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.

A tweet is clickbait if (1) its tweet title or information required to understand what the content of the article is, and (2) the tweet exaggerates the article to create misleading expectations for the reader. 

We provided definitions and examples to the annotators.

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.

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

In total, we inspected more than 250,000 annotations.

Most publishers follow the general distribution. Buzzfeed and Breitbart are notable exceptions.

TV networks are particularly non-clickbait.

The corpus has been used for the clickbait challenge 2017.

The corpus was used for the clickbait challenge 2017.