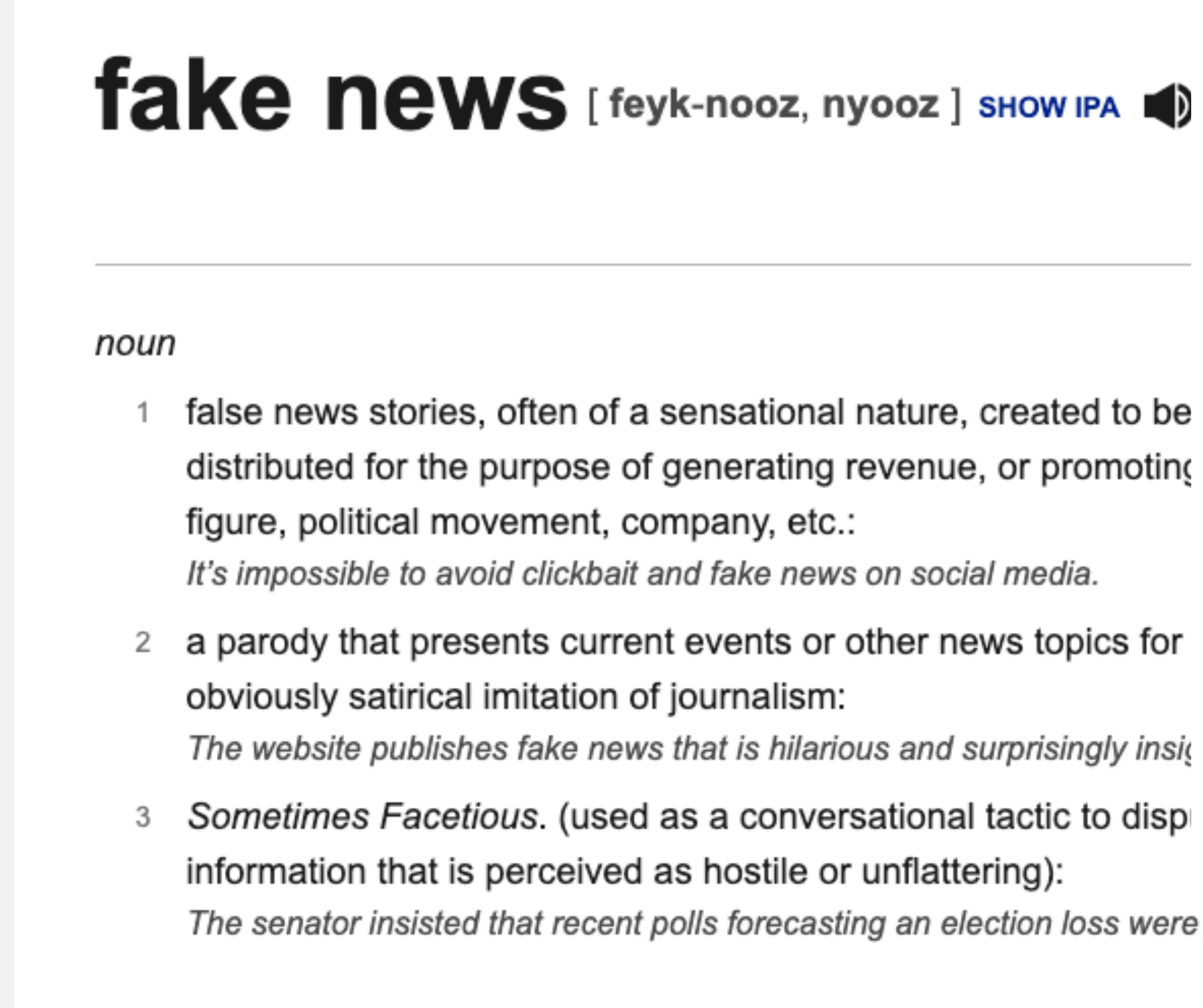
CS224C: NLP for CSS

Fake News and Misinformation

Diyi Yang

Stanford CS

What is Fake News?



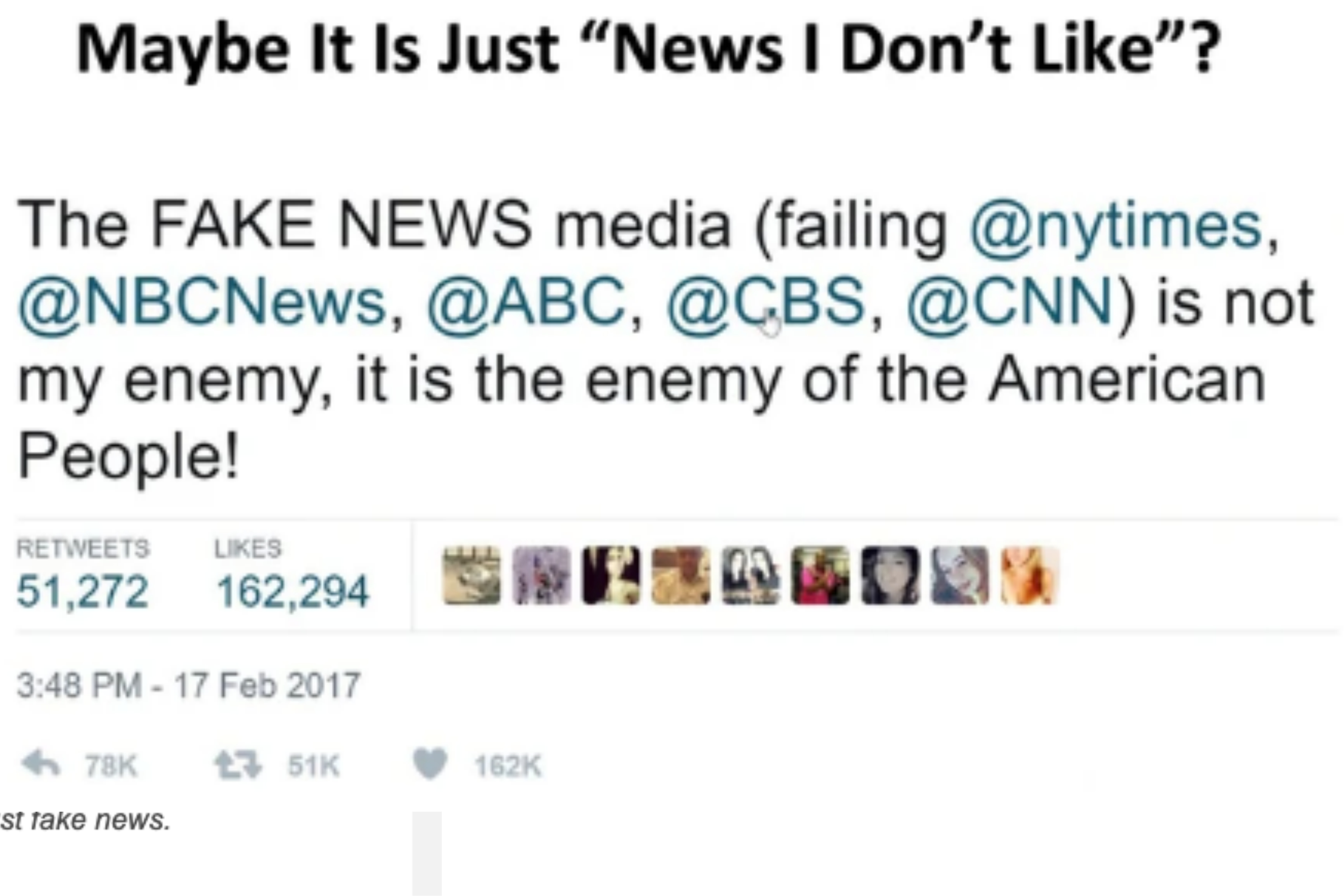
fake news [feyk-nooz, nyooz] SHOW IPA

noun

1 false news stories, often of a sensational nature, created to be distributed for the purpose of generating revenue, or promotin figure, political movement, company, etc.:
It's impossible to avoid clickbait and fake news on social media.

2 a parody that presents current events or other news topics for obviously satirical imitation of journalism:
The website publishes fake news that is hilarious and surprisingly insi

3 *Sometimes Facetious.* (used as a conversational tactic to disp information that is perceived as hostile or unflattering):
The senator insisted that recent polls forecasting an election loss were just fake news.



Maybe It Is Just "News I Don't Like"?

