Crowdsourcing a Corpus for Clickbait Spoiling

July 4th, 2019 • Jana Puschmann

1. Referee: Prof. Dr. Benno Stein
2. Referee: PD Dr. Andreas Jakoby
Clickbait

The term “clickbait” refers to social media messages that are foremost designed to entice their readers into clicking an accompanying link to the posters’ website, at the expense of informativeness and objectiveness.

- Potthast et. al. [2018]
Clickbait

BuzzFeed 🌐 @BuzzFeed · Jun 24
These "Spider-Man: Far From Home" Posters Have A Huge Mistake, And Samuel L. Jackson Had The Best Response

Samuel L. Jackson Had The Perfect Response To This "Spider-Man: ...
"Last time I trusted someone, I lost an eye."
buzzfeed.com

https://twitter.com/BuzzFeed/status/1143221248257748993
Clickbait

Die Ehefrau hilft, den Fall zu lösen – und rettet ihrem Gatten damit das Leben.

K.o. durch Lachs: Ein Mann schwebt in Lebensgefahr – hat er eine Fischallergie?
Clickbait

HuffPost 🚨 @HuffPost · 4h
And "Game of Thrones" star Maisie Williams' next role is...

Maisie Williams' Next Role Sounds As Badass As Arya Stark In 'Gam...
But new show "Two Weeks to Live" is a world away from Westeros.

https://twitter.com/HuffPost/status/1143895724645593089
Clickbait

The Independent 📰 @Independent · 2h
Netflix to lose its most-streamed show in 2020

Netflix is losing its most-streamed show in 2020
Netflix is about to lose its most successful licensed television series. Bosses are no doubt floundering following the news that all nine seasons... independent.co.uk

https://twitter.com/Independent/status/1143793015523123201
Excel + Tableau: A Beautiful Partnership

Amazon Prime Day is coming!
Sign up for the BuzzFeed Shopping newsletter to get all the latest deals delivered to your inbox.
Combat Clickbait

The Independent 🦅 @Independent • 2h
Netflix to lose its most-streamed show in 2020

Netflix is losing its most-streamed show in 2020
Netflix is about to lose its most successful licensed television series. Bosses are no doubt floundering following the news that all nine seasons...

independent.co.uk
Combat Clickbait: Warning

The Independent @Independent · 2h
⚠️ This post contains media that was detected as clickbait. Read more.

Netflix to lose its most-streamed show in 2020

Netflix is losing its most-streamed show in 2020
Netflix is about to lose its most successful licensed television series. Bosses are no doubt floundering following the news that all nine seasons...

independent.co.uk

5 5 8
Combat Clickbait: Block Media

The Independent 🎈 @Independent · 2h
Netflix to lose its most-streamed show in 2020

This post contains media that was detected as clickbait. Why is this media considered clickbait?

View

5️⃣ 5️⃣ 8️⃣
Combat Clickbait: Manual Spoiling

Saved You A Click @SavedYouAClick · Jan 29
Store it upside down to evenly distribute the oil.

WBAL-TV 11 Baltimore 🔄 @wbaltv11
You’ve probably been storing peanut butter wrong all along bit.ly/2DuRDY7

https://twitter.com/SavedYouAClick/status/1090226980740628480
Combat Clickbait: Automated Spoiling

Netflix to lose its most-streamed show in 2020

Netflix is about to lose its most successful licensed television series. Bosses are no doubt floundering following the news that all nine seasons...
Corpus Construction

Crowdsourcing a Corpus for Clickbait Spoiling
Crowdsourcing Process on Amazon MTurk
Base Corpus: Webis-Clickbait-17

• 38,517 annotated Twitter tweets and their related articles
• Each tweet was rated by 5 annotators on a 4-point scale
• All 1,845 articles with a „truthMean“ higher than 0.8 were adopted
Base Corpus: Webis-Clickbait-18

• All 5,787 clickbait-spoiler-pairs from Facebook, reddit and Twitter and their related articles were adopted
• Defined as clickbait only by the person who posted the spoiler

https://twitter.com/SavedYouAClick/status/1096773022449582080
Base Corpus: Pre-Annotation

• A base corpus of 7,632 clickbaits and their articles was constructed

• 433 entries could not be spoiled

• 7,199 clickbait entries were annotated in the crowdsourcing process
Crowdsourcing Task: Instructions

Clickbait Spoiling

Find a fitting spoiler sentence for the given clickbait headline.

