

Bachelor's thesis defence

# 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

2 11 62

# Clickbait

BuzzFeed  @E  stern 4 Std. · 

These "Spider-Man" Fans Are Going To Kill You

Samuel L. Jackson

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



STERN.DE

Samuel L. Jackson

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

buzzfeed.com

2 Kommentare

2 Gefällt mir

Kommentieren

Teilen

# Clickbait

BuzzFeed  @BuzzFeed 4 Std. · 7  
These "Spider-Man: Far From Home" photos show Samuel L. Jackson's best moment.  
  
STERN.DE  
Samuel L. Jackson Had The K.o. durch Lack einer Fischallergie  
"Last time I trusted someone like that."  
buzzfeed.com

HuffPost  @HuffPost · 4h  
And "Game of Thrones" star Maisie Williams' next role is...  
  
**Maisie Williams' Next Role Sounds As Badass As Arya Stark In 'Game of Thrones'**  
But new show "Two Weeks to Live" is a world away from Westeros.  
huffpost.com

2 11 62 Gefällt mir 5 8 59

# Clickbait

BuzzFeed  @BuzzFeed 4 Std. ago  
These "Spider-Man: Far From Home" Chapters Have A Hidden Secret: Die Ehefrau hilft, den Fall zu lösen – und Samuel L. Jackson Had The Best Response.

HuffPost  And "Game of Thrones" Maisie Williams' New Show Is About to Kill Netflix

STERN.DE  
Samuel L. Jackson Had The Perfect Response To His Eye Injury: "Last time I trusted someone with a lachan eye..."

buzzfeed.com

2 11 Gefällt mir 5 Kommentiere

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

NETFLIX NETFLIX NETFLIX NETFLIX NETFLIX NETFLIX NETFLIX NETFLIX NETFLIX

**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

The screenshot shows the BuzzFeed website homepage. At the top, there's a navigation bar with links for News, Quizzes, TV & Movies, Shopping, Trending, Videos, and Newsletters. To the right of the navigation is a search icon and three decorative circular icons labeled 'LOL', 'win', and '↗'. The main banner features a red background with white text: 'Want To Learn Something New About Your Favorite Restaurant?' followed by 'Excel + Tableau: A Beautiful Partnership' and a 'GET THE WHITEPAPER' button. Below the banner, there's a 'FOOD' category section with a 'QUIZ' button and a small video thumbnail showing two people. A promotional graphic for 'Amazon Prime Day 2019' is visible on the left, and a newsletter sign-up form is on the right.

Want To Learn Something New About Your Favorite Restaurant?

**BuzzFeed** News Quizzes TV & Movies Shopping Trending Videos Newsletters

LOL win ↗

++ + a b | e a u

## Excel + Tableau: A Beautiful Partnership

GET THE WHITEPAPER

FOOD

QUIZ

ANZEIGE

Amazon Prime Day is coming!

Sign up for the BuzzFeed Shopping newsletter to get all the latest deals delivered to your inbox.

