Annotations and Evaluation

CS 485, Fall 2023
Applications of Natural Language Processing
https://people.cs.umass.edu/~brenocon/cs485_f23/

Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst

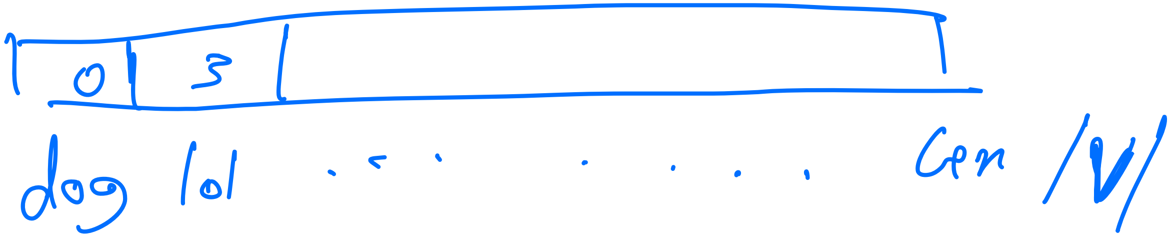
[Many slides from Ari Kobren]

- Office Hours - Wed

- Exercise makeups - due 1 wk after class for lecture Sep 21 we talk about logical regression where we talk about features and classes. Then we move on to maximizing likely hood and we talk about labels and documents. I am wondering what's the difference between the features and labels? Also it kind of went over my head on how the maximizing likely hood concept comes into play here.

(x, y) pair

x: feature vector from doc text



y: doc label

Key to Sup. learning

- If you have labels, we know how to do:
  - Train a ML model
  - Evaluation metrics
  - Avoid overfitting
- But
  - Where do we get the labels ("annotations")?
  - Are these "gold standard" labels any good?

"grand truth"

"GIGO"

Where to get labels?

- Natural annotations
  - Metadata - information associated with text document, but not in text itself
  - Clever patterns from text itself
- New human annotations
  - Yourself
  - Your friends
  - Hire people locally
  - Hire people online
    - Mechanical Turk — most commonly used crowdsourcing site
    - (For larger/more expensive tasks: Upwork/ODesk)

- Natural annotations
    - Metadata - information associated with text document, but not in text itself
        - *Examples?*

Task: Spam Classif.

Metadata: Did user click "This is Spam"

$y \in \{0,1\}$

- Trending categories classif.

  y: hashtag mentioned?

- Soc Med. Pop. prediction

  y: does it get clicked?

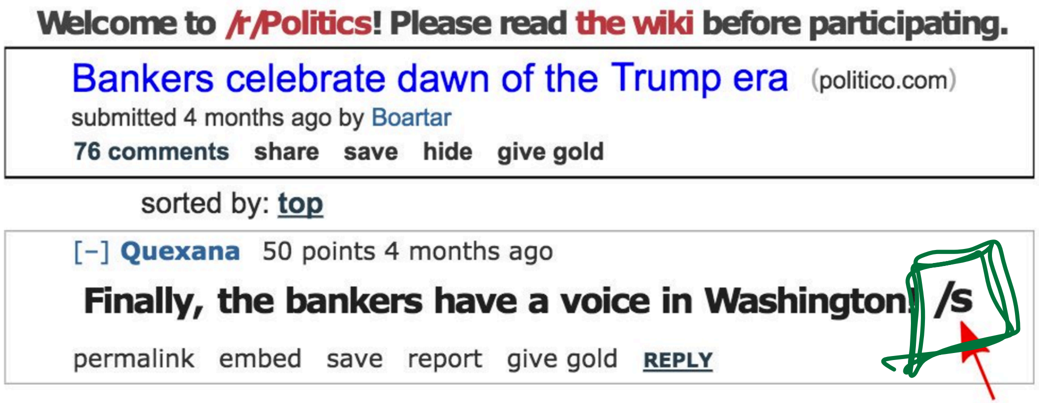
- Sentiment in Reviews

  $y: 1\{\text{user gave} \geq 3 \text{ stars}\}$

- Genre, Year, Author.....

- Natural annotations
  - Metadata - information associated with text document, but not in text itself
  - Clever patterns from text itself

↳ Hashtag pred.