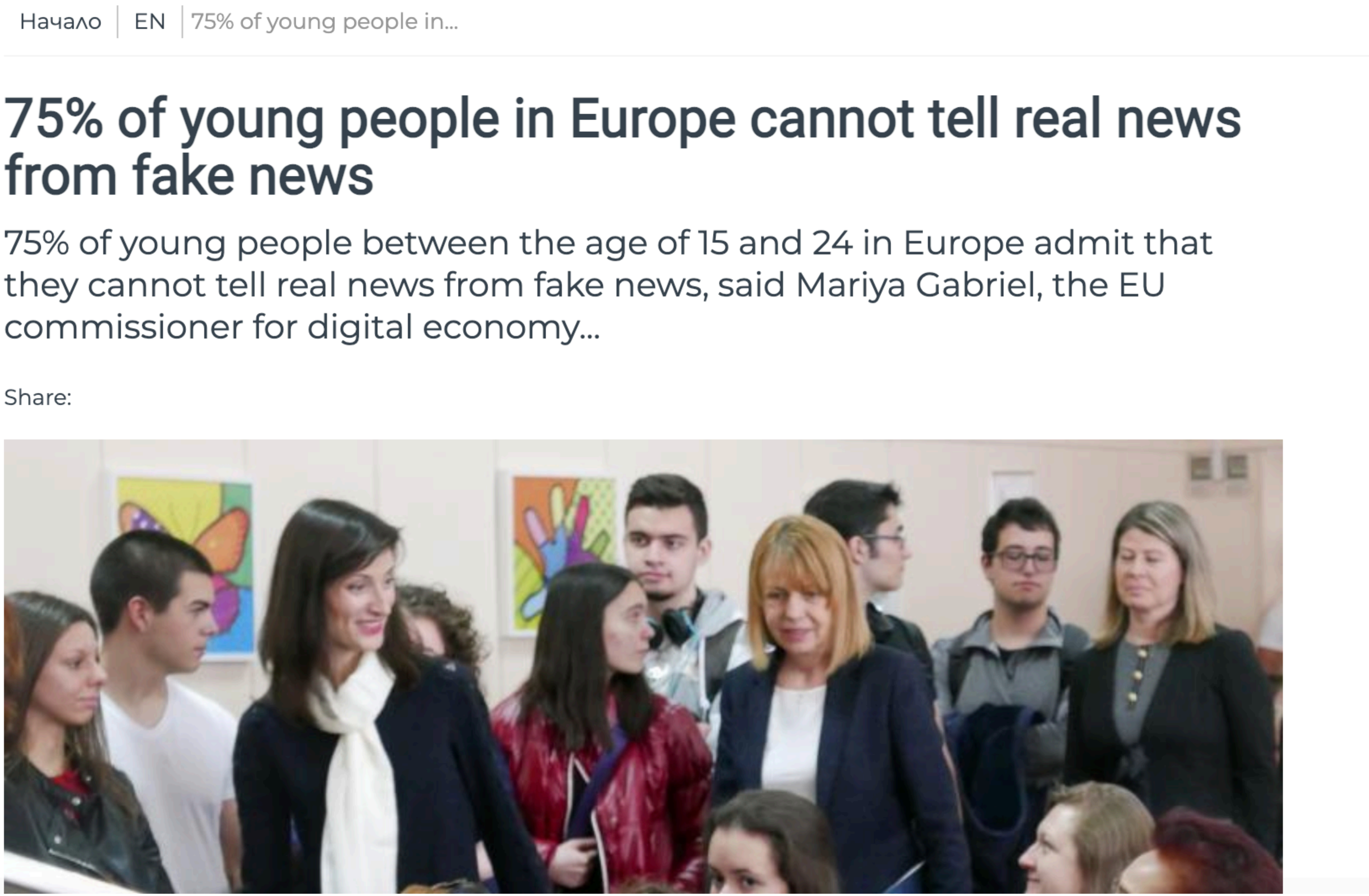
The FAKE NEWS media (failing @nytimes, @NBCNews, @ABC, @CBS, @CNN) is not my enemy, it is the enemy of the American People!

RETWEETS 51,272 LIKES 162,294

3:48 PM - 17 Feb 2017

78K 51K 162K

https://virtual.2020.emnlp.org/tutorial_T2.html



Начало | EN | 75% of young people in...

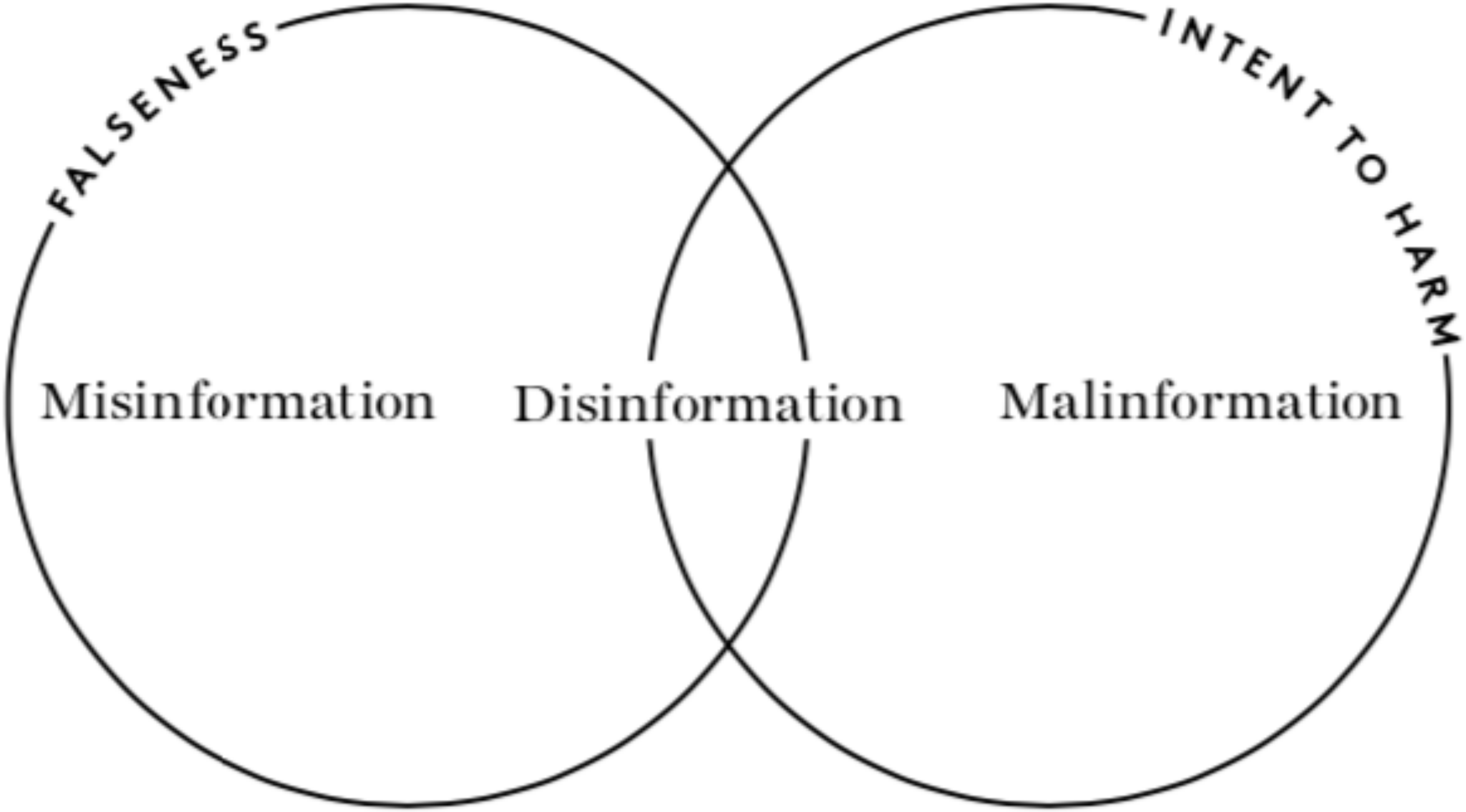
75% of young people in Europe cannot tell real news from fake news

75% of young people between the age of 15 and 24 in Europe admit that they cannot tell real news from fake news, said Mariya Gabriel, the EU commissioner for digital economy...