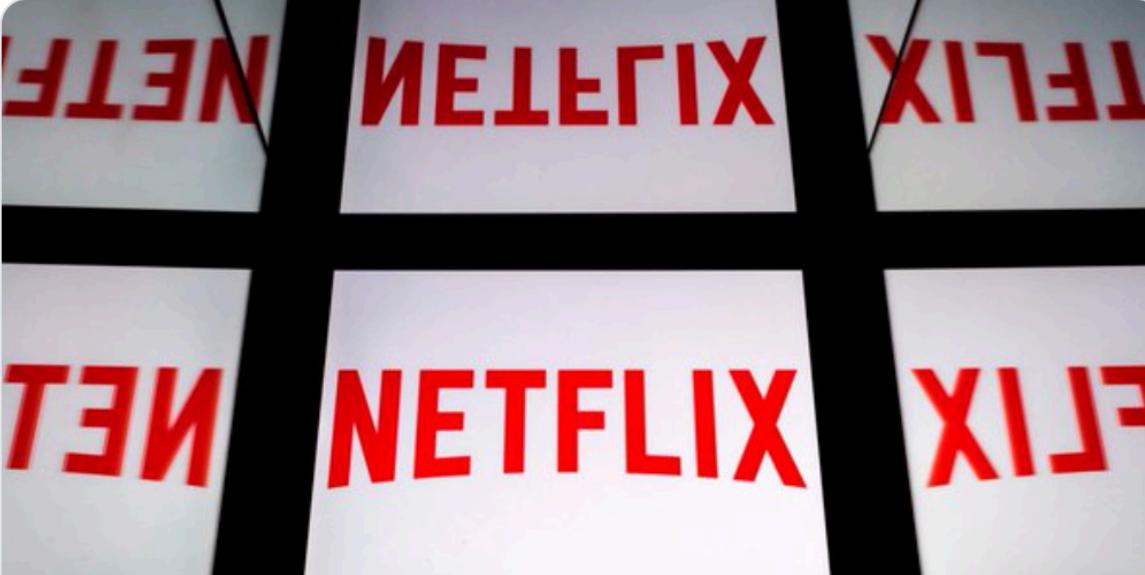
Your email address

Sign up

PRIME DAY 2019

# 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](https://independent.co.uk)

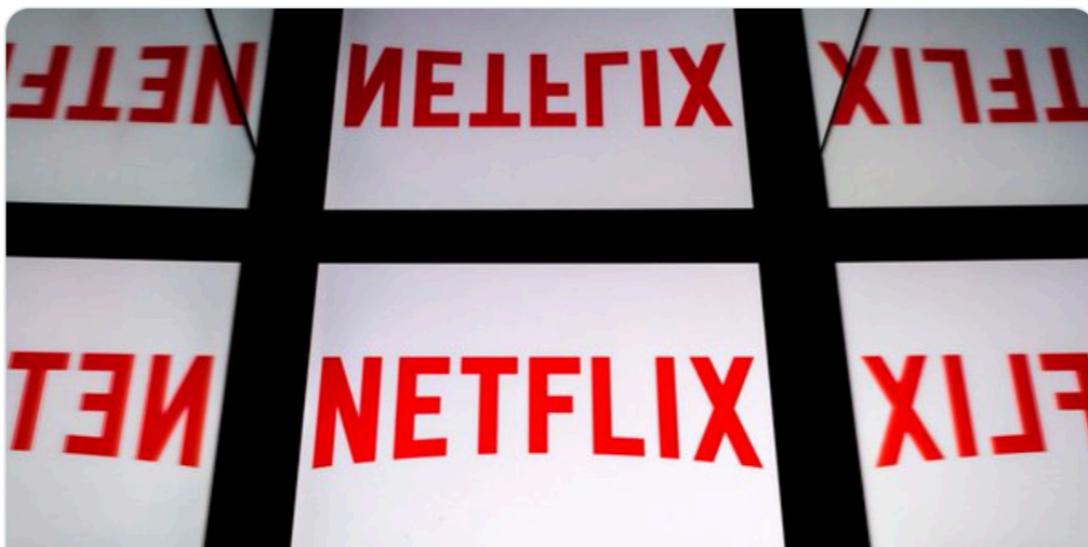
5 5 8

# 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](https://independent.co.uk)

5 5 8 9

# 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](#)

5    5    8    

# 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://bit.ly/2DuRDY7)



24



193



1.2K



# Combat Clickbait: Automated Spoiling

The Independent  @Independent · 2h

 This post contains media that was detected as clickbait. [Read more.](#)

Spoiler: "Netflix to lose hit Steve Carell sitcom The Office in 2020"

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...

