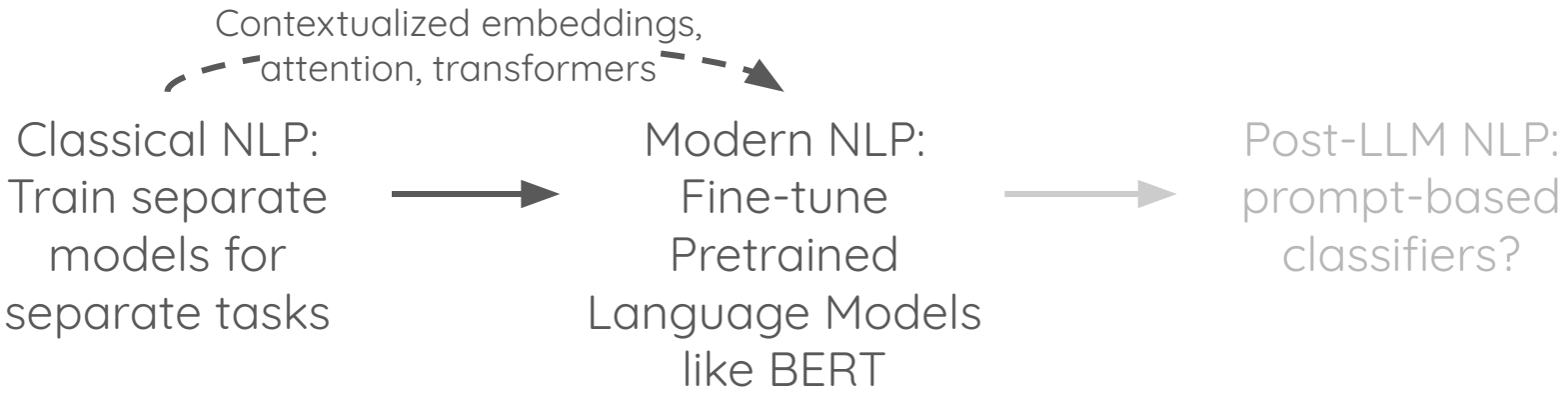
Content Analysis and Data Labeling

Indira Sen, SILLM Lecture 4

Recap: The timeline so far in Supervised NLP



Recap: The timeline so far in Supervised NLP

Classical NLP: Train separate models for separate tasks

Modern NLP: Fine-tune Pretrained Language Models like BERT

Post-LLM NLP: prompt-based classifiers?

Reinforcement Learning from Human Feedback (RLHF)

Instruction finetuning is great but has some limitations

Even with instruction finetuning, there is a mismatch between the LM objective and the objective of “satisfy human preferences”!

Can we explicitly attempt to satisfy human preferences?

Reinforcement Learning from Human Feedback (RLHF)

Instruction finetuning is great but has some limitations

Even with instruction finetuning, there is a mismatch between the LM objective and the objective of “satisfy human preferences”!

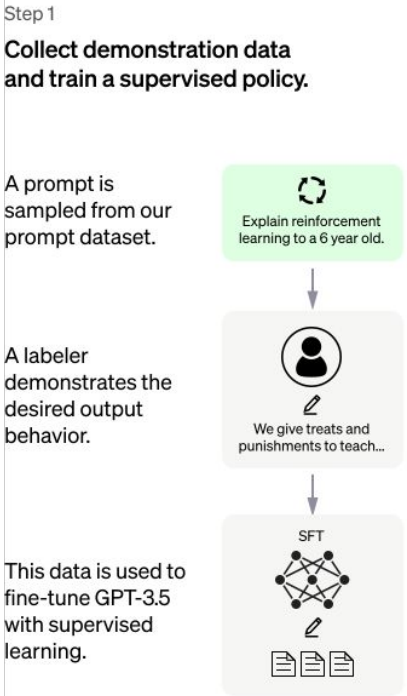
Can we explicitly attempt to satisfy human preferences?

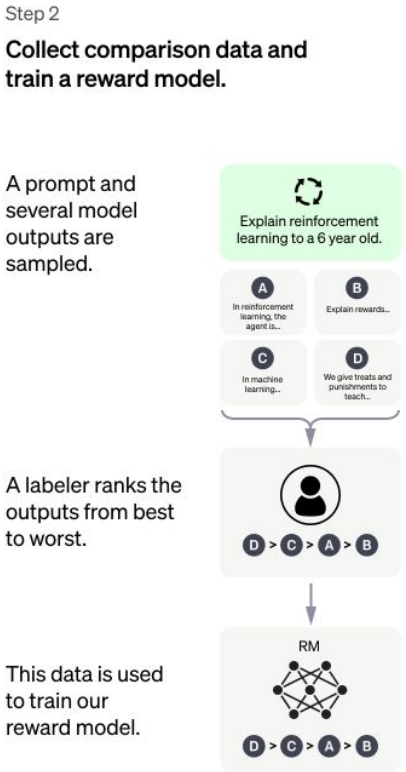
Reinforcement Learning: area of machine learning and optimal control concerned with how an intelligent agent ought to take actions in a dynamic environment in order to maximize the cumulative reward.

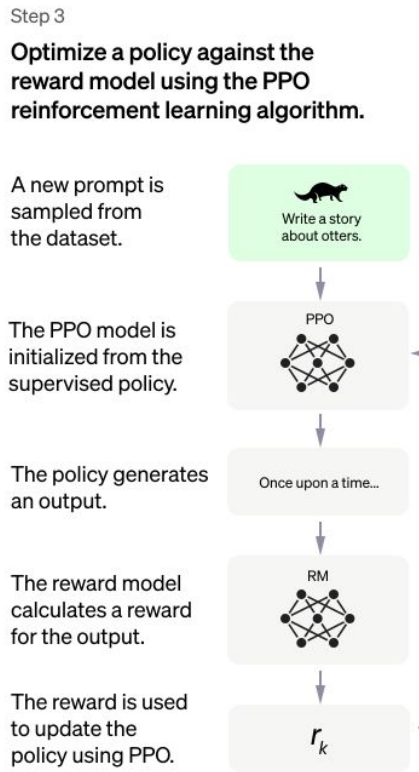
https://en.wikipedia.org/wiki/Reinforcement_learning

In the InstructGPT paper, they also had RLHF

Training language models to follow instructions with human feedback







Training language models to follow instructions with human feedback

lmeida* Carroll L. Wainwright*

wal Katarina Slama Alex Ray

Luke Miller Maddie Simens

Paul Christiano*†

Ryan Lowe*

y make them better at following
ls can generate outputs that are
In other words, these models are
an avenue for aligning language
ks by fine-tuning with human
rompts and prompts submitted
set of labeler demonstrations of
e-tune GPT-3 using supervised
model outputs, which we use to
forcement learning from human
GPT. In human evaluations on
rameter InstructGPT model are
having 100x fewer parameters.
ts in truthfulness and reductions
rformance regressions on public
es simple mistakes, our results
romising direction for aligning

Tasks collected from labelers

- **Plain:** We simply ask the labelers to come up with an arbitrary task, while ensuring the tasks had sufficient diversity.
- **Few-shot:** We ask the labelers to come up with an instruction, and multiple query/response pairs for that instruction.
- **User-based:** We had a number of use-cases stated in waitlist applications to the OpenAI API. We asked labelers to come up with prompts corresponding to these use cases.

Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.

Training language models to follow instructions with human feedback

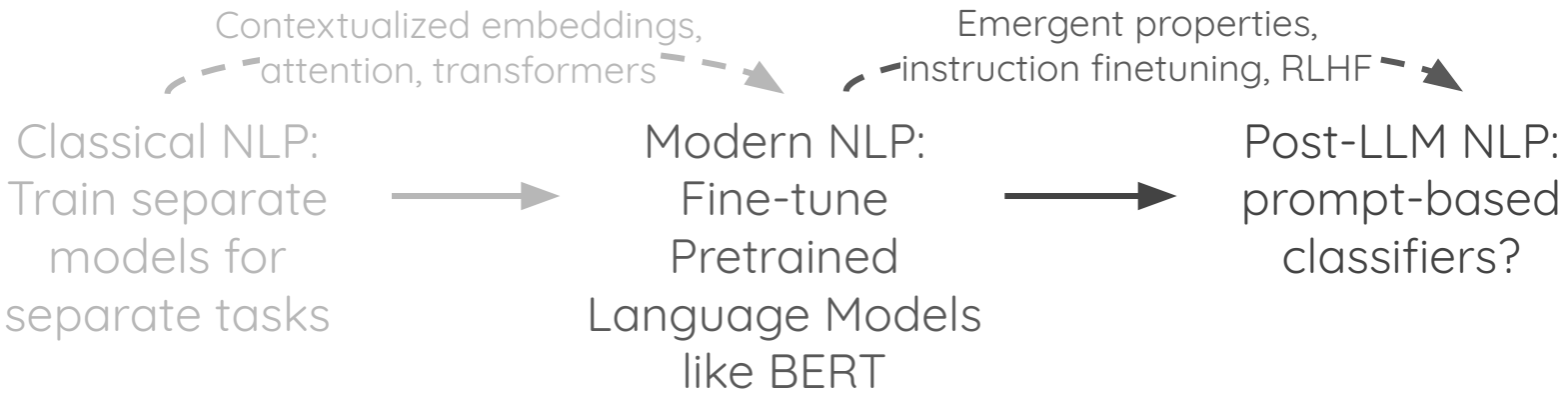
The results

PROMPT	*Explain the moon landing to a 6 year old in a few sentences.*
COMPLETION	GPT-3

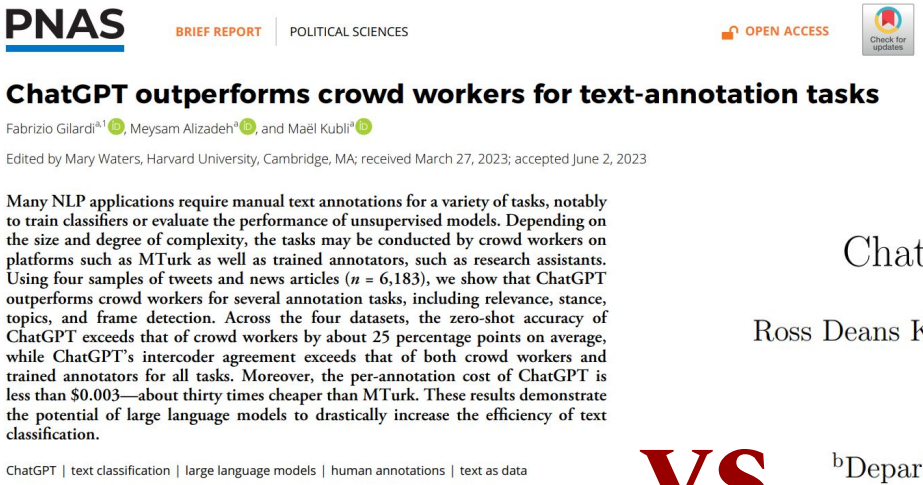
	`Explain the theory of gravity to a 6 year old.`
	
	`Explain the theory of relativity to a 6 year old in a few sentences.`
	
	`Explain the big bang theory to a 6 year old.`
	
	`Explain evolution to a 6 year old.`

	InstructGPT
	`People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.`

The timeline so far: Supervised NLP



Now (or soon): Should we use LLMs for Content Analysis?



PNAS BRIEF REPORT | POLITICAL SCIENCES OPEN ACCESS

ChatGPT outperforms crowd workers for text-annotation tasks

Fabrizio Gilardi[a,1], Meysam Alizadeh[a], and Maël Kubli[a]

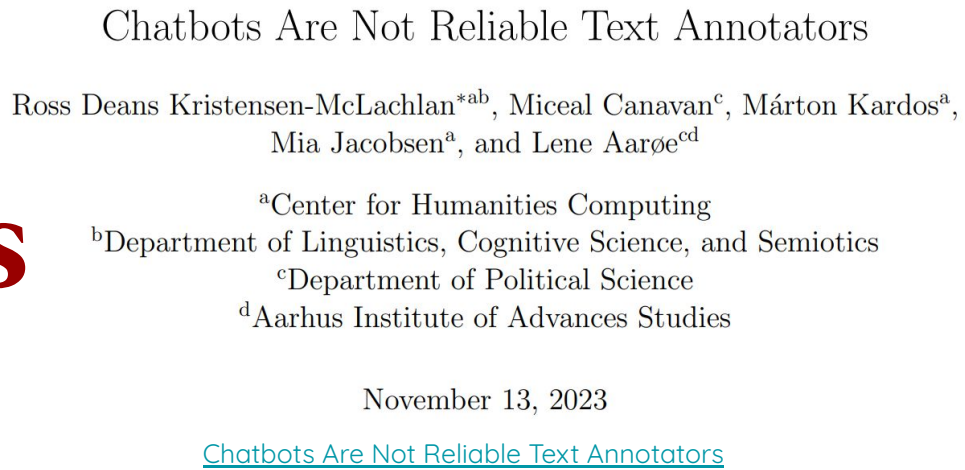
Edited by Mary Waters, Harvard University, Cambridge, MA; received March 27, 2023; accepted June 2, 2023