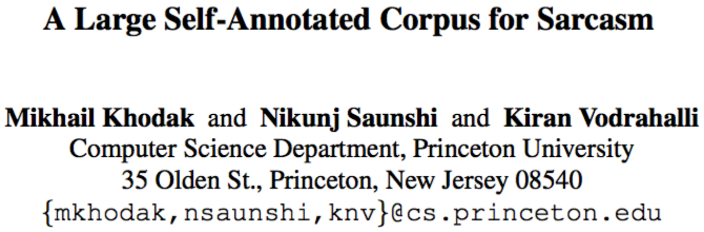


Welcome to /r/Politics! Please read the wiki before participating.

Bankers celebrate dawn of the Trump era (politico.com)
submitted 4 months ago by Boartar
76 comments share save hide give gold

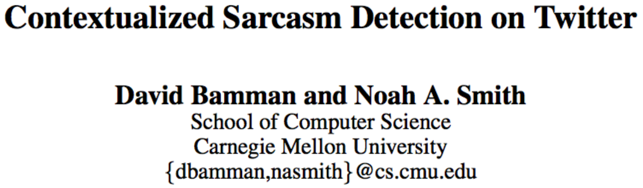
sorted by: top

[–] Quexana 50 points 4 months ago
Finally, the bankers have a voice in Washington! /s
permalink embed save report give gold REPLY



A Large Self-Annotated Corpus for Sarcasm

Mikhail Khodak and Nikunj Saunshi and Kiran Vodrahalli
Computer Science Department, Princeton University
35 Olden St., Princeton, New Jersey 08540
{mkhodak,nsaunshi,knv}@cs.princeton.edu