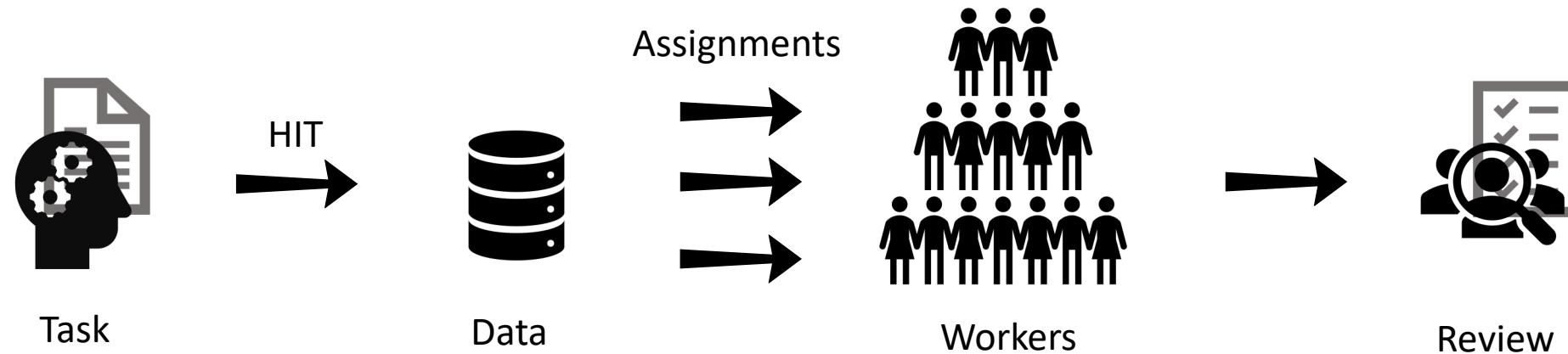
[independent.co.uk](https://independent.co.uk)

5 5 8

# 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

Some people are such food snobs [link](#)

not click baiting (0.0)     slightly click baiting (0.33)     considerably click baiting (0.66)     heavily click baiting (1.0)

crowd sourcing task design

```
{"id": "608999590243741697",
"truthJudgments": [0.33, 1.0, 1.0, 0.66, 0.33],
"truthMean" : 0.6666667,
"truthMedian": 0.6666667,
"truthMode" : 1.0,
"truthClass" : "clickbait"}
```

truth.jsonl

# 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

A screenshot of a Twitter feed. At the top, a tweet from the account **Saved You A Click** (@SavedYouAClick) dated Feb 16 is shown. The tweet text is: "order of the phoenix (or deathly hallows part 2 he couldn't decide)". To the left of the tweet is a large hand cursor icon pointing towards it. Below this is a reply from the account **Mental Floss** (@mental\_floss). The reply text is: "Daniel Radcliffe's Favorite Harry Potter Movie Isn't a Fan Favorite — bit.ly/2tj3bYB". The reply includes a small profile picture of Daniel Radcliffe. At the bottom of the screenshot are standard Twitter interaction icons: a speech bubble with '13' (comments), a retweet icon with '42', a heart icon with '655' (likes), and an envelope icon.

# 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 wait 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?

Source: <http://godtoday.com/one-superstar-left-julianne-hough-fuming-with-rage-now-she-refuses-to-work-with-her/>

# 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 wait 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?

Source: <http://godtoday.com/one-superstar-left-julianne-hough-fuming-with-rage-now-she-refuses-to-work-with-her/>