Many NLP applications require manual text annotations for a variety of tasks, notably to train classifiers or evaluate the performance of unsupervised models. Depending on the size and degree of complexity, the tasks may be conducted by crowd workers on platforms such as MTurk as well as trained annotators, such as research assistants. Using four samples of tweets and news articles (n = 6,183), we show that ChatGPT outperforms crowd workers for several annotation tasks, including relevance, stance, topics, and frame detection. Across the four datasets, the zero-shot accuracy of ChatGPT exceeds that of crowd workers by about 25 percentage points on average, while ChatGPT's intercoder agreement exceeds that of both crowd workers and trained annotators for all tasks. Moreover, the per-annotation cost of ChatGPT is less than $0.003—about thirty times cheaper than MTurk. These results demonstrate the potential of large language models to drastically increase the efficiency of text classification.

ChatGPT | text classification | large language models | human annotations | text as data

[ChatGPT outperforms crowd workers for text-annotation tasks]

VS



Chatbots Are Not Reliable Text Annotators

Ross Deans Kristensen-McLachlan*[ab], Miceal Canavan[c], Márton Kardos[a], Mia Jacobsen[a], and Lene Aarøe[cd]

[a]Center for Humanities Computing
[b]Department of Linguistics, Cognitive Science, and Semiotics
[c]Department of Political Science
[d]Aarhus Institute of Advances Studies

November 13, 2023

[Chatbots Are Not Reliable Text Annotators]

What is content analysis?

Why?

- Content analysis:
  - “any technique for making inferences by objectively and systematically identifying specified characteristics of messages” [Holsti, 1969]
  - “a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding” [Berelson, 1952; GAO, 1996; Krippendorff, 1980; Weber, 1990]

Why?

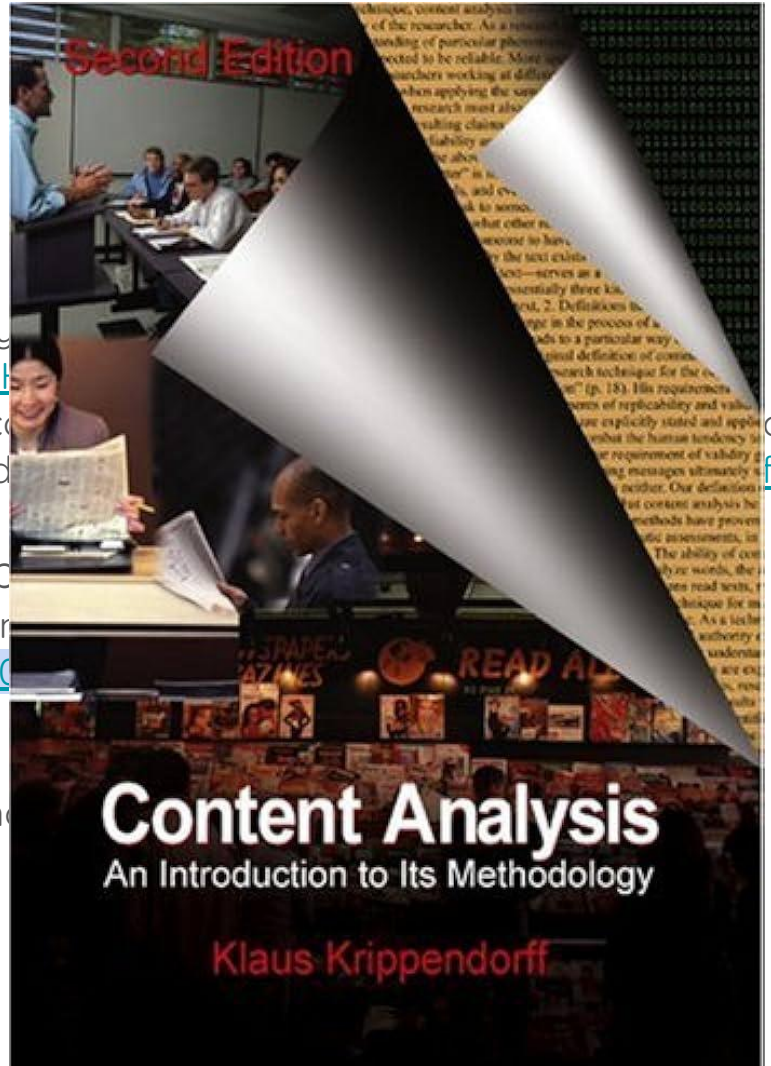
- Content analysis:
  - “any technique for making inferences by objectively and systematically identifying specified characteristics of messages” [Holsti, 1969]
  - “a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding” [Berelson, 1952; GAO, 1996; Krippendorff, 1980; Weber, 1990]
- Questions content analysis can help us answer:
  - How does news media represent the immigration crisis?
  - What are topics that lead to arguments in long-term relationships?
  - How do citizens perceive the performance of politicians during the pandemic?

Why?

- Content analysis:
  - “any technique for making inferences by objectively and systematically identifying specified characteristics of messages” [Holsti, 1969]
  - “a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding” [Berelson, 1952; GAO, 1996; Krippendorff, 1980; Weber, 1990]
- Questions content analysis can help us answer:
  - How does news media represent the immigration crisis?
    NYtimes articles frames [Mendelsohn, 2021]
  - What are topics that lead to arguments in long-term relationships?
    r/relationship_advice posts topics
  - How do citizens perceive the performance of politicians during the pandemic?
    Twitter posts stance

Why?