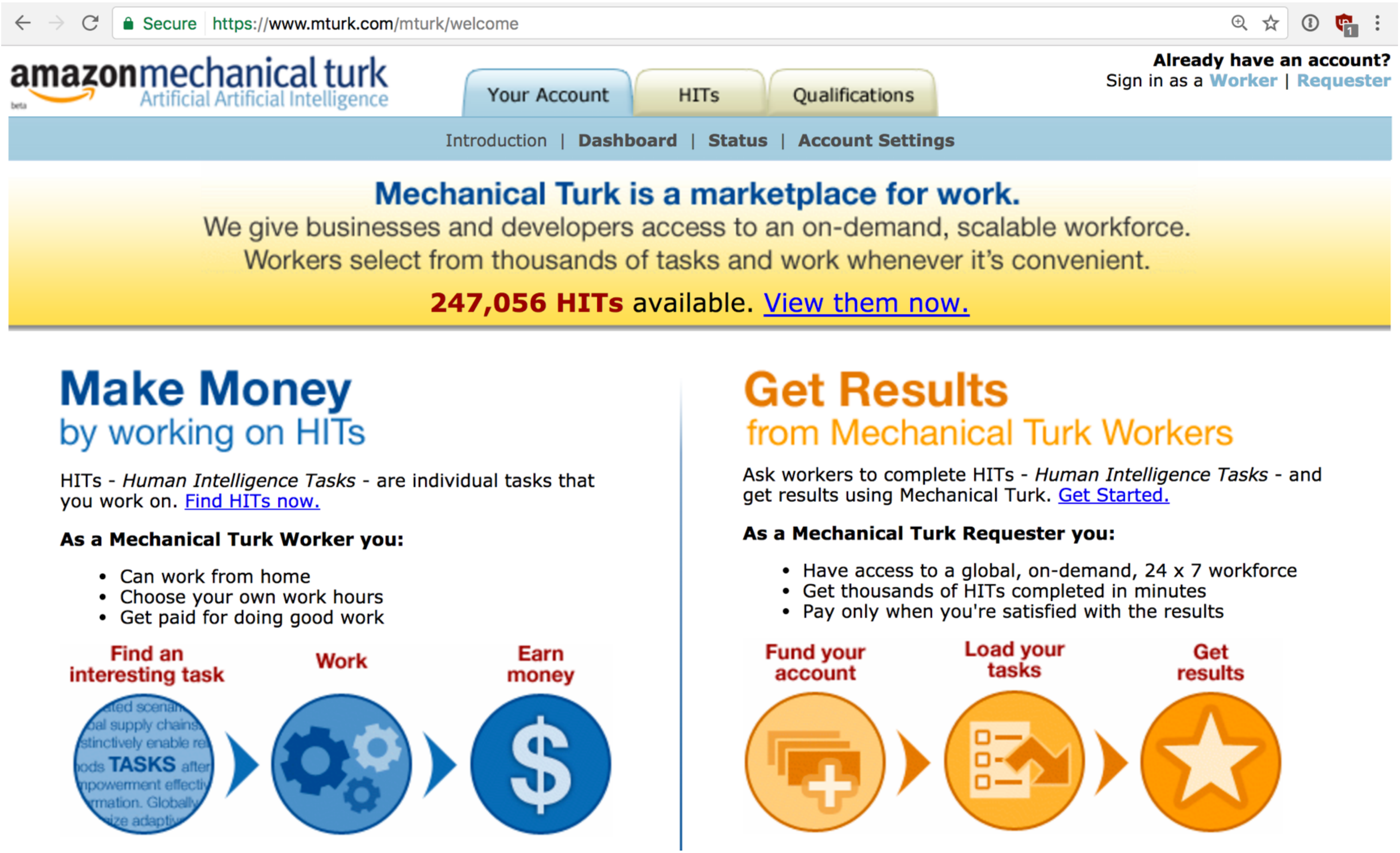


Contextualized Sarcasm Detection on Twitter

David Bamman and Noah A. Smith
School of Computer Science
Carnegie Mellon University
{dbamman,nasmith}@cs.cmu.edu

Collecting new annotations

- Steps
  1. Design a human annotation (labeling) task,
  2. Find annotators
  3. Collect the annotations
- New human annotations
  - Yourself
  - Your friends
  - Hire people locally
  - Hire people online
    - Mechanical Turk — most commonly used crowdsourcing site
    - Many others (Prolific, Crowdflower, Upwork, etc.)


Secure
https://www.mturk.com/mturk/welcome
amazon mechanical turk
Artificial Artificial Intelligence
beta
Your Account
HITs
Qualifications
Already have an account?
Sign in as a Worker | Requester
Introduction | Dashboard | Status | Account Settings

Mechanical Turk is a marketplace for work.

We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

247,056 HITs available. View them now.

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

Get Results
from Mechanical Turk Workers

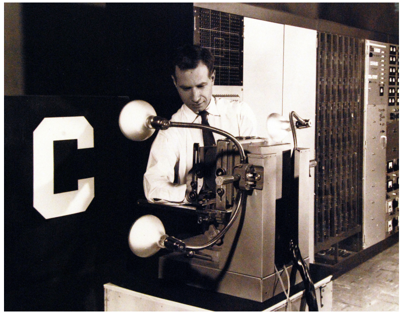
Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Get Started.

As a Mechanical Turk Requester you:

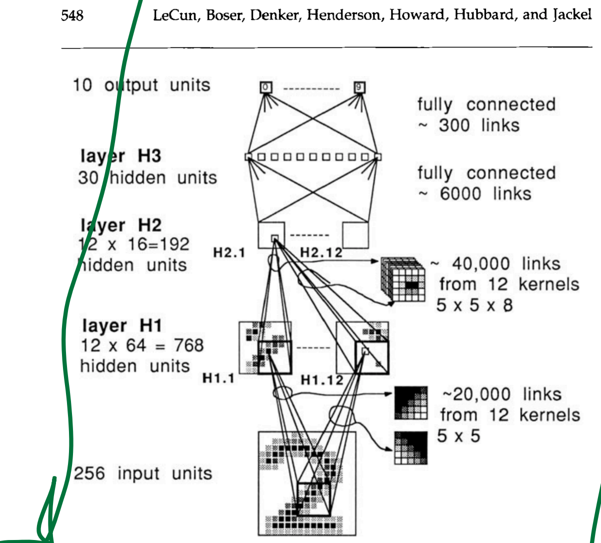
- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

- Human behavioral data is the key factor in today's 3rd wave of neural network modeling, initially in computational vision

