In this stage, we also did not review each video twice, but both reviewers worked in close collaboration: 25 "obvious" non-clickbait videos and 20 "obvious" clickbait videos[5] were handpicked for review based on their teasers. Video selection was done based on reviewers' prior experience and by browsing videos. Both reviewers agreed in their discussions on the non-clickbait videos, but opinions differed on two "clickbait" videos, where their content turned out to be highly relevant to the teaser, though the titles were considered too brief on their own.

Observations. Recall that clickbait pertains to how a video is advertised, whereas the video's actual content (quality) is not in question. This is why our video review took special note of the video teaser. The title and thumbnail are comparable to Twitter teaser messages: characteristics, such as excessive use of capitalization and punctuation, other highlighting and use of certain words like 'this' and 'unthinkable', emotional writing, and deliberate ambiguities or omission of information also appear on most clickbaits we observed. The short video descriptions are a unique addition, and they fall into three categories: (1) blank or same as/paraphrase of title; (2) additional information about content; and (3) encouragement for channel subscription. The thumbnail images very often contain extra textual information, especially in the case of user-defined custom thumbnails. Many of the clickbait videos employ thumbnails with brightly colored extra text, extreme facial expressions or emojis, and unnatural, surprising, or suggestive pictures. The aforementioned characteristics can be utilized for detection.

However, the teaser information alone does not suffice to detect clickbait videos. Of the 87 videos labeled as non-clickbait, 9 have titles that meet the aforementioned criteria (excessive capitalization or punctuation) with 5 medium-to-severe cases. In spite of that, the teasers have a strong relation to the video's content and hence were not considered as clickbait. This may result in a high false-positive rate when crowdsourcing annotations based on teasers alone.

Providing the video and user reactions like comments as well helps to resolve ambiguous cases at the cost of a significantly higher workload. Watching, say, the first minute of a video suffices to resolve all of the aforementioned false positive cases. Reviewing comments, however, was not so useful: the types of comments we identified include discussions about the video's topic, questions or remarks directed to the uploader, feedback on the video itself, and random thoughts. Reviewing all comments on a video is feasible only for less than, say, 20 comments, but selecting comments for review, e.g., based on a sentiment analysis, will prove difficult: content-related expressions of anger, e.g., against an evil antagonist from a short film, are intermingled with meta discussions about the video (quality). A better approach may be selecting keywords or phrases (e.g., when a commenter explicitly says 'clickbait') from the comments. Other video properties were unhelpful, which may be due to sample size.

Altogether, we will proceed with crowdsourcing based on a two-stage process, where in the first stage, video teasers will be reviewed, which can be done at a glance and therefore at reasonable costs, and in the second stage, the clickbait identified will be reviewed more in-depth to ensure a low false positive rate by asking the workers to watch the first 30 to 60 seconds of a video plus some extracts from the comments.

As an aside, we discovered a kind of clickbait special to video platforms: "Staybait" refers to videos that foretell or promise something exciting in the video to build up suspense, but fail to deliver, or barely mention it later. The two examples of staybait we found were both from vlogs (personal experience documentation in video form).

Conclusion

Automatic clickbait detection on YouTube is still far out of reach, since a reliable training corpus needs to be constructed. Doing so requires access to the base population of videos, which we represent using the YouTube 8M corpus, video meta data, which we crawl for that corpus, and a reliable annotation procedure, which we developed as part of this study. As we do not expect to annotate the entire YouTube 8M corpus, what is still missing, is a reliable way of sampling from YouTube 8M so that the sampling strategy is unbiased, yet, yields a non-trivial amount of clickbait videos as part of a reasonably-sized sample of, say, 100,000 videos. For Twitter, Potthast et al. (2018) solved this problem by looking only at the top news publishers. Given the high diversity of successful YouTube channels, however, it is questionable whether the same strategy can be applied here as well.

[5] 5 more videos were deleted on YouTube and had to be omitted.

References

Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.; and Vijayanarasimhan, S. 2016. Youtube-8m: A large-scale video classification benchmark. *CoRR* abs/1609.08675.

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

Anand, A.; Chakraborty, T.; and Park, N. 2016. We used neural networks to detect clickbaits: You won't believe what happened next! *CoRR* abs/1612.01340.

Biyani, P.; Tsioutsiouliklis, K.; and Blackmer, J. 2016. " 8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. In *AAAI Conference on Artificial Intelligence*.

Chakraborty, A.; Paranjape, B.; Kakarla, S.; and Ganguly, N. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. *CoRR* abs/1610.09786.

Potthast, M.; Köpsel, S.; Stein, B.; and Hagen, M. 2016. Clickbait Detection. In *Proceedings of the 38th European Conference on Information Retrieval (ECIR 16)*, 810–817. Springer.

Potthast, M.; Gollub, T.; Komlossy, K.; Schuster, S.; Wiegmann, M.; Garces, E.; Hagen, M.; and Stein, B. 2018. Crowdsourcing a Large Corpus of Clickbait on Twitter. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 18)*.

Wei, W., and Wan, X. 2017. Learning to identify ambiguous and misleading news headlines. *CoRR* abs/1705.06031.