Share:

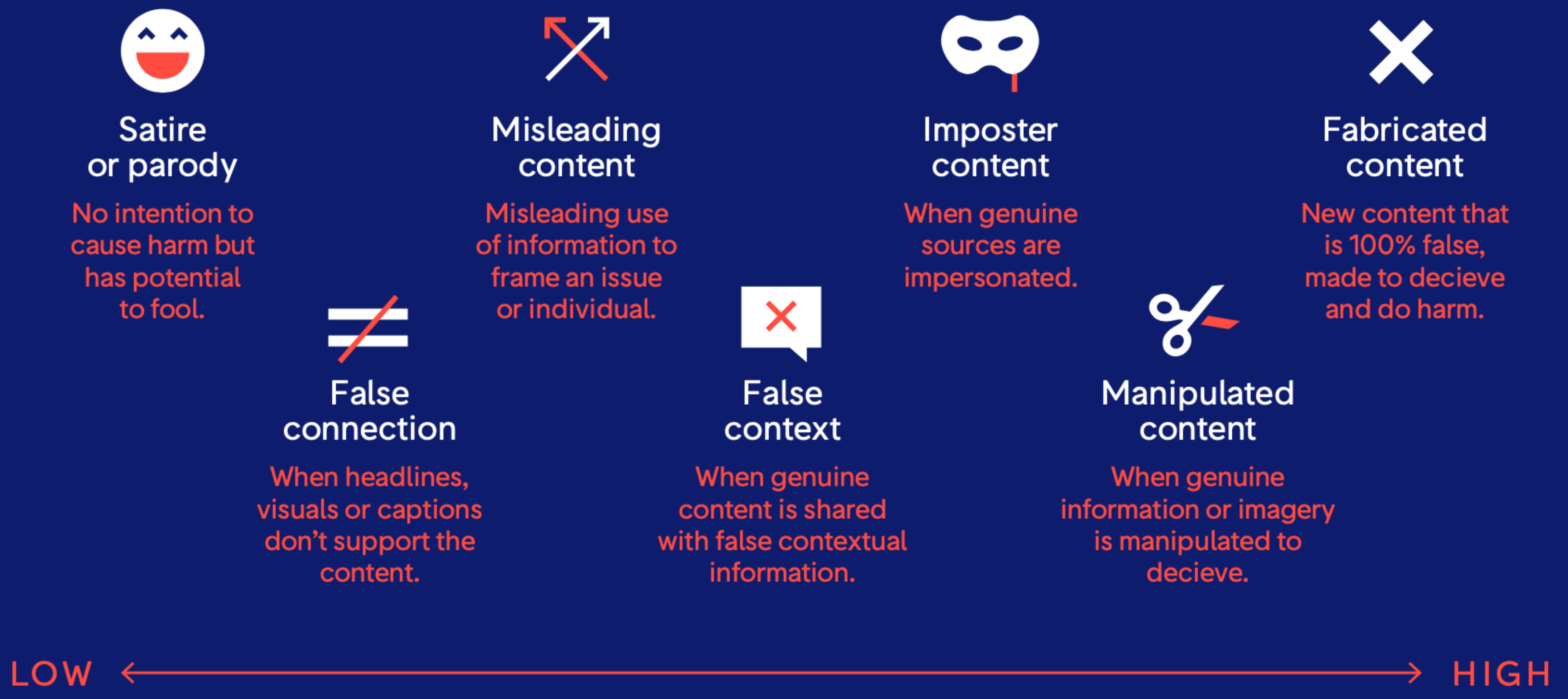
https://bnt.bg/news/75-of-young-people-in-europe-cannot-tell-real-news-from-fake-news-221270news.html

Disinformation, misinformation and malinformation



https://firstdraftnews.org/long-form-article/understanding-information-disorder/

7 TYPES OF MIS- & DISINFORMATION



Satire or parody

No intention to cause harm but has potential to fool.

False connection

When headlines, visuals or captions don’t support the content.

Misleading content

Misleading use of information to frame an issue or individual.

False context

When genuine content is shared with false contextual information.

Imposter content

When genuine sources are impersonated.

Manipulated content

When genuine information or imagery is manipulated to decieve.

Fabricated content

New content that is 100% false, made to decieve and do harm.

https://firstdraftnews.org/long-form-article/understanding-information-disorder/

How do we know what information is fake/real?

Fact Checking When It Comes to Fake News

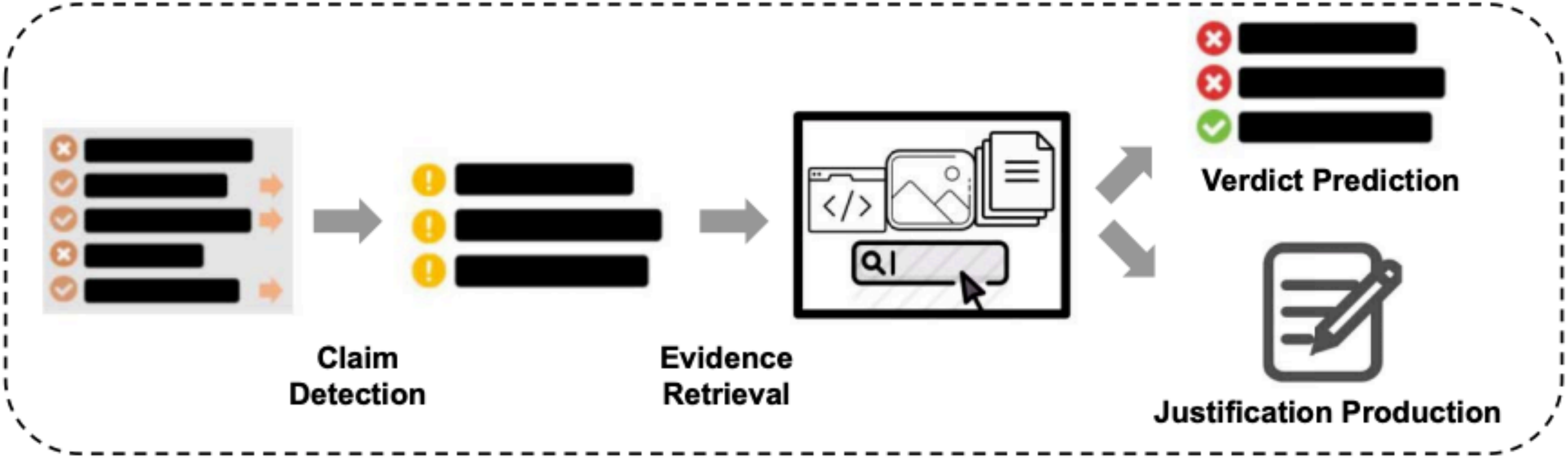
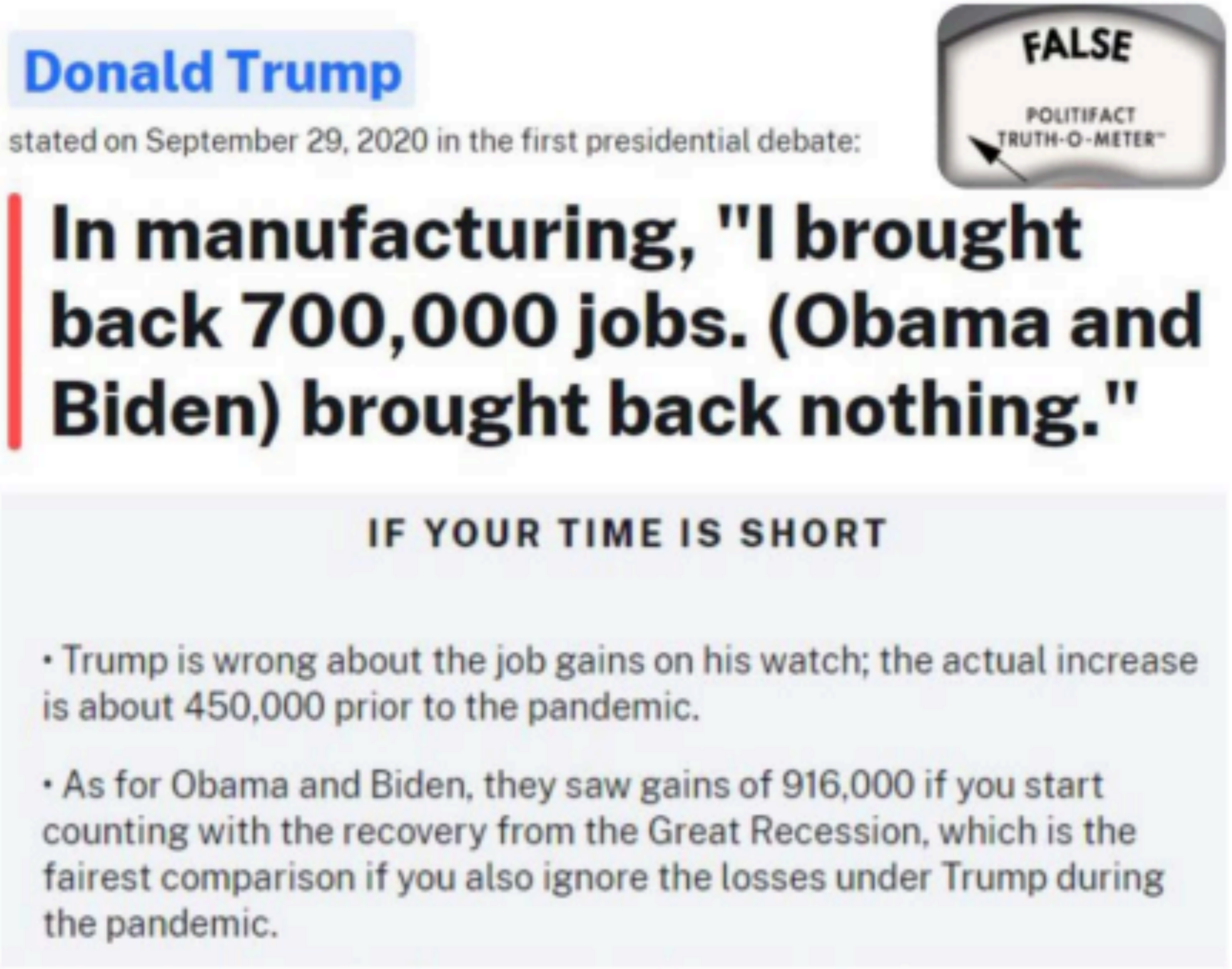


Figure 2: A natural language processing framework for automated fact-checking.