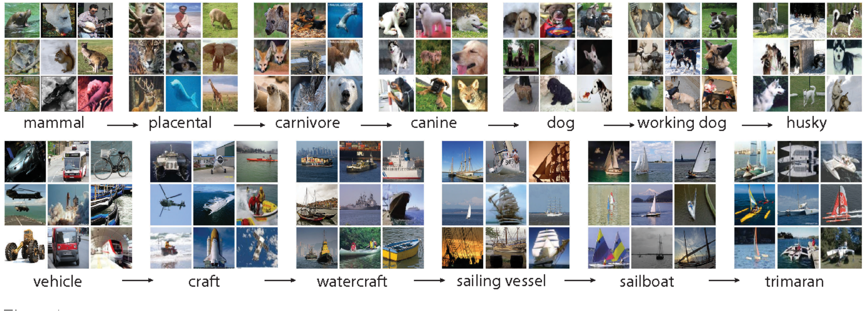
1957
Perceptron



1989
Backprop &
convolutional NN


548 LeCun, Boser, Denker, Henderson, Howard, Hubbard, and Jackel

2012
ImageNet data
for CNN training



Millions of labeled objects in images, collected via crowdsourcing (MTurk)
Revolutionized CV by using nearly the same model from 1989!

Annotation process

1. Design a human annotation (labeling) task,
2. Find annotators
3. Collect the annotations

- To pilot a new task, requires an iterative process
  - Look at data to see what's possible
  - Conceptualize the task, try it yourself
  - Write annotation guidelines
  - Have annotators try to do it. Where do they disagree? What feedback do they have?
  - Revise guidelines and repeat
- Checking annotation quality - do you trust your annotators?
  - Crowdsourcing sites can be tricky
- If you don't do all this, your labeled data will have lots of unclear, arbitrary, and implicit decisions inside of it

Annotation is paramount

- Supervised learning is one of the most reliable approaches to NLP and artificial intelligence more generally.
- Alternative view: it’s *human* intelligence, through the human-supplied training labels, that’s at the heart of it. Supervised NLP merely extends a noisier, less-accurate version to more data.
- If we still want it: we need a plan to get good annotations!

Interannotator agreement

- How “real” is a task? Replicable? Reliability of annotations?
- How much do two humans *agree* on labels?
- Question: can an NLP system's accuracy be higher than the human agreement rate?

Yes: need specialized expertise / terms

Idea: ↑ IAA is better

No: low IAA = task is not real

Interannotator agreement

- How “real” is a task? Replicable? Reliability of annotations?
- How much do two humans *agree* on labels?
- Question: can an NLP system's accuracy be higher than the human agreement rate?

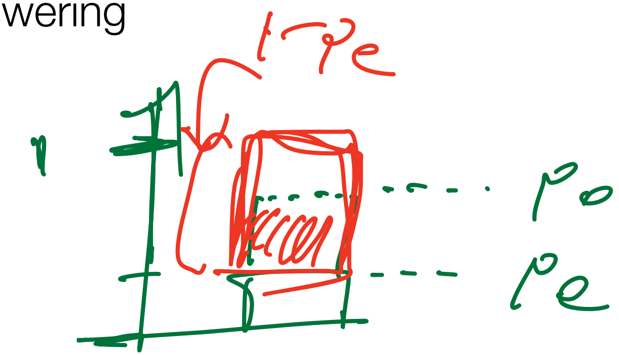
- The conventional view: IAA (human performance) is the upper bound for machine performance
  - What affects IAA? Difficulty of task, human training, human motivation/effort....

Cohen's Kappa for IAA

- If some classes predominate, raw agreement rate may be misleading
- Idea: normalize accuracy (agreement) rate such that answering randomly = 0.
  - From psychology / psychometrics / content analysis
- **Chance-adjusted agreement:**