In this task you are asked to copy sentences from the given article and paste them into the answer text area below.

Above each article, a "clickbait" headline is stated. Clickbait headlines aim to create a "curiosity gap" by providing just enough information to make readers curious, but not enough to satisfy their curiosity without clicking through to the linked content.

From the article, copy those sentences which provide enough information to close the curiosity gap created by the clickbait headline, i.e., which spoil the clickbait headline.

Please consider the following guidelines

- If possible choose only one sentence; try to be precise.
- If you find more than one spoiler candidate, please choose the one that you think fits best.
- Copy whole sentences, even if the answer can be summarized in a word.
- To spoil a listicle, please copy the top three entries.
  - If the list is ordered, please spoil the three highest ranked entries.
- If you did not find a spoiler sentence, please tick the checkbox below the submission box.
Crowdsourcing Task

Clickbait

One Superstar Left Julianne Hough FUMING With Rage; Now She Refuses To Work With Her! – God Today

Article

One Superstar Left Julianne Hough FUMING With Rage; Now She Refuses To Work With Her!

Julianne Hough made a huge announcement recently. She was chosen to be a host for the 2016 Miss USA competition. Julianne has shot her way up to fame with her amazing eyes, beautiful voice, and perfect dance moves. Julianne is considered by many to be a breeze to work with, but one super star got on her nerves recently. Before I get into that Julianne released a statement about the Miss USA pageant.

"I feel fortunate to be a part of encouraging and uplifting them to make a positive difference in the world," she said. "I'm also excited to contribute what I've learned in over a decade in front of the camera and behind the scenes as a performer and choreographer on live television to the Miss USA broadcast." She cannot way to be a host at this amazing show.

Julianne is an amazing actress, singer, and dancer. Yet, one super star begs to differ. Selena Gomez, formerly on the Disney channel, called Julianne a snob during an interview. She said that Julianne bribed her way up the ladder. Julianne could not believe this unprovoked act! She had never thought that she did Selena any harm, but now Julianne is holding a grudge. She refuses to work with Selena, and I am sure it will go on forever. Are you a fan of Julianne?

Crowdsourcing Task

Clickbait

One Superstar Left Julianne Hough FUMING With Rage; Now She Refuses To Work With Her! – God Today

Article

One Superstar Left Julianne Hough FUMING With Rage; Now She Refuses To Work With Her!

Julianne Hough made a huge announcement recently. She was chosen to be a host for the 2016 Miss USA competition. Julianne has shot her way up to fame with her amazing eyes, beautiful voice, and perfect dance moves. Julianne is considered by many to be a breeze to work with, but one super star got on her nerves recently. Before I get into that Julianne released a statement about the Miss USA pageant.

"I feel fortunate to be a part of encouraging and uplifting them to make a positive difference in the world," she said. "I'm also excited to contribute what I've learned in over a decade in front of the camera and behind the scenes as a performer and choreographer on live television to the Miss USA broadcast." She cannot way to be a host at this amazing show.

Julianne is an amazing actress, singer, and dancer. Yet, one super star begs to differ. Selena Gomez, formerly on the Disney channel, called Julianne a snob during an interview. She said that Julianne bribed her way up the ladder. Julianne could not believe this unprovoked act! She had never thought that she did Selena any harm, but now Julianne is holding a grudge. She refuses to work with Selena, and I am sure it will go on forever. Are you a fan of Julianne?

Crowdsourcing Task: Spoiler Annotation

**Answer**

Selena Gomez, formerly on the Disney channel, called Julianne a snob during an interview.

- I did not find a fitting spoiler sentence

Submit
Crowdsourcing Task: Spoiler Annotation

Answer

Spoiler sentence(s) ...

☐ I did not find a fitting spoiler sentence

If you did not find a fitting sentence, please answer this question.

It was not possible to find a fitting spoiler sentence because ...

☐ the content is too complex to summarize in only one sentence.

☐ the content does not relate to the headline.

☐ I do not consider this clickbait.