# 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

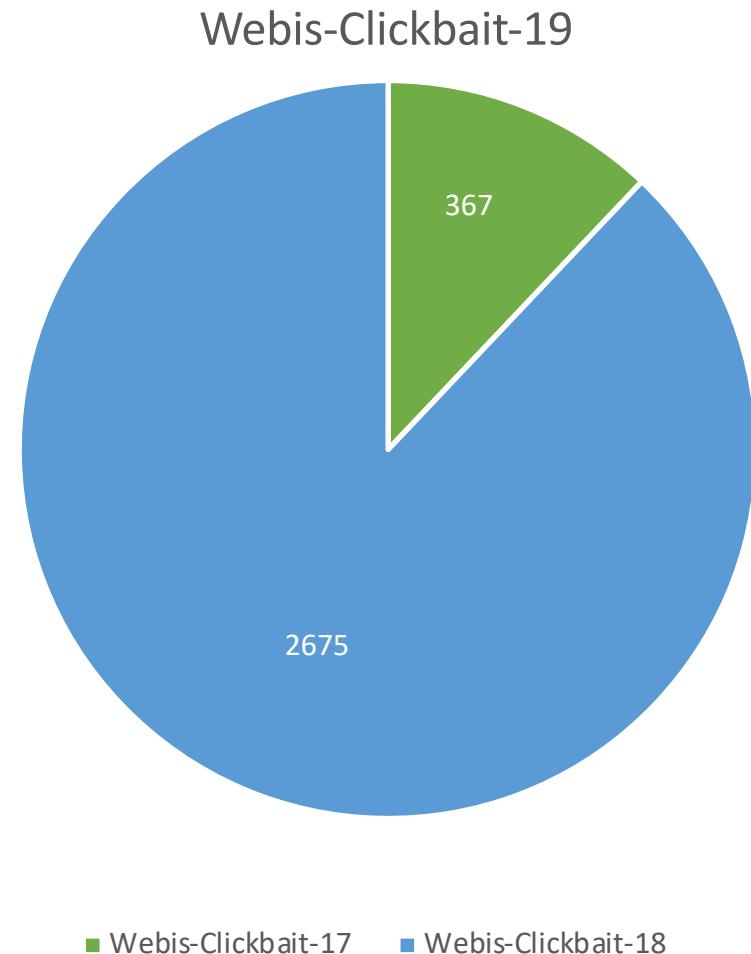
# Crowdsourcing Task: Review

---

Worker assignment review			
	Approved	Rejected	Σ
Spoiler found	3477	2387	5864
No spoiler found	821	1204	2025
Empty submissions	0	22	22
Σ	4298	3613	7911

# 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

	Random Ranking	Naive Ranking	Cosine Similarity	Logistic Regression
Precision@1				
Precision@2				
Precision@3				
Precision@4				
Precision@5				
Precision@6				
Precision@7				
Precision@8				
Precision@9				
Precision@10				
Average Rank				

# Random Ranking

- Ranks the sentences of an article in a random order

# Random Ranking

	Random Ranking	Naive Ranking	Cosine Similarity	Logistic Regression
Precision@1	8.02			
Precision@2	14.40			
Precision@3	20.97			
Precision@4	27.32			
Precision@5	32.94			
Precision@6	38.40			
Precision@7	44.28			
Precision@8	49.01			
Precision@9	53.32			
Precision@10	57.82			
Average Rank	12.99			

# Naive Ranking

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

# Naive Ranking

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

# Cosine Similarity

- **Assumption:** Sentences that similar to the clickbait are more likely to spoil it, than sentences that are not