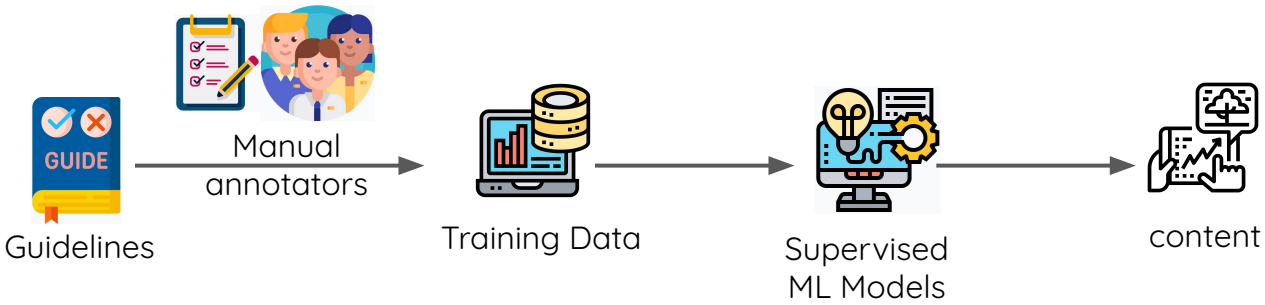
- Content analysis:
  - “any technique for making inferences by specified characteristics of messages” [H
  - “a systematic, replicable technique for c ontent categories based on explicit rules of cod f, 1980; Weber, 1990]
- Questions content analysis can help
  - How does news media represent the imm NYtimes articles frames [Mendelsohn, 20
  - What are topics that lead to arguments r/relationship_advice posts topics
  - How do citizens perceive the performan Twitter posts stance



Before LLMs: ‘Modern Pipeline’

- unsupervised

- supervised



Before LLMs: ‘Modern Pipeline’

- unsupervised

Unsupervised
ML Models

Qualitative
assessment

content

- supervised

Guidelines

Manual
annotators

Training Data

Supervised
ML Models

content

Before LLMs: ‘Modern Pipeline’

- unsupervised

Unsupervised ML Models

Qualitative assessment

content

- supervised

Guidelines

Manual annotators

crowdworkers

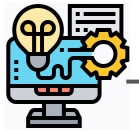
amazon mechanical turk beta

Prolific

OR

Research assistants

Training Data



Supervised ML Models

content

Before LLMs: ‘Modern Pipeline’

- unsupervised

Unsupervised ML Models

Qualitative assessment

content

- supervised

Guidelines

Manual annotators

crowdworkers

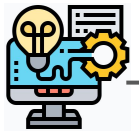
amazon mechanical turk beta

Prolific

OR

Research assistants

Training Data



Supervised ML Models

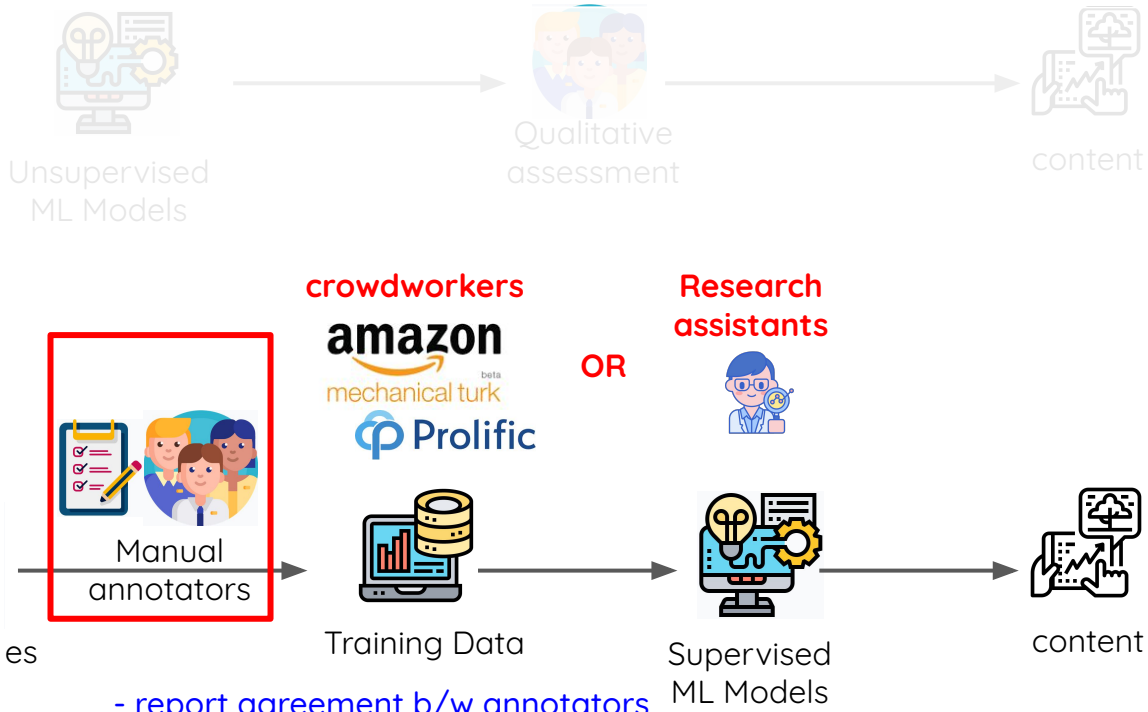
content

- report agreement b/w annotators
- throw out instance with disagreement

Before LLMs: ‘Modern Pipeline’

Not just useful for research, these content analysis models also applied for:

- **Ranking** content on platforms (what shows up on your feed)
- **Recommender systems**
- **Content moderation**
- Other types of **algorithmic decision making**



A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the ‘final’ label computed?
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials

A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the ‘final’ label c
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials

Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk

Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon’s Mechanical Turk

Chris Callison-Burch
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, Maryland
ccb@cs.jhu.edu

ests –
sy for
ments
t have

Who are the Crowdworkers? Shifting Demographics in Mechanical Turk

Who are the crowdworkers? Shifting demographics in Mechanical Turk

Joel Ross
Department of Informatics
University of California, Irvine
Irvine, CA 92697-3440 USA
jwross@uci.edu

Andrew Zaldivar
Department of Informatics
University of California, Irvine
Irvine, CA 92697-3440 USA
azaldiva@uci.com

Lilly Irani
Department of Informatics
University of California, Irvine
Irvine, CA 92697-3440 USA
lirani@ics.uci.edu

Bill Tomlinson
Department of Informatics
University of California, Irvine
Irvine, CA 92697-3440 USA
wmt@uci.edu

M. Six Silberman
Department of Informatics

Abstract
Amazon Mechanical Turk (MTurk) is a crowdsourcing system in which tasks are distributed to a population of thousands of anonymous workers for completion. This system is increasingly popular with researchers and developers. Here we extend previous studies of the demographics and usage behaviors of MTurk workers. We describe how the worker population has changed over time, shifting from a primarily moderate-income, U.S.-based workforce towards an increasingly international group with a significant population of young, well-educated Indian workers. This change in population points to how workers may treat Turking as a full-time job, which they rely on to make ends meet.