☐ Other: Enter your reason here

Submit
## Crowdsourcing Task: Review

<table>
<thead>
<tr>
<th>Worker assignment review</th>
<th>Approved</th>
<th>Rejected</th>
<th>(\sum)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spoiler found</td>
<td>3477</td>
<td>2387</td>
<td>5864</td>
</tr>
<tr>
<td>No spoiler found</td>
<td>821</td>
<td>1204</td>
<td>2025</td>
</tr>
<tr>
<td>Empty submissions</td>
<td>0</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>(\sum)</td>
<td>4298</td>
<td>3613</td>
<td>7911</td>
</tr>
</tbody>
</table>
Webis Clickbait Corpus 2019

• The crowdsourcing process led to the Webis-Clickbait-19 corpus, which consists of 3,042 articles.
Webis Clickbait Corpus 2019

{
    "uuid": "8ceed760-2fdb-49ec-b8a9-cd08ddd83426",
    "originalCorpus": "webis-clickbait-17",
    "postId": "810551255659585536",
    "postPlatform": "Twitter",
    "postMedia": ["media/photo_810551253520551936.jpg"],
    "postText": ["Washington Capitals' \u2018Social Night ..."],
    "targetTitle": "Washington Capitals' \u2018Social Night ...",
    "targetDescription": "The Washington Capitals messed up ...",
    "targetKeywords": ",",
    "targetMedia": null,
    "targetParagraphs": ["A very smart woman once advised ..."],
    "targetUrl": null,
    "humanSpoiler": null,
    "spoilerSentences": ["An interaction with one fan got ..."],
    "spoilerSentencePositions": [[2, 0, 150], [2, 151, 240]]
}
Clickbait Spoiling Experiments

Corpus Analysis
Clickbait Spoiling

<table>
<thead>
<tr>
<th></th>
<th>Random Ranking</th>
<th>Naive Ranking</th>
<th>Cosine Similarity</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision@1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Rank in %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Random Ranking

• Ranks the sentences of an article in a random order
Random Ranking

<table>
<thead>
<tr>
<th></th>
<th>Random Ranking</th>
<th>Naive Ranking</th>
<th>Cosine Similarity</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision@1</td>
<td>8.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@2</td>
<td>14.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@3</td>
<td>20.97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@4</td>
<td>27.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@5</td>
<td>32.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@6</td>
<td>38.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@7</td>
<td>44.28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@8</td>
<td>49.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@9</td>
<td>53.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@10</td>
<td>57.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Rank</td>
<td>12.99</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Precision@n in %)
Naive Ranking

• **Assumption:** Sentences in the beginning of an article are more likely to spoil a clickbait than the following sentences
## Naive Ranking

<table>
<thead>
<tr>
<th></th>
<th>Random Ranking</th>
<th>Naive Ranking</th>
<th>Cosine Similarity</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision@1</td>
<td>8.02</td>
<td>6.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@2</td>
<td>14.40</td>
<td>22.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@3</td>
<td>20.97</td>
<td>35.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@4</td>
<td>27.32</td>
<td>45.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@5</td>
<td>32.94</td>
<td>53.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@6</td>
<td>38.40</td>
<td>60.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@7</td>
<td>44.28</td>
<td>67.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@8</td>
<td>49.01</td>
<td>72.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@9</td>
<td>53.32</td>
<td>76.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@10</td>
<td>57.82</td>
<td>80.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Rank</td>
<td>12.99</td>
<td>7.73</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Precision@n in %]
Cosine Similarity

• **Assumption**: Sentences that similar to the clickbait are more likely to spoil it, than sentences that are not
## Cosine Similarity