# Cosine Similarity

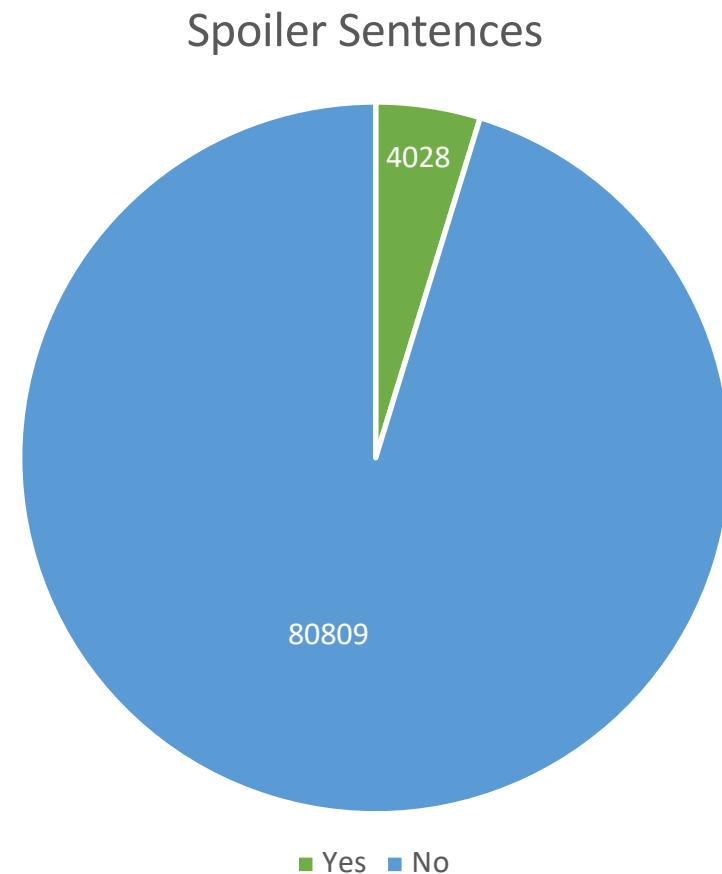
	Random Ranking	Naive Ranking	Cosine Similarity	Logistic Regression
Precision@1	8.02	6.28	12.89	
Precision@2	14.40	22.22	27.94	
Precision@3	20.97	35.04	40.04	
Precision@4	27.32	45.30	49.28	
Precision@5	32.94	53.52	58.71	
Precision@6	38.40	60.82	64.50	
Precision@7	44.28	67.19	70.45	
Precision@8	49.01	72.42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

# Logistic Regression Model

- **Assumption:** Creating a classifier based on both features from the previous approaches will increase the performance