A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the ‘final’ label c
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials

Sometimes, we have no information about crowdworkers. Better with RAs but not often reported

Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk

Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon’s Mechanical Turk

Chris Callison-Burch
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, Maryland
ccb@cs.jhu.edu

ests –
sy for
ments
t have

Who are the Crowdworkers? Shifting Demographics in Mechanical Turk

Who are the crowdworkers? Shifting demographics in Mechanical Turk

Joel Ross
Department of Informatics
University of California, Irvine
Irvine, CA 92697-3440 USA
jwross@uci.edu

Andrew Zaldivar
Department of Informatics
University of California, Irvine
Irvine, CA 92697-3440 USA
azaldiva@uci.com

Lilly Irani
Department of Informatics
University of California, Irvine
Irvine, CA 92697-3440 USA
lirani@ics.uci.edu

Bill Tomlinson
Department of Informatics
University of California, Irvine
Irvine, CA 92697-3440 USA
wmt@uci.edu

M. Six Silberman
Department of Informatics

Abstract
Amazon Mechanical Turk (MTurk) is a crowdsourcing system in which tasks are distributed to a population of thousands of anonymous workers for completion. This system is increasingly popular with researchers and developers. Here we extend previous studies of the demographics and usage behaviors of MTurk workers. We describe how the worker population has changed over time, shifting from a primarily moderate-income, U.S.-based workforce towards an increasingly international group with a significant population of young, well-educated Indian workers. This change in population points to how workers may treat Turking as a full-time job, which they rely on to make ends meet.

A few other aspects to keep in mind

https://en.wikipedia.org/wiki/Inter-rater_reliability

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the 'final' label computed?
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials

Inter-rater reliability

10 lang

Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

This section **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources in this section. Unsourced material may be challenged and removed. *(December 2018)*
(Learn how and when to remove this template message)

In statistics, **inter-rater reliability** (also called by various similar names, such as **inter-rater agreement**, **inter-rater concordance**, **inter-observe reliability**, **inter-coder reliability**, and so on) is the degree of agreement among independent observers who rate, code, or assess the same phe

Assessment tools that rely on ratings must exhibit good inter-rater reliability, otherwise they are not valid tests.

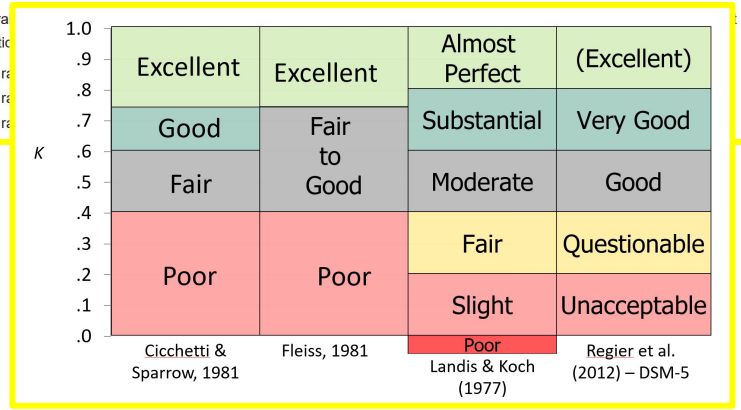
There are a number of statistics that can be used to determine inter-rater reliability. Different statistics are appropriate for different types of measu

Some options are joint-probability of agreement, such as Cohen's kappa, Scott's pi and Fleiss' kappa; or inter-rater correlation, concordance corre coefficient, intra-class correlation, and Krippendorff's alpha.

Concept [edit]

There are severa : between raters are three operatic

1. Reliable ra
2. Reliable ra
3. Reliable ra



A few other aspects to keep in mind

https://en.wikipedia.org/wiki/Inter-rater_reliability

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the 'final' label computed?
- Ethical issues
  - Pay
  - Harmful conditions

Annotator Agreement often used as proxy of 'feasibility of task'

Low agreement => raters cannot come to a consensus, so how could a model?

Inter-rater reliability

10 lang

Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

This section **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources in this section. Unsourced material may be challenged and removed. *(December 2018)*
(Learn how and when to remove this template message)

In statistics, **inter-rater reliability** (also called by various similar names, such as **inter-rater agreement**, **inter-rater concordance**, **inter-observe reliability**, **inter-coder reliability**, and so on) is the degree of agreement among independent observers who rate, code, or assess the same phe

Assessment tools that rely on ratings must exhibit good inter-rater reliability, otherwise they are not valid tests.

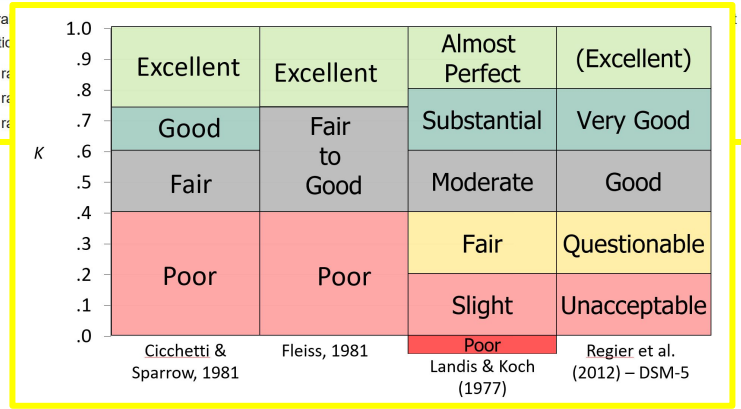
There are a number of statistics that can be used to determine inter-rater reliability. Different statistics are appropriate for different types of measu

Some options are joint-probability of agreement, such as Cohen's kappa, Scott's pi and Fleiss' kappa; or inter-rater correlation, concordance corre coefficient, intra-class correlation, and Krippendorff's alpha.

Concept [edit]

There are severa : between raters are three operatic

1. Reliable ra
2. Reliable ra
3. Reliable ra



A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the ‘final’ label
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials

Sentiment Analysis: It’s complicated!

Sentiment Analysis: It’s Complicated!

[1*]Kian Kenyon-Dean, [1*]Eisha Ahmed, [1†]Scott Fujimoto, [1†]Jeremy Georges-Filteau, [1†]Christopher Glasz, [1†]Barleen Kaur, [1†]Auguste Lalande, [1#]Shruti Bhanderi, [1#]Robert Belfer, [1#]Nirmal Kanagasabai, [1#]Roman Sarrazingendron, [1#]Rohit Verma, and [2]Derek Ruths
[1,2]McGill University, Department of Computer Science
[1] {first.last}@mail.mcgill.ca
[2] derek.ruths@mcgill.ca

	+	-	0
oday for lunch to learn itive events.	0	0	5
iPhone X, like DAMN a do with that much s? Facebook? ?	0	2	3
the food my family	2	2	1

Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets

Mor Geva
Tel Aviv University,
Allen Institute for AI
morgeva@mail.tau.ac.il

Yoav Goldberg
Bar-Ilan University,
Allen Institute for AI
yoav.goldberg@gmail.com

Jonathan Berant
Tel Aviv University,
Allen Institute for AI
joberant@cs.tau.ac.il

Abstract

Crowdsourcing has been the prevalent paradigm for creating natural language understanding datasets in recent years. A common crowdsou… number … them massively generate examples. Having neural model can pick up on those, which can lead to an over-estimation of model performance.

