

Clickbait

Detection

supported by  | 



What's clickbait?

Merriam Webster: *something (such as a headline) designed to make readers want to click a hyperlink especially when the link leads to content of dubious value or interest.*

Wikipedia: *Clickbait is a pejorative term describing web content that is aimed at generating online advertising revenue, especially at the expense of quality or accuracy, relying on sensationalist headlines or eye-catching thumbnail pictures to attract click-throughs and to encourage forwarding of the material over online social networks.*

Why care?

- Clickbait messes with your mind by exploiting cognitive biases to achieve its purpose
- Clickbait hardly ever delivers on its promises (esp. with respect to emotional reactions)
- Clickbait violates journalistic codes of ethics (i.e., news should be informative up front)
- Clickbait is on the rise in social networks (e.g., it may become a new form of spam)

Our contributions

- First proof-of-concept on clickbait detection
 - Constructed large corpus of clickbait on Twitter
 - Extracted features targeting clickbait
 - Trained a linear classifier to detect clickbait
 - Evaluated classification performance (random forest on top 1000 features as per χ^2)
- Precision: 0.76**
Recall: 0.76
ROC-AUC: 0.79

Webis Twitter Clickbait Corpus 2016:

Type	Publisher (top 20 as per retweets)	Retweets in 2014 (millions)	Twitter Clickbait Corpus (week 24, 2015)						
			Tweets w/ links	Random sample	Clickbait Annotation			Majority	
			A	B	C				
Newspapers	Washington Post	6.4	691	150	57%	21%	39%	41%	
	Forbes	11.5	721	150	51%	29%	35%	38%	
	Daily Mail	6.9	516	150	40%	28%	12%	22%	
	New York Times	23.8	875	150	39%	13%	25%	21%	
	The Telegraph	6.4	699	150	31%	13%	23%	21%	
	Wall Street Journal	6.5	747	150	42%	12%	20%	19%	
	The Guardian	14.0	744	150	26%	11%	19%	15%	
Internet media	Business Insider	6.5	779	150	65%	35%	49%	51%	
	Huffington Post	11.6	770	150	53%	34%	45%	46%	
	The Independent	5.8	530	150	55%	37%	42%	45%	
	BuzzFeed	10.0	695	150	68%	40%	32%	42%	
	Mashable	20.6	803	150	53%	22%	29%	33%	
	Yahoo!	8.2	195	150	47%	10%	25%	23%	
	Bleacher Report	10.2	196	150	35%	5%	11%	9%	
TV networks	ESPN	6.9	142	142	57%	11%	24%	24%	
	BBC News	39.6	694	150	41%	12%	15%	17%	
	CNN	15.0	345	150	41%	13%	18%	17%	
	NBC News	9.7	408	150	20%	3%	12%	14%	
	ABC News	17.6	279	150	16%	9%	9%	9%	
	Fox News	10.2	378	150	19%	7%	8%	8%	
			Σ	247.4	11,207	2,992	43%	19%	25%

What's next?

- Project funded by Google Digital News Initiative
- Crowdsource a more diverse corpus
- Understanding and analyzing clickbait better
- Consider other platforms (e.g., Facebook, YouTube, publisher front pages)
- Consider personalized detection

Contact

Martin Potthast, Sebastian Köpse, Benno Stein, and Matthias Hagen
 Bauhaus-Universität Weimar
 Germany
 martin.potthast@uni-weimar.de
 www.webis.de