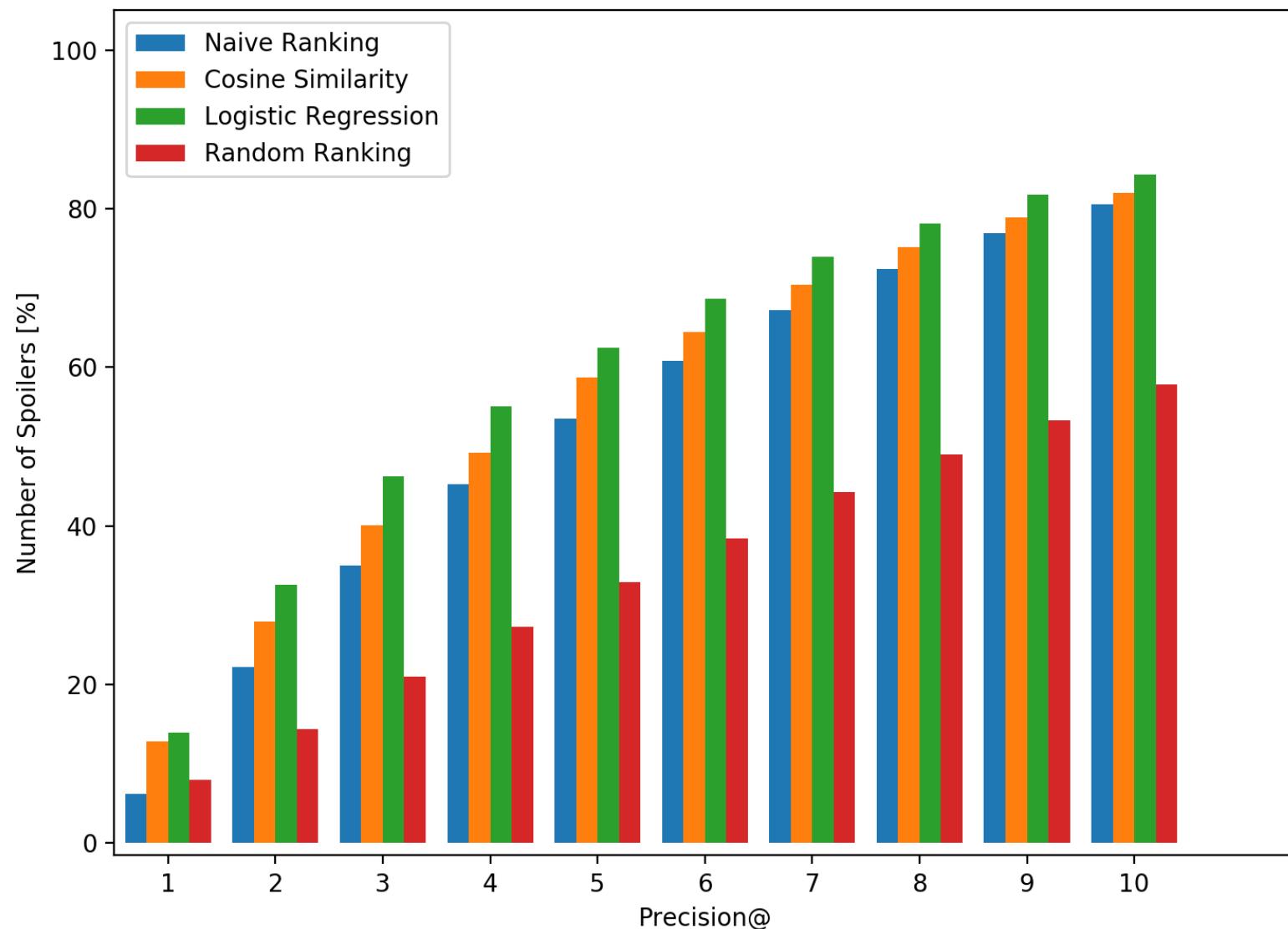
# Logistic Regression Model

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



# Logistic Regression

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



# 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



How old did Turing get?



All

Images

News

Videos

Shopping

More

Settings

Tools

About 14.900.000 results (0,65 seconds)