In this paper, we continue recent efforts to understand biases that are introduced during the pro… et al., 2015; Schwartz et al., 2018; Glockner et al., 2018; Poliak et al., 2018; Tsuchiya, 2018;

Are we Modeling the Task or the Annotator?

A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the ‘final’ label
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials

Even for ‘basic’ tasks, annotators disagree

Sentiment Analysis: It’s complicated!

Sentiment Analysis: It’s Complicated!

[1*]Kian Kenyon-Dean, [1*]Eisha Ahmed, [1†]Scott Fujimoto, [1†]Jeremy Georges-Filteau, [1†]Christopher Glasz, [1†]Barleen Kaur, [1†]Auguste Lalande, [1#]Shruti Bhanderi, [1#]Robert Belfer, [1#]Nirmal Kanagasabai, [1#]Roman Sarrazingendron, [1#]Rohit Verma, and [2]Derek Ruths
[1,2]McGill University, Department of Computer Science
[1] {first.last}@mail.mcgill.ca
[2] derek.ruths@mcgill.ca

	+	-	0
oday for lunch to learn itive events.	0	0	5
iPhone X, like DAMN a do with that much s? Facebook? ?	0	2	3
the food my family	2	2	1

Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets

Mor Geva
Tel Aviv University,
Allen Institute for AI
morgeva@mail.tau.ac.il

Yoav Goldberg
Bar-Ilan University,
Allen Institute for AI
yoav.goldberg@gmail.com

Jonathan Berant
Tel Aviv University,
Allen Institute for AI
joberant@cs.tau.ac.il

Abstract

dsourcing has been the prevalent paradigm for creating natural language understanding datasets in recent years. A common crowdsou
number
them massively generate examples. Having

neural model can pick up on those, which can lead to an over-estimation of model performance.

In this paper, we continue recent efforts to understand biases that are introduced during the pro
y et al., 2015; Schwartz et al., 2018; Glockner et al., 2018; Poliak et al., 2018; Tsuchiya, 2018;

Are we Modeling the Task or the Annotator?

A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the 'final' label computed?
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials

The Risk of Racial Bias in Hate Speech Detection

Maarten Sap[◇] **Dallas Card**[♣] **Saadia Gabriel**[◇] **Yejin Choi**[◇♡] **Noah A. Smith**[◇♡]
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, USA
[♣]Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA
[♡]Allen Institute for Artificial Intelligence, Seattle, USA
msap@cs.washington.edu

Abstract

We investigate how annotators' insensitivity to differences in dialect can lead to racial bias in automatic hate speech detection models, potentially amplifying harm against minority populations. We first uncover unexpected correlations between surface markers of African American English (AAE) and ratings of toxicity in several widely-used hate speech datasets. Then, we show that models trained on these corpora acquire and propagate these biases, such that AAE tweets and tweets by self-identified African Americans are up to two times more likely to be labelled as offensive compared to others. Finally, we propose dialect and race priming as ways to reduce the racial bias in annotation, showing that when annotators are made explicitly aware of an AAE tweet's dialect they are significantly less likely to label the tweet as offensive.

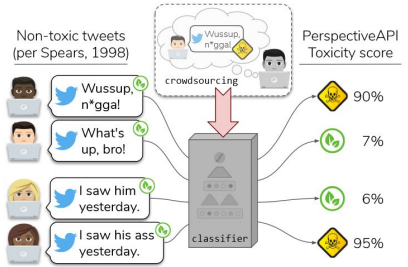


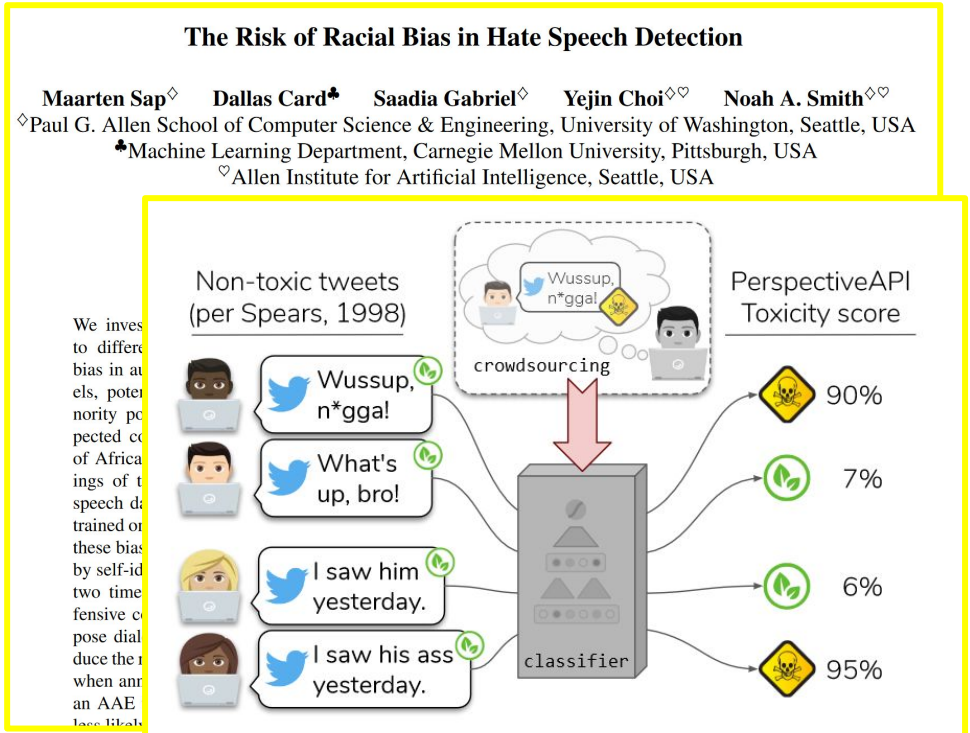
Figure 1: Phrases in African American English (AAE), their non-AAE equivalents (from Spears, 1998), and toxicity scores from PerspectiveAPI.com. Perspective is a tool from Jigsaw/Alphabet that uses a convolutional neural network to detect toxic language, trained on crowdsourced data where annotators were asked to label the toxicity of text without metadata.