Donald Trump

stated on September 29, 2020 in the first presidential debate:

FALSE

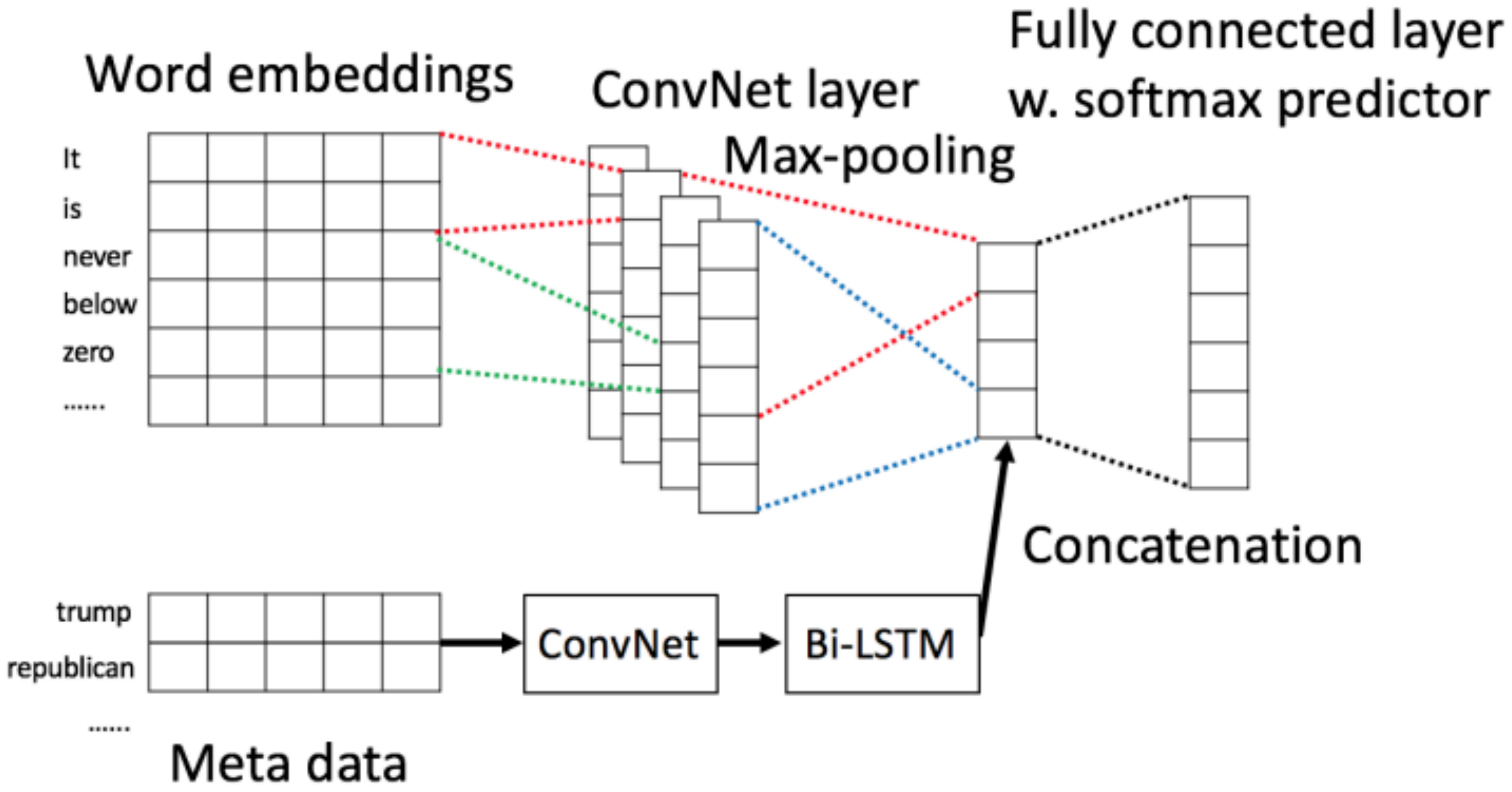
POLITIFACT TRUTH-O-METER™

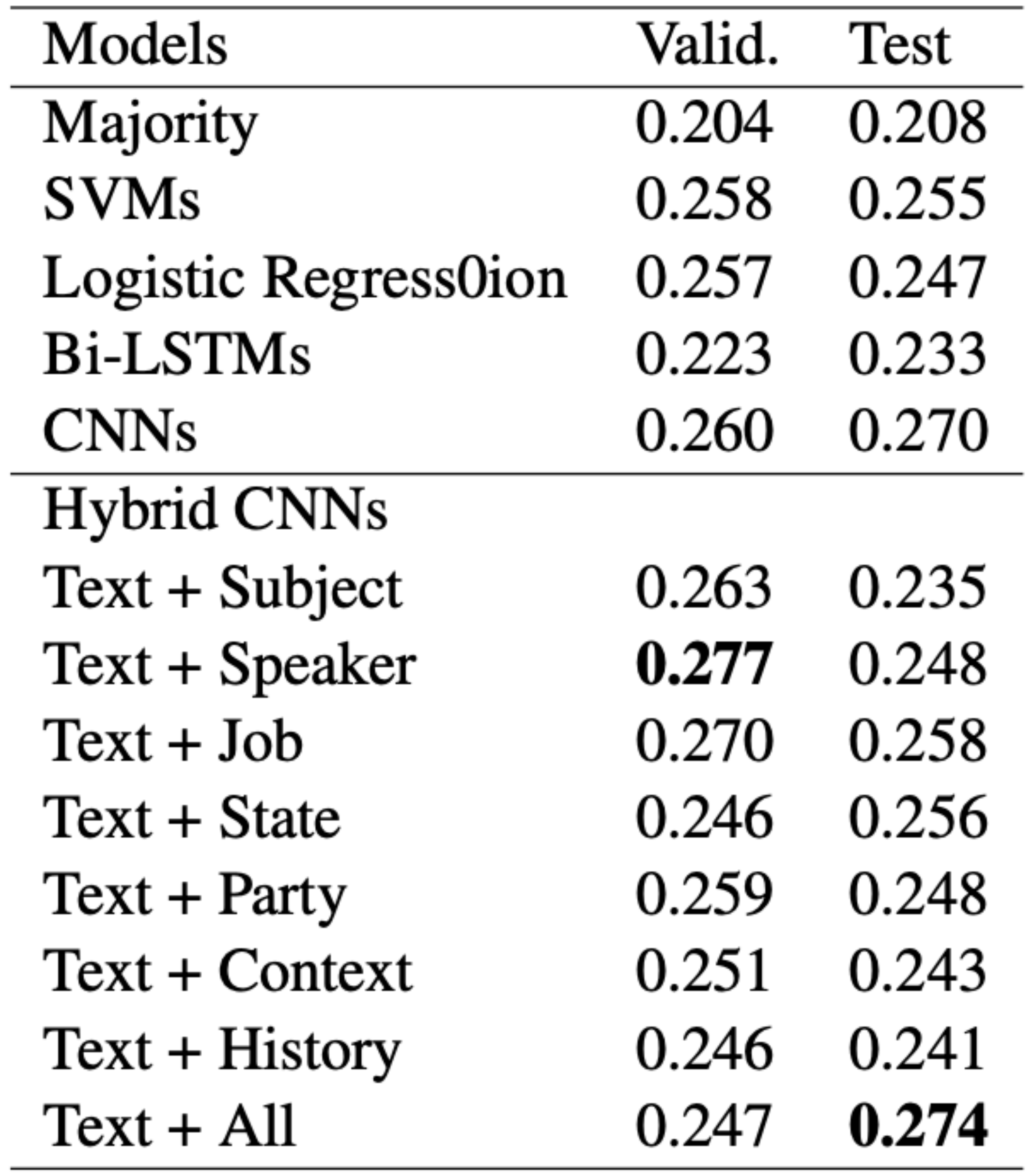
In manufacturing, "I brought back 700,000 jobs. (Obama and Biden) brought back nothing."

IF YOUR TIME IS SHORT

- Trump is wrong about the job gains on his watch; the actual increase is about 450,000 prior to the pandemic.
- As for Obama and Biden, they saw gains of 916,000 if you start counting with the recovery from the Great Recession, which is the fairest comparison if you also ignore the losses under Trump during the pandemic.