Alan Turing / Age at death

41 years

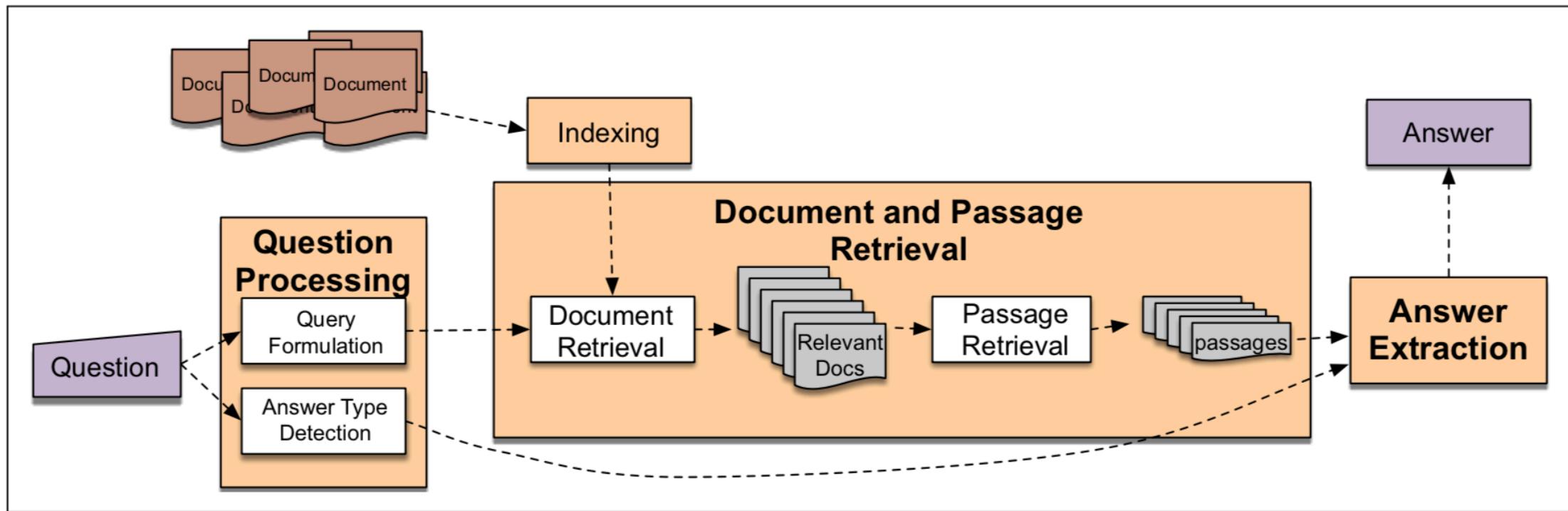
1912–1954



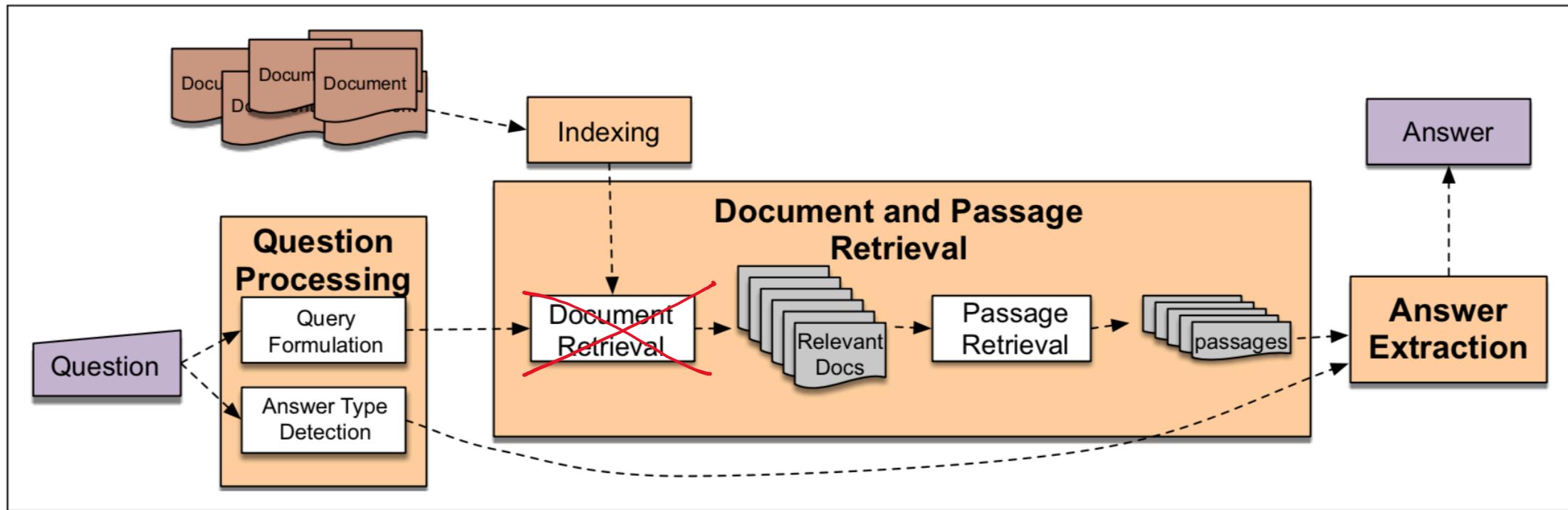
# Relation between Clickbait and Questions

- 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



# Open-Domain Question Answering



# Thank you for listening

Questions?

# References

- Martin Potthast, Tim Gollub, Matthias Hagen, and Benno Stein. The clickbait challenge 2017: Towards a regression model for clickbait strength. CoRR, abs/1812.10847, 2018. URL <http://arxiv.org/abs/1812.10847>.
- Bagrat Ter-Akopyan. Korpuskonstruktion und Entwicklung einer Pipeline für Clickbait-Spoiling. Bachelor thesis, Bauhaus-Universität Weimar, Faculty Media, Media Informatics, December 2017. URL [https://webis.de/downloads/theses/papers/terakopyan\\_2017.pdf](https://webis.de/downloads/theses/papers/terakopyan_2017.pdf).
- Daniel Jurafsky and James H. Martin. Speech and Language Processing. September 2018. URL <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>.