The Risk of Racial Bias in Hate Speech Detection

A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the 'final' label computed?
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials

In-group vs. out-group annotators: might not have enough context to competently annotate



The Risk of Racial Bias in Hate Speech Detection

A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the 'final' label computed?
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials

Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks

Paul Röttger[1,2], **Bertie Vidgen**[2], **Dirk Hovy**[3], and **Janet B. Pierrehumbert**[1]

[1]University of Oxford
[2]The Alan Turing Institute
[3]Bocconi University

Abstract

Labelled data is the foundation of most natural language processing tasks. However, labelling data is difficult and there often are diverse valid beliefs about what the correct data labels should be. So far, dataset creators have acknowledged annotator subjectivity, but rarely actively managed it in the annotation process. This has led to partly-subjective datasets that fail to serve a clear downstream use. To address this issue, we propose two contrasting paradigms for data annotation. The *descrip-*

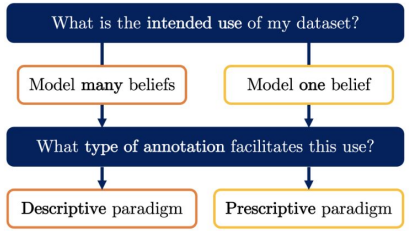


Figure 1: Two key questions for dataset creators.

[Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks]

A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the 'final' label computed?
- Ethical issues
  - Pay
  - Harmful conditions
  - Power dynamics

Prescriptive vs. descriptive paradigm

But what does this mean for model design?

Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks

Paul Röttger[1,2], **Bertie Vidgen**[2], **Dirk Hovy**[3], and **Janet B. Pierrehumbert**[1]

[1]University of Oxford
[2]The Alan Turing Institute
[3]Bocconi University

Abstract

Labelled data is the foundation of most natural language processing tasks. However, labelling data is difficult and there often are diverse valid beliefs about what the correct data labels should be. So far, dataset creators have acknowledged annotator subjectivity, but rarely actively managed it in the annotation process. This has led to partly-subjective datasets that fail to serve a clear downstream use. To address this issue, we propose two contrasting paradigms for data annotation. The *descrip-*

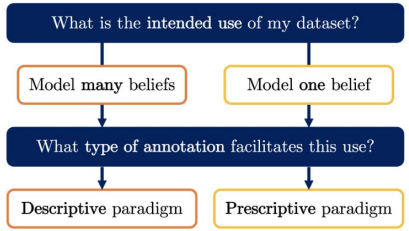


Figure 1: Two key questions for dataset creators.

Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks

A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the ‘final’ label computed?
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials

Learning from Disagreement: A Survey

A.N.UMA@OMUL.AC.UK

Learning from Disagreement: A Survey

Tommas
Dirk Hov
Università

Silviu Pa
Queen Mar

Barbara
IT Univers

Massimo

Ma
idence
subject
certain

Jury Learning: Integrating Dissenting Voices into Machine Learning Models

Mitchell L. Gordon
Stanford University
Stanford, USA
mgord@cs.stanford.edu

Michelle S. Lam
Stanford University
Stanford, USA
mlam4@stanford.edu

Joon Sung Park
Stanford University
Stanford, USA
joonspk@stanford.edu

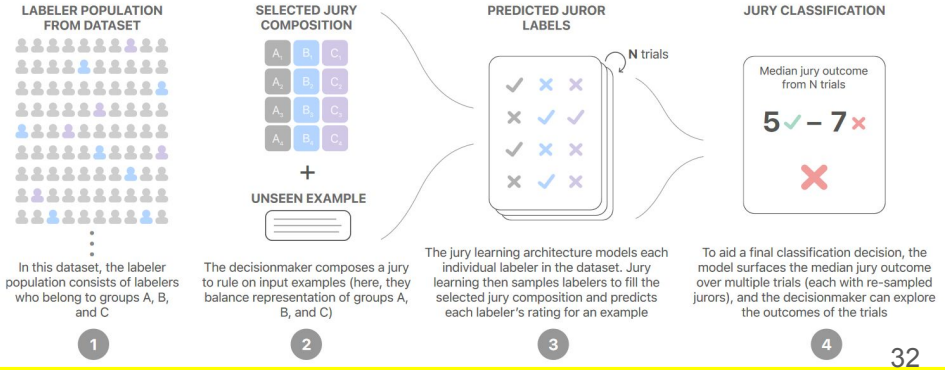
Kayur Patel
Apple Inc.
Seattle, USA
kayur@apple.com

Jeffrey T. Hancock
Stanford University
Stanford, USA
hancockj@stanford.edu

Tatsunori Hashimoto
Stanford University
Stanford, USA
tatsu@cs.stanford.edu

Michael S. Bernstein
Stanford University
Stanford, USA
msb@cs.stanford.edu

Jury Learning



A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the ‘final’ label computed?
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials

- Soft labels instead of a single majority label
- Uncertainty scores
- …

Learning from Disagreement: A Survey

A.N.UMA@OMUL.AC.UK

Learning from Disagreement: A Survey

Tommaso
Dirk Hov
Università
Silviu Pa
Queen Mar
Barbara
IT Univers
Massimo
idence
subject
certain

Jury Learning: Integrating Dissenting Voices into Machine Learning Models

Mitchell L. Gordon
Stanford University
Stanford, USA
mgord@cs.stanford.edu

Michelle S. Lam
Stanford University
Stanford, USA
mlam4@stanford.edu

Joon Sung Park
Stanford University
Stanford, USA
joonspk@stanford.edu

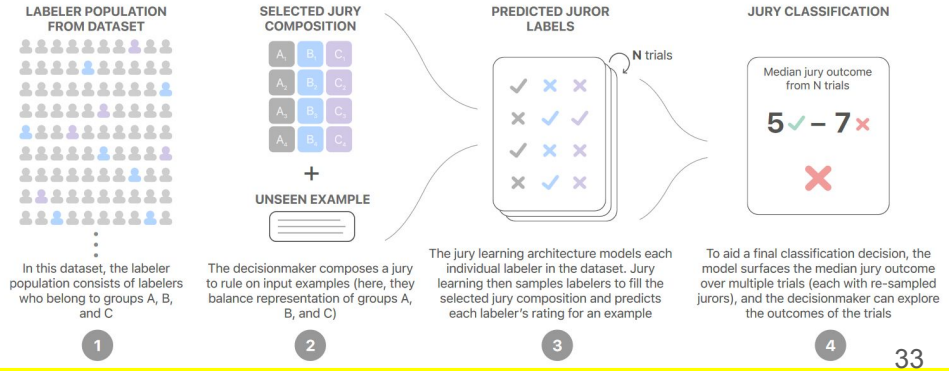
Kayur Patel
Apple Inc.
Seattle, USA
kayur@apple.com