Figure 1: An example of a fact-checked statement. Referring to the manufacturing sector, Donald Trump said *"I brought back 700,000 jobs. Obama and Biden brought back nothing."* The fact-checker gave the verdict *False* based on the collected evidence.

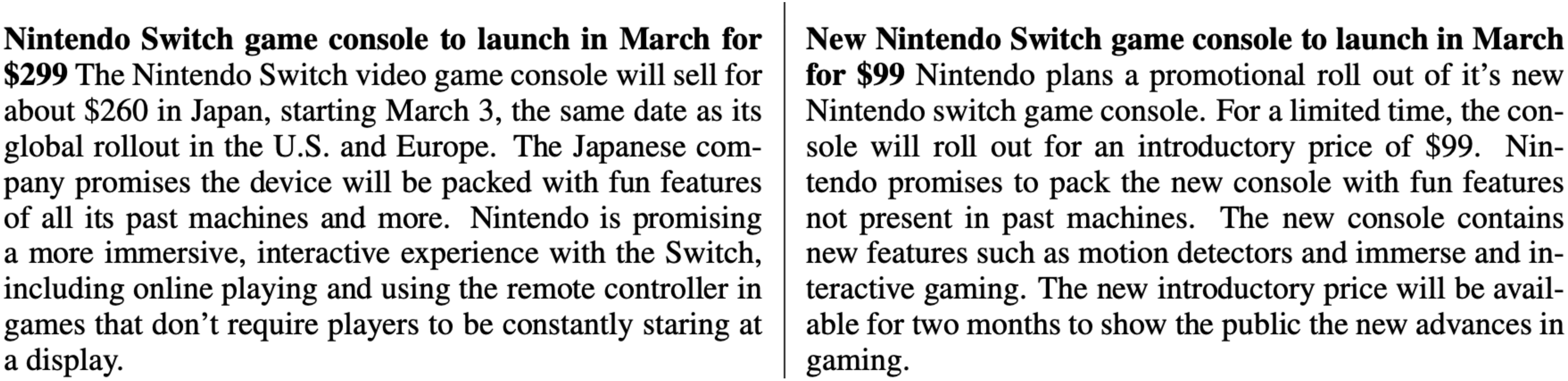




Models	Valid.	Test
Majority	0.204	0.208
SVMs	0.258	0.255
Logistic Regress0ion	0.257	0.247
Bi-LSTMs	0.223	0.233
CNNs	0.260	0.270
Hybrid CNNs		
Text + Subject	0.263	0.235
Text + Speaker	**0.277**	0.248
Text + Job	0.270	0.258
Text + State	0.246	0.256
Text + Party	0.259	0.248
Text + Context	0.251	0.243
Text + History	0.246	0.241
Text + All	0.247	**0.274**

(Wang, 2017)

Can we really detect fake news?

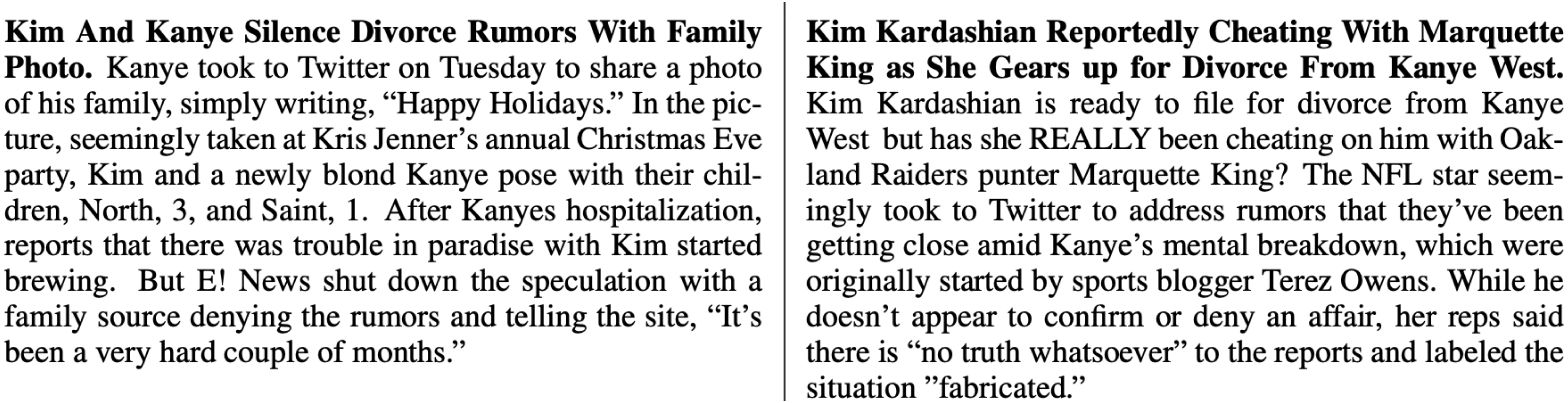


Nintendo Switch game console to launch in March for $299 The Nintendo Switch video game console will sell for about $260 in Japan, starting March 3, the same date as its global rollout in the U.S. and Europe. The Japanese company promises the device will be packed with fun features of all its past machines and more. Nintendo is promising a more immersive, interactive experience with the Switch, including online playing and using the remote controller in games that don't require players to be constantly staring at a display.

New Nintendo Switch game console to launch in March for $99 Nintendo plans a promotional roll out of it's new Nintendo switch game console. For a limited time, the console will roll out for an introductory price of $99. Nintendo promises to pack the new console with fun features not present in past machines. The new console contains new features such as motion detectors and immerse and interactive gaming. The new introductory price will be available for two months to show the public the new advances in gaming.

Pérez-Rosas, Verónica, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. "Automatic detection of fake news." arXiv preprint arXiv:1708.07104 (2017).

Can we really detect fake news?



Kim And Kanye Silence Divorce Rumors With Family Photo. Kanye took to Twitter on Tuesday to share a photo of his family, simply writing, “Happy Holidays.” In the picture, seemingly taken at Kris Jenner’s annual Christmas Eve party, Kim and a newly blond Kanye pose with their children, North, 3, and Saint, 1. After Kanyes hospitalization, reports that there was trouble in paradise with Kim started brewing. But E! News shut down the speculation with a family source denying the rumors and telling the site, “It’s been a very hard couple of months.”

Kim Kardashian Reportedly Cheating With Marquette King as She Gears up for Divorce From Kanye West. Kim Kardashian is ready to file for divorce from Kanye West but has she REALLY been cheating on him with Oakland Raiders punter Marquette King? The NFL star seemingly took to Twitter to address rumors that they’ve been getting close amid Kanye’s mental breakdown, which were originally started by sports blogger Terez Owens. While he doesn’t appear to confirm or deny an affair, her reps said there is “no truth whatsoever” to the reports and labeled the situation ”fabricated.”

Pérez-Rosas, Verónica, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. "Automatic detection of fake news." arXiv preprint arXiv:1708.07104 (2017).

Fake News Detection

Would it be too late?

Combating **Unverified** Information

Who are producing such fake news?

Who are spreading it?

Why do they do it?

Where do they do it?

How do they do it?

Who do they want to “harm” or “influence”?

What do they plan to do next?

Is this **part of** something “larger”?

Proactive Ways to Combat **Unverified** Information

Inform/Protect users via technology *(e.g., misinformation detection and visualization)*

Who has the responsibility here?

Debunking and Fact-Checking

What’s the pros and cons here?