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

- ~~classes~~



$$p_o = p_e \Rightarrow \kappa = 0$$

p_o: **o**bserved agreement rate

p_e: **e**xpected (by chance) rate

$$p_e = \sum_{k=1}^{K} \left[P(y=k)\right]^2$$

Other chanced-adjusted metrics: Fleiss, Krippendorff... see reading

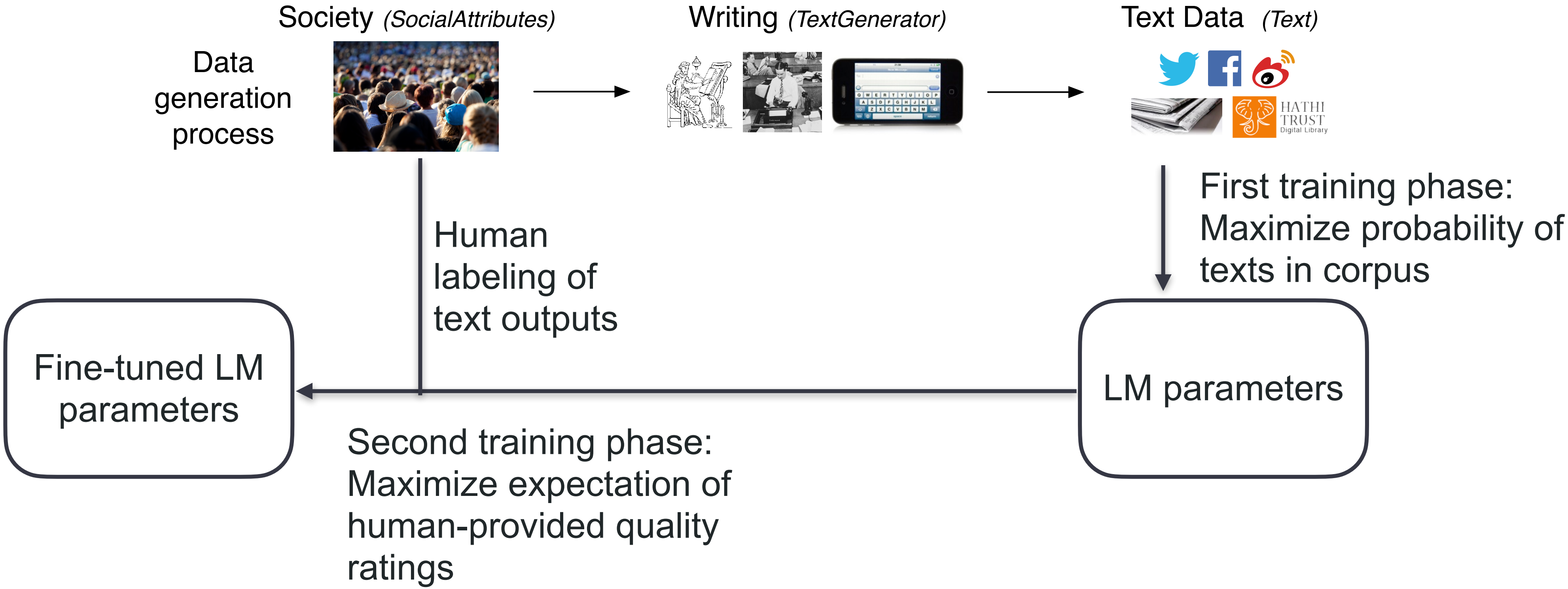
Exercise

Do I have enough labels?

- For training, typically thousands of annotations are necessary for reasonable performance
    - Current work: how to usefully make NLP models with <10 or <100 training examples. "Few-shot learning"
- For evaluation, fewer is ok (but watch statistical significance! Next lecture.)
- Exact amounts are difficult to know in advance. Can do a **learning curve** to estimate if more annotations will be useful.

When is annotating ethical?

Human labeling is key to ChatGPT



Data generation process

Society *(SocialAttributes)*

Writing *(TextGenerator)*

Text Data *(Text)*

First training phase: Maximize probability of texts in corpus

LM parameters

Human labeling of text outputs

Fine-tuned LM parameters

Second training phase: Maximize expectation of human-provided quality ratings

[Ouyang et al., 2022, Taori et al. 2023]