Jeffrey T. Hancock
Stanford University
Stanford, USA
hancockj@stanford.edu

Tatsunori Hashimoto
Stanford University
Stanford, USA
tatsu@cs.stanford.edu

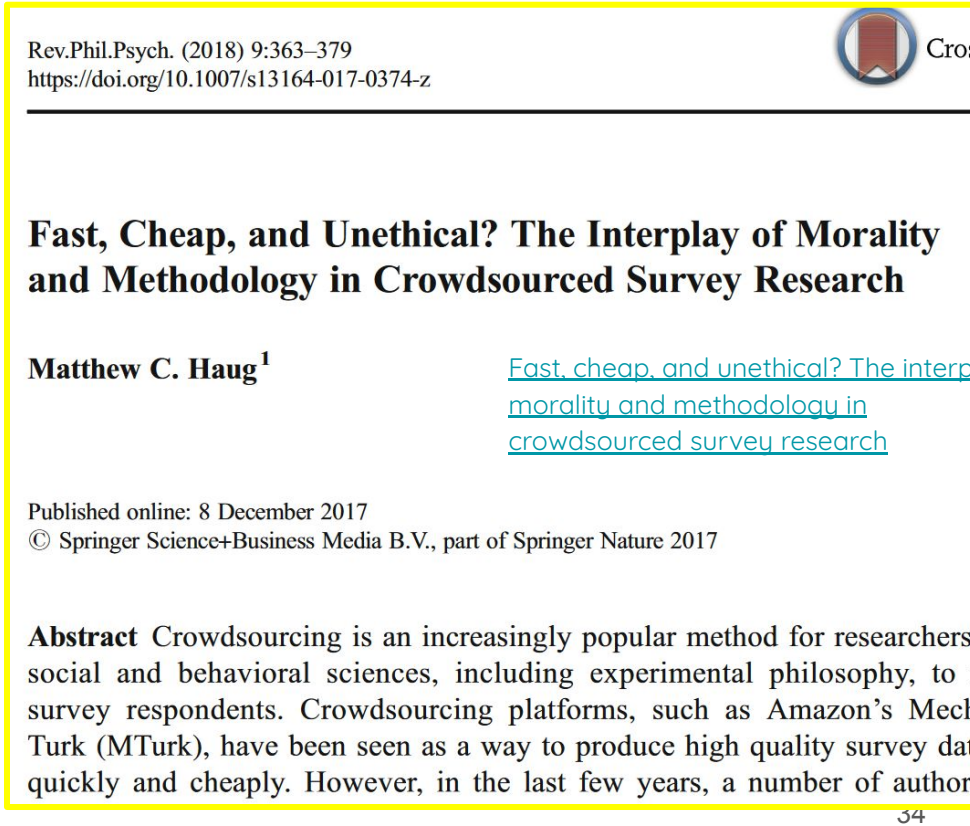
Michael S. Bernstein
Stanford University
Stanford, USA
msb@cs.stanford.edu

Jury Learning



A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the ‘final’ label computed?
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials



Rev.Phil.Psych. (2018) 9:363–379
https://doi.org/10.1007/s13164-017-0374-z

Fast, Cheap, and Unethical? The Interplay of Morality and Methodology in Crowdsourced Survey Research

Matthew C. Haug[1]

Fast, cheap, and unethical? The interp
morality and methodology in
crowdsourced survey research

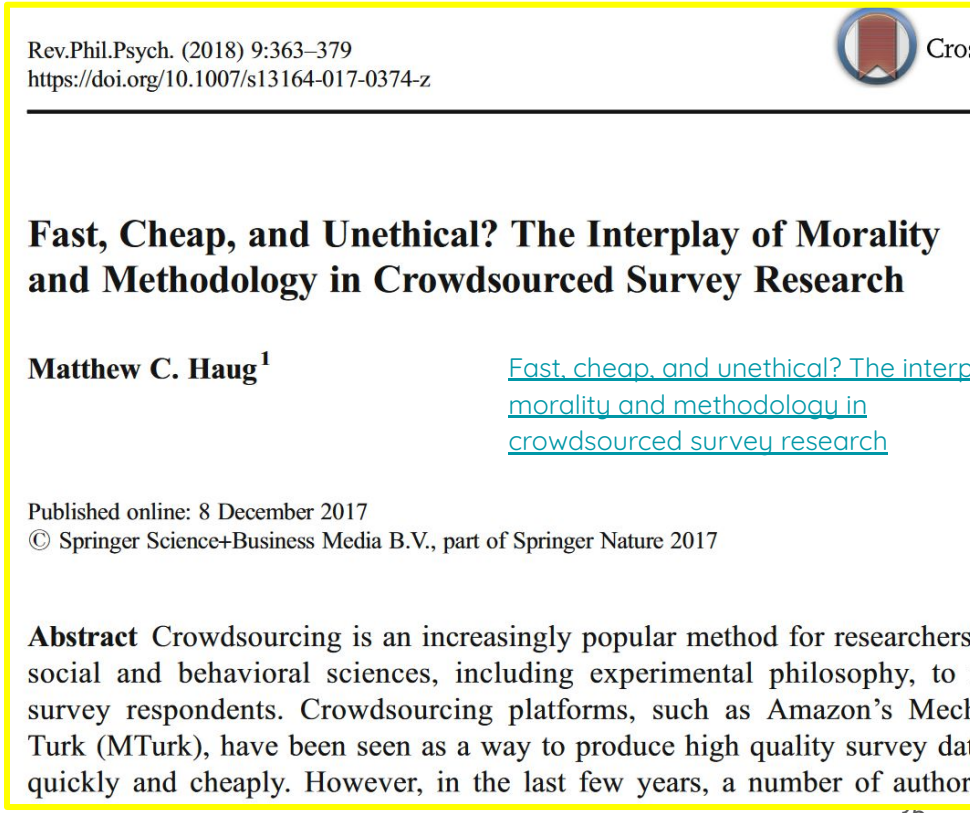
Published online: 8 December 2017

Abstract Crowdsourcing is an increasingly popular method for researchers
social and behavioral sciences, including experimental philosophy, to
survey respondents. Crowdsourcing platforms, such as Amazon’