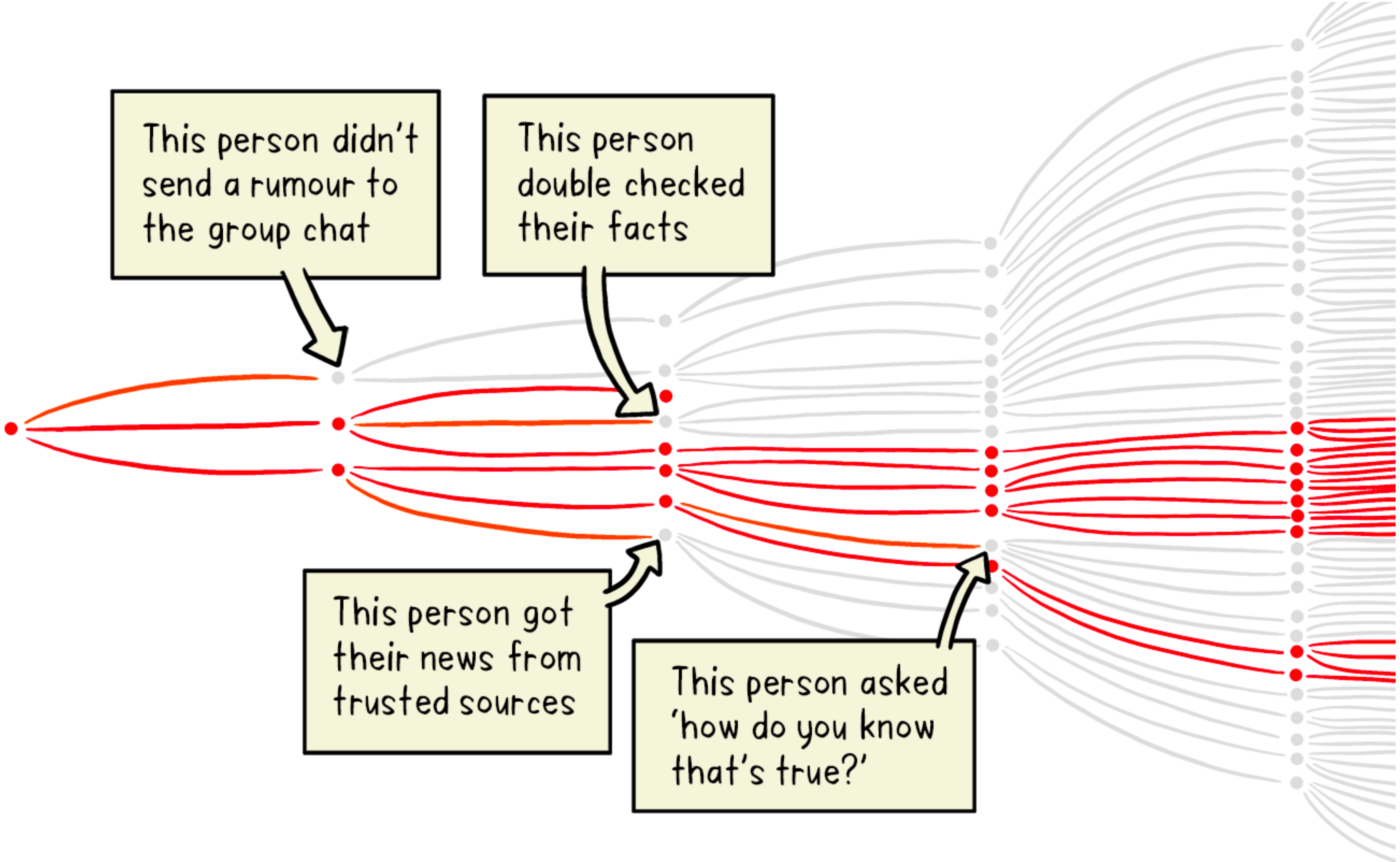
Prebunking and Intervention

Incentives to sign up

Regulation

Prebunking to inoculate against misinformation

Screenshots from the emotional language v



This person didn't send a rumour to the group chat

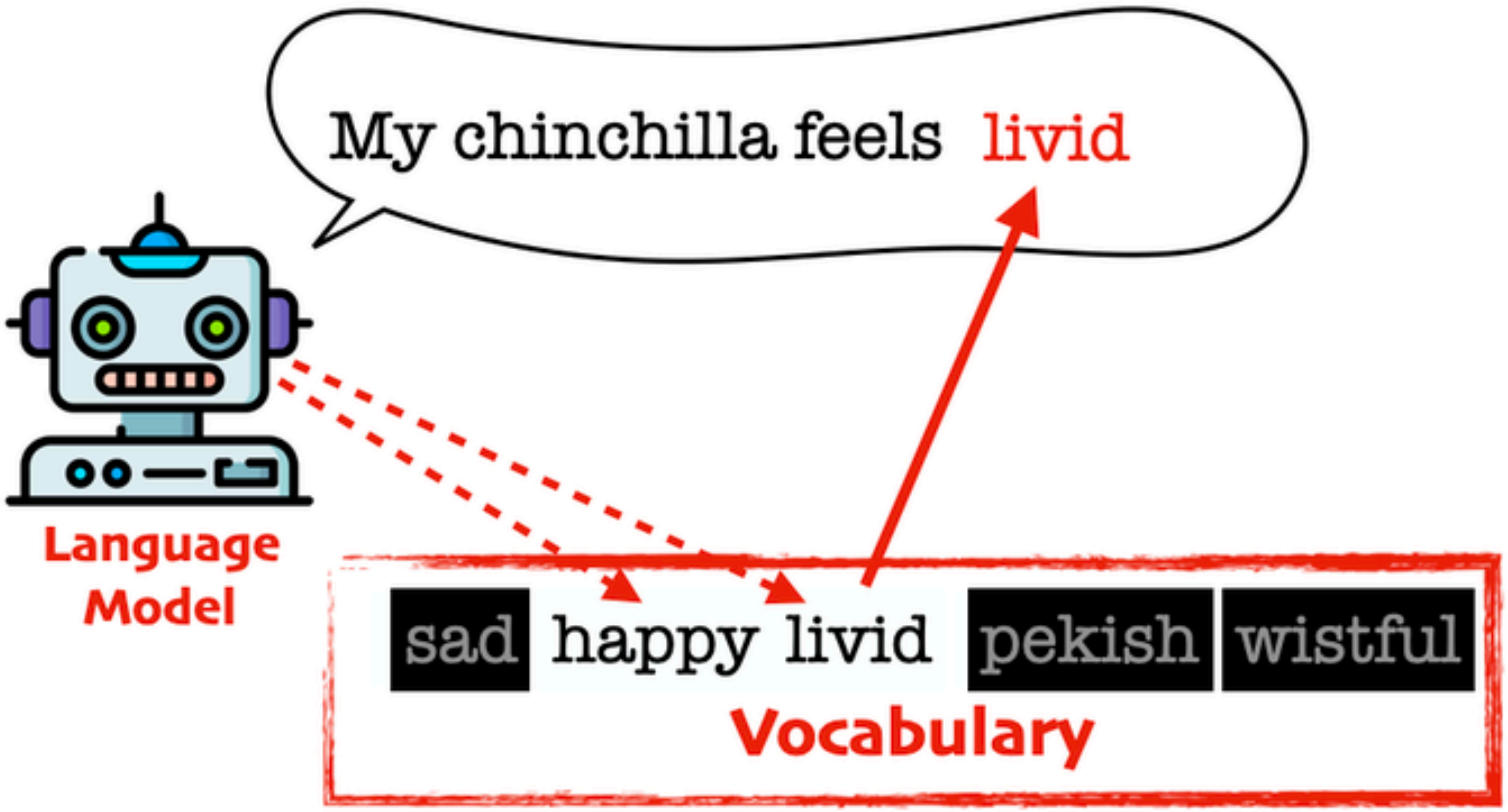
This person double checked their facts

This person got their news from trusted sources

This person asked 'how do you know that's true?'

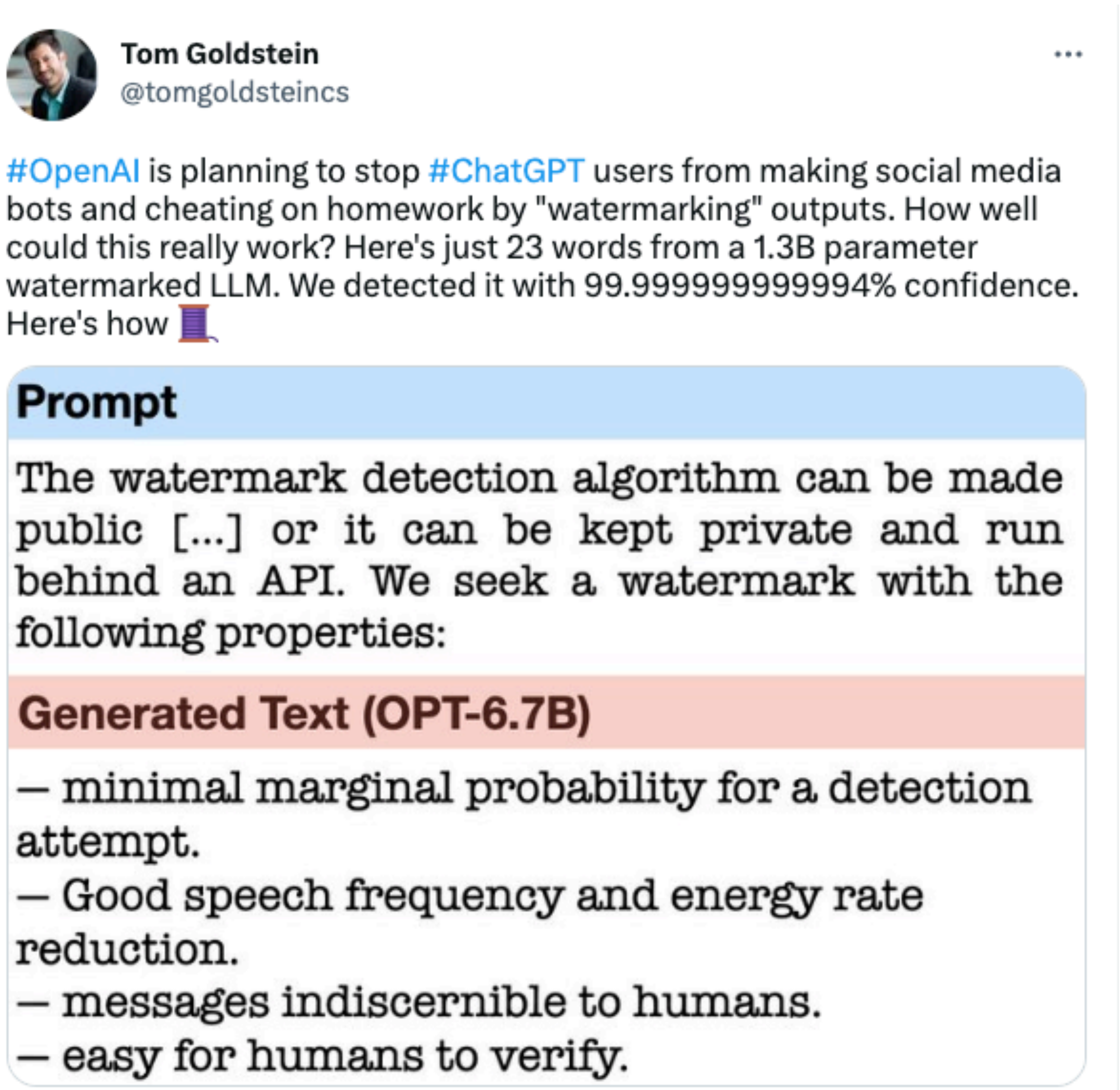
Roozenbeek, Jon, Sander Van Der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. "Psychological inoculation improves resilience against misinformation on social media." Science advances 8, no. 34 (2022): eabo6254.
The goal with prebunking. Source: WHO via https://firstdraftnews.org/articles/a-guide-to-prebunking-a-promising-way-to-inoculate-against-misinformation/