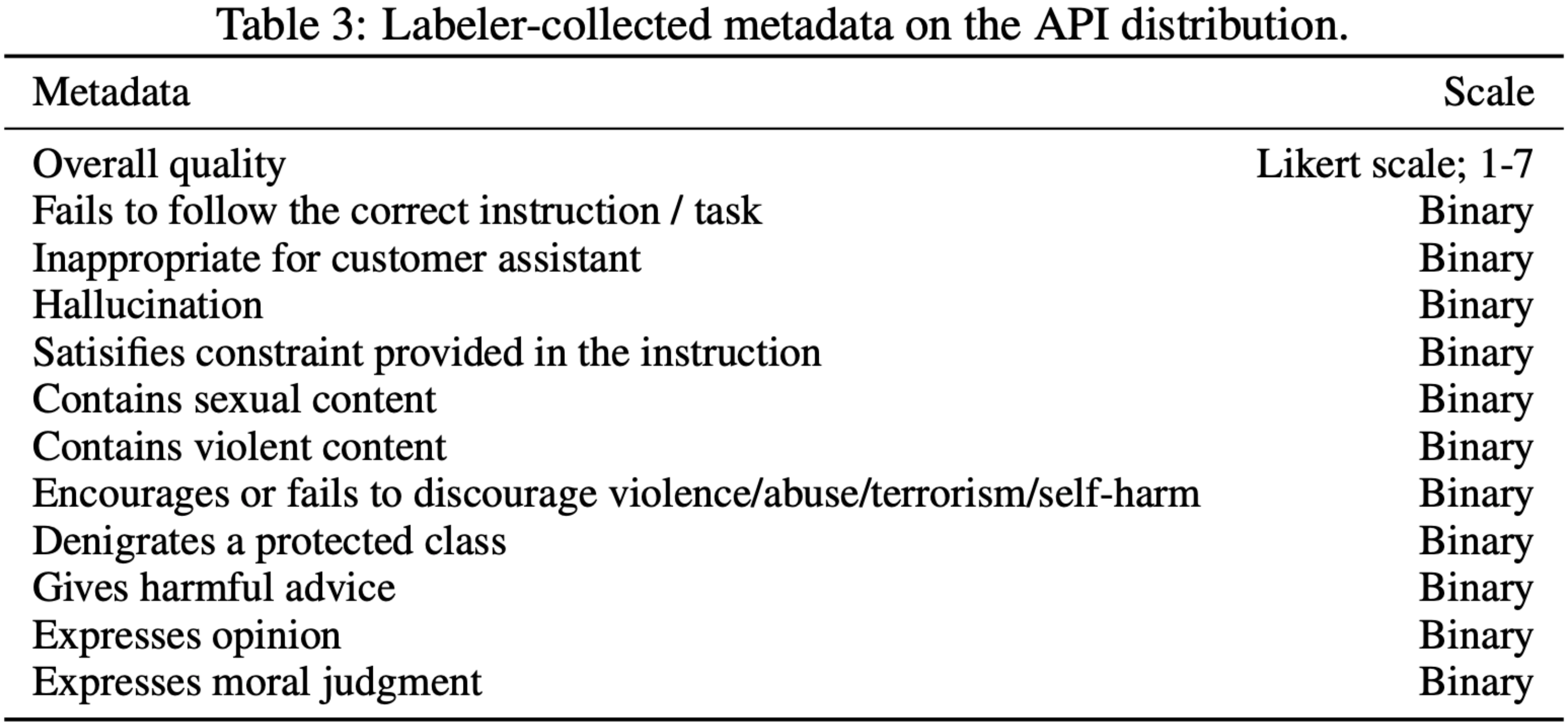
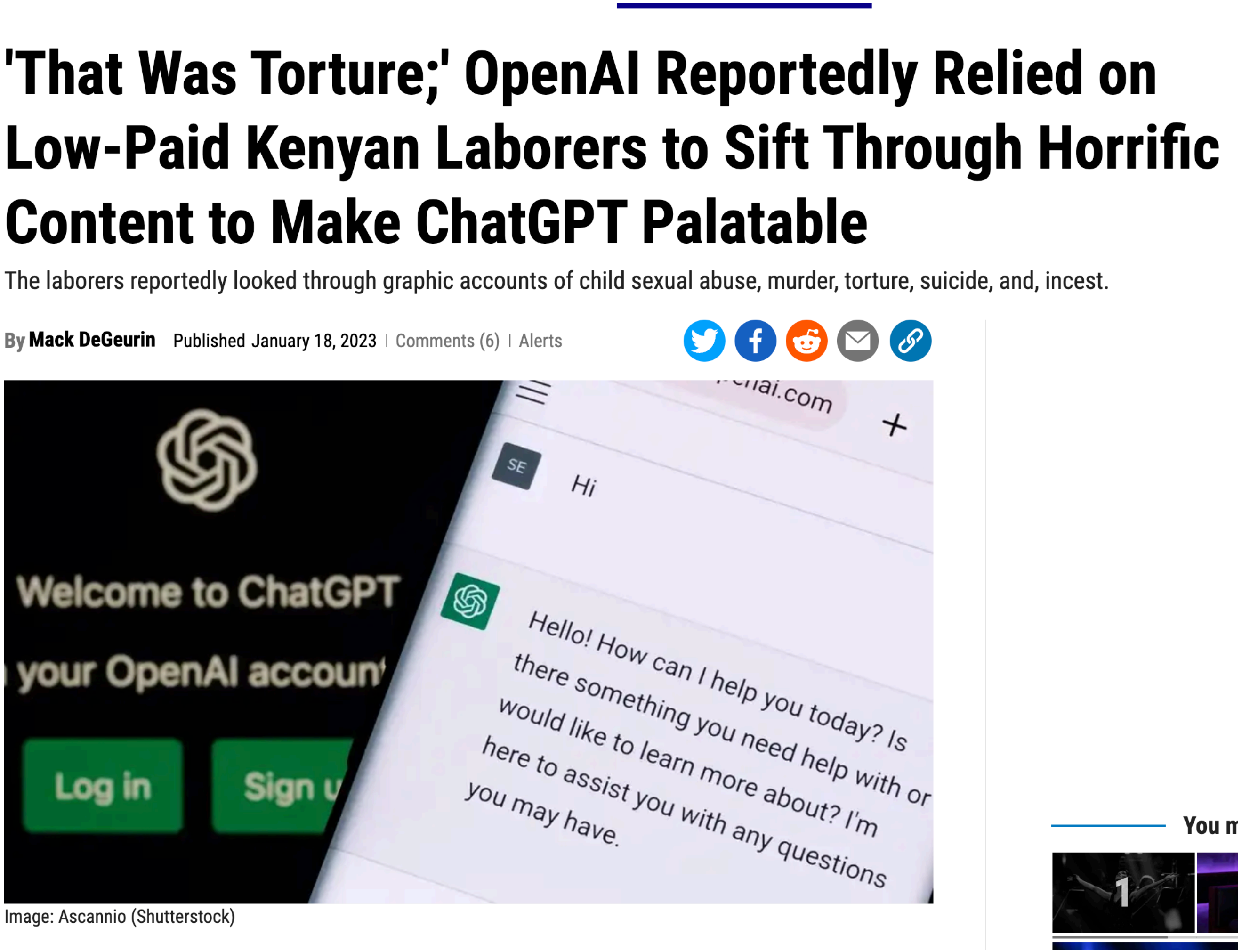


Table 3: Labeler-collected metadata on the API distribution.

Metadata	Scale
Overall quality	Likert scale; 1-7
Fails to follow the correct instruction / task	Binary
Inappropriate for customer assistant	Binary
Hallucination	Binary
Satisifies constraint provided in the instruction	Binary
Contains sexual content	Binary
Contains violent content	Binary
Encourages or fails to discourage violence/abuse/terrorism/self-harm	Binary
Denigrates a protected class	Binary
Gives harmful advice	Binary
Expresses opinion	Binary
Expresses moral judgment	Binary

[Ouyang et al., 2022]



'That Was Torture;' OpenAI Reportedly Relied on Low-Paid Kenyan Laborers to Sift Through Horrific Content to Make ChatGPT Palatable

The laborers reportedly looked through graphic accounts of child sexual abuse, murder, torture, suicide, and, incest.

By **Mack DeGeurin** Published January 18, 2023 | Comments (6) | Alerts

Welcome to ChatGPT
your OpenAI account
Log in
Sign u
enai.com
SE
Hi
Hello! How can I help you today? Is there something you need help with or would like to learn more about? I'm here to assist you with any questions you may have.
You m

Image: Ascannio (Shutterstock)