<table>
<thead>
<tr>
<th></th>
<th>Random Ranking</th>
<th>Naive Ranking</th>
<th>Cosine Similarity</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision@1</td>
<td>8.02</td>
<td>6.28</td>
<td>12.89</td>
<td></td>
</tr>
<tr>
<td>Precision@2</td>
<td>14.40</td>
<td>22.22</td>
<td>27.94</td>
<td></td>
</tr>
<tr>
<td>Precision@3</td>
<td>20.97</td>
<td>35.04</td>
<td>40.04</td>
<td></td>
</tr>
<tr>
<td>Precision@4</td>
<td>27.32</td>
<td>45.30</td>
<td>49.28</td>
<td></td>
</tr>
<tr>
<td>Precision@5</td>
<td>32.94</td>
<td>53.52</td>
<td>58.71</td>
<td></td>
</tr>
<tr>
<td>Precision@6</td>
<td>38.40</td>
<td>60.82</td>
<td>64.50</td>
<td></td>
</tr>
<tr>
<td>Precision@7</td>
<td>44.28</td>
<td>67.19</td>
<td>70.45</td>
<td></td>
</tr>
<tr>
<td>Precision@8</td>
<td>49.01</td>
<td>72.42</td>
<td>75.12</td>
<td></td>
</tr>
<tr>
<td>Precision@9</td>
<td>53.32</td>
<td>76.92</td>
<td>78.96</td>
<td></td>
</tr>
<tr>
<td>Precision@10</td>
<td>57.82</td>
<td>80.60</td>
<td>81.95</td>
<td></td>
</tr>
<tr>
<td>Average Rank</td>
<td>12.99</td>
<td>7.73</td>
<td>7.06</td>
<td></td>
</tr>
</tbody>
</table>

[Precision@n in %]
Logistic Regression Model

• **Assumption:** Creating a classifier based on both features from the previous approaches will increase the performance
Logistic Regression Model

- Only approximately 5% of all sentences are part of a spoiler
Logistic Regression

<table>
<thead>
<tr>
<th>Precision@n</th>
<th>Random Ranking</th>
<th>Naive Ranking</th>
<th>Cosine Similarity</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision@1</td>
<td>8.02</td>
<td>6.28</td>
<td>12.89</td>
<td>13.91</td>
</tr>
<tr>
<td>Precision@2</td>
<td>14.40</td>
<td>22.22</td>
<td>27.94</td>
<td>32.58</td>
</tr>
<tr>
<td>Precision@3</td>
<td>20.97</td>
<td>35.04</td>
<td>40.04</td>
<td>46.25</td>
</tr>
<tr>
<td>Precision@4</td>
<td>27.32</td>
<td>45.30</td>
<td>49.28</td>
<td>55.06</td>
</tr>
<tr>
<td>Precision@5</td>
<td>32.94</td>
<td>53.52</td>
<td>58.71</td>
<td>62.46</td>
</tr>
<tr>
<td>Precision@6</td>
<td>38.40</td>
<td>60.82</td>
<td>64.50</td>
<td>68.61</td>
</tr>
<tr>
<td>Precision@7</td>
<td>44.28</td>
<td>67.19</td>
<td>70.45</td>
<td>73.93</td>
</tr>
<tr>
<td>Precision@8</td>
<td>49.01</td>
<td>72.42</td>
<td>75.12</td>
<td>78.11</td>
</tr>
<tr>
<td>Precision@9</td>
<td>53.32</td>
<td>76.92</td>
<td>78.96</td>
<td>81.79</td>
</tr>
<tr>
<td>Precision@10</td>
<td>57.82</td>
<td>80.60</td>
<td>81.95</td>
<td>84.29</td>
</tr>
<tr>
<td>Average Rank</td>
<td>12.99</td>
<td>7.73</td>
<td>7.06</td>
<td>6.71</td>
</tr>
</tbody>
</table>

[Precision@n in %]
Future Work and Outlook

Possible approaches to continue this work
Future Work in Clickbait Spoiling

• Formulation of further features

• Incorporation of the findings from Bagrat Ter-Akopyans bachelor’s thesis

• OR

• Use Open-Domain Question Answering to spoil clickbait
About 14,900,000 results (0.65 seconds)

Alan Turing / Age at death

41 years
1912–1954
Relation between Clickbait and Questions

• What Happened to Frank Ocean’s Staircase? (Direct)

• How Angelina Jolie Told Brad Pitt She Wanted a Divorce (Indirect)

• How did Angelina Julie tell Brad Pitt She Wanted a Divorce?

• This is the worst Arab state for women

• Which is the worst Arab state for women?
Open-Domain Question Answering

Jurafsky et. al. [2018]
Open-Domain Question Answering

Jurafsky et. al. [2018]
Thank you for listening

Questions?
References