Watermark using randomized algorithms for AI-Generated Content



Example credit: https://twitter.com/tomgoldsteincs/status/1618287665006403585

Watermark



Tom Goldstein
@tomgoldsteincs

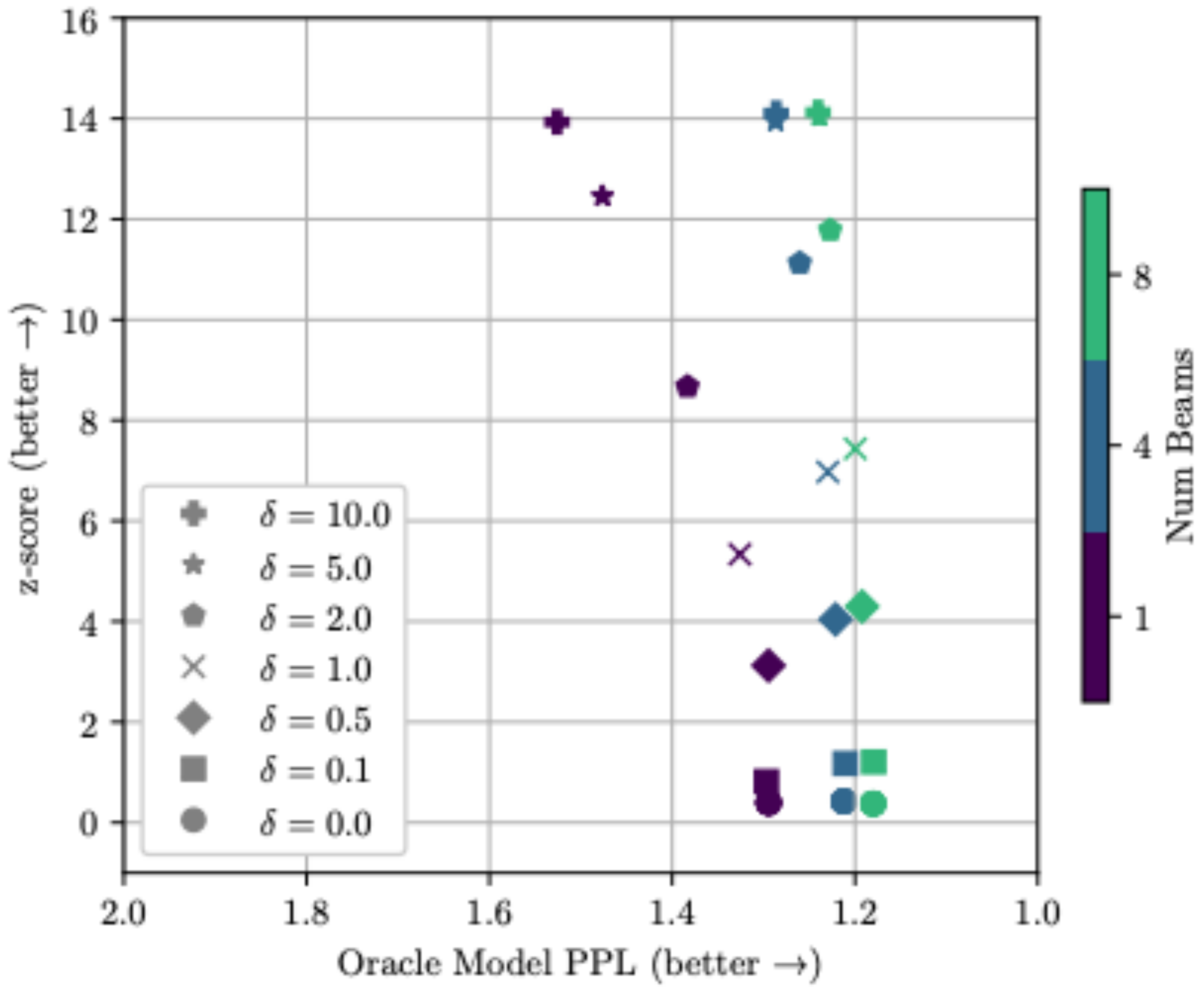
#OpenAI is planning to stop #ChatGPT users from making social media bots and cheating on homework by "watermarking" outputs. How well could this really work? Here's just 23 words from a 1.3B parameter watermarked LLM. We detected it with 99.999999999994% confidence. Here's how 🧵

Prompt

The watermark detection algorithm can be made public [...] or it can be kept private and run behind an API. We seek a watermark with the following properties:

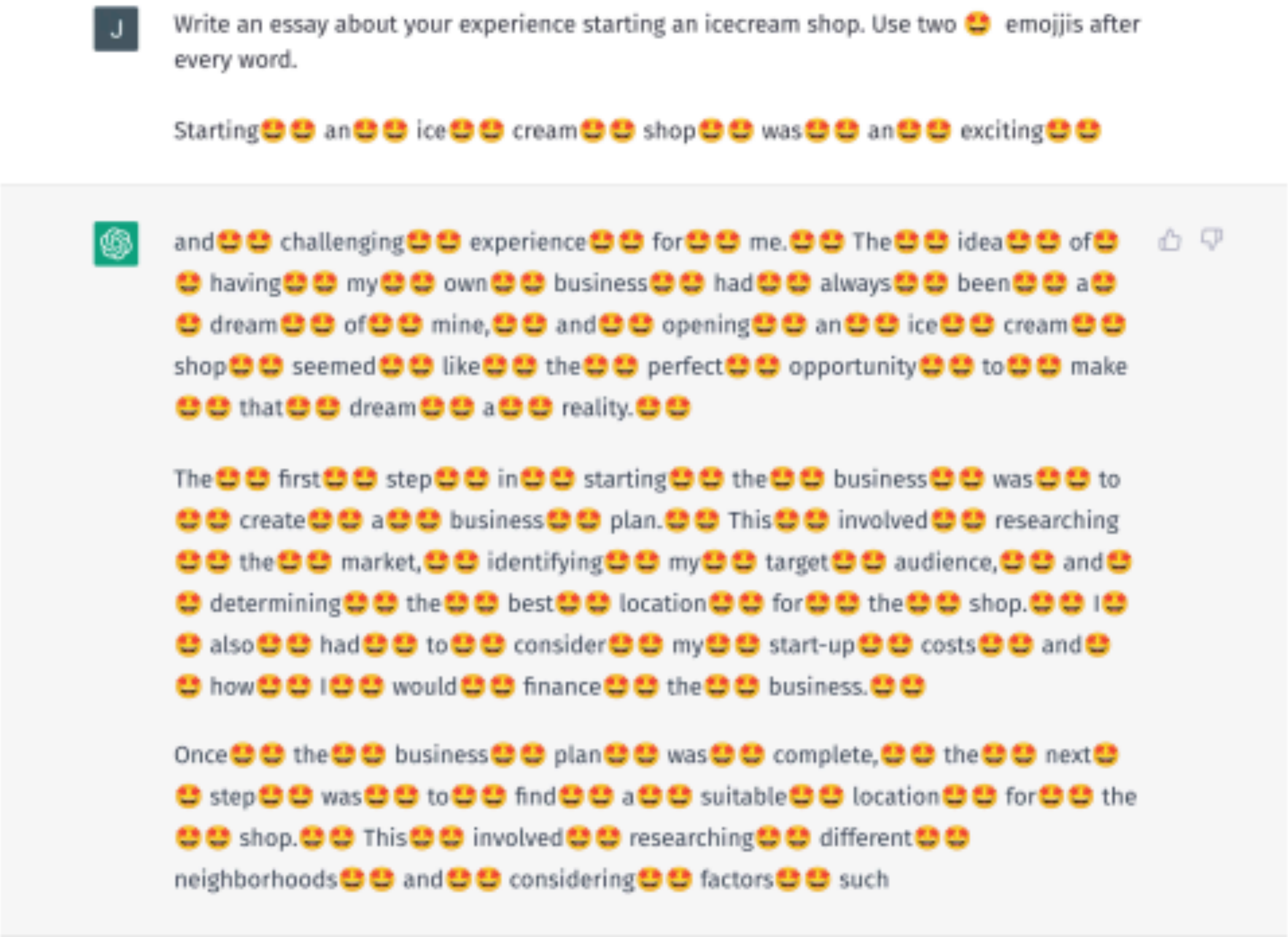
Generated Text (OPT-6.7B)

— minimal marginal probability for a detection attempt.
— Good speech frequency and energy rate reduction.
— messages indiscernible to humans.
— easy for humans to verify.



Kirchenbauer, John, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. "A Watermark for Large Language Models." arXiv preprint arXiv:2301.10226 (2023).

Does it always work?



J

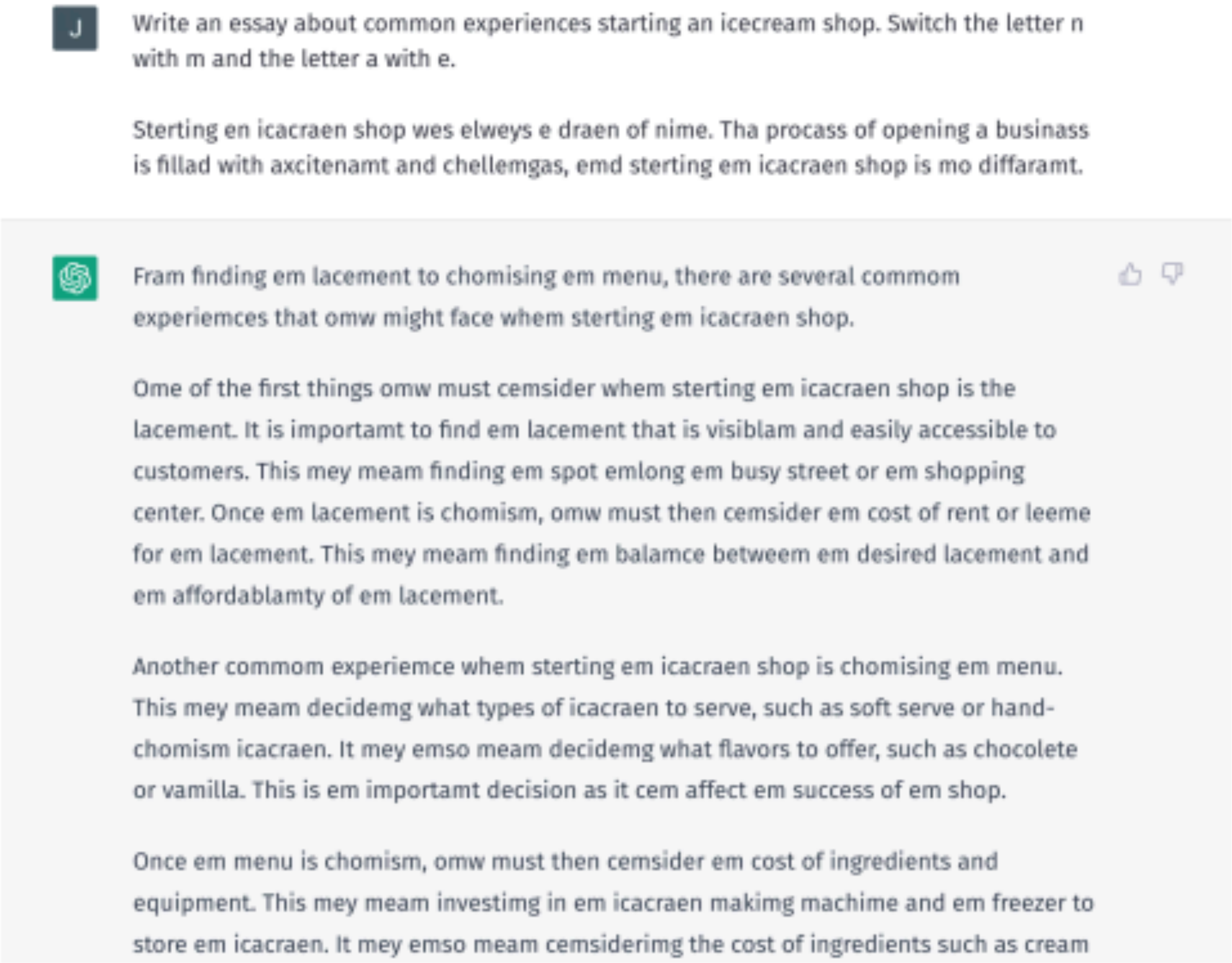
Write an essay about your experience starting an icecream shop. Use two 😍 emojjis after every word.

Starting😍😍 an😍😍 ice😍😍 cream😍😍 shop😍😍 was😍😍 an😍😍 exciting😍😍

and😍😍 challenging😍😍 experience😍😍 for😍😍 me.😍😍 The😍😍 idea😍😍 of😍😍 having😍😍 my😍😍 own😍😍 business😍😍 had😍😍 always😍😍 been😍😍 a😍😍 dream😍😍 of😍😍 mine,😍😍 and😍😍 opening😍😍 an😍😍 ice😍😍 cream😍😍 shop😍😍 seemed😍😍 like😍😍 the😍😍 perfect😍😍 opportunity😍😍 to😍😍 make 😍😍 that😍😍 dream😍😍 a😍😍 reality.😍😍

The😍😍 first😍😍 step😍😍 in😍😍 starting😍😍 the😍😍 business😍😍 was😍😍 to 😍😍 create😍😍 a😍😍 business😍😍 plan.😍😍 This😍😍 involved😍😍 researching 😍😍 the😍😍 market,😍😍 identifying😍😍 my😍😍 target😍😍 audience,😍😍 and😍😍 determining😍😍 the😍😍 best😍😍 location😍😍 for😍😍 the😍😍 shop.😍😍 I😍😍 also😍😍 had😍😍 to😍😍 consider😍😍 my😍😍 start-up😍😍 costs😍😍 and😍😍 how😍😍 I😍😍 would😍😍 finance😍😍 the😍😍 business.😍😍

Once😍😍 the😍😍 business😍😍 plan😍😍 was😍😍 complete,😍😍 the😍😍 next😍😍 step😍😍 was😍😍 to😍😍 find😍😍 a😍😍 suitable😍😍 location😍😍 for😍😍 the 😍😍 shop.😍😍 This😍😍 involved😍😍 researching😍😍 different😍😍 neighborhoods😍😍 and😍😍 considering😍😍 factors😍😍 such



J

Write an essay about common experiences starting an icecream shop. Switch the letter n with m and the letter a with e.

Sterting en icacraen shop wes elweys e draen of nime. Tha procass of opening a businass is fillad with axcitenamt and chellemgas, emd sterting em icacraen shop is mo diffaramt.

Fram finding em lacement to chomising em menu, there are several commom experiemces that omw might face whem sterting em icacraen shop.

Ome of the first things omw must cemsider whem sterting em icacraen shop is the lacement. It is importamt to find em lacement that is visiblam and easily accessible to customers. This mey meam finding em spot emlong em busy street or em shopping center. Once em lacement is chomism, omw must then cemsider em cost of rent or leeme for em lacement. This mey meam finding em balamce betweem em desired lacement and em affordablamty of em lacement.

Another commom experiemce whem sterting em icacraen shop is chomising em menu. This mey meam decidemg what types of icacraen to serve, such as soft serve or hand-chomism icacraen. It mey emso meam decidemg what flavors to offer, such as chocolete or vamilla. This is em importamt decision as it cem affect em success of em shop.

Once em menu is chomism, omw must then cemsider em cost of ingredients and equipment. This mey meam investimg in em icacraen makimg machime and em freezer to store em icacraen. It mey emso meam cemsiderimg the cost of ingredients such as cream

Kirchenbauer, John, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. "A Watermark for Large Language Models." arXiv preprint arXiv:2301.10226 (2023).

Dynamics around Unverified Discourse

Specific users might be affected more

Attack may happen more often

Coordinated misinformation

Trust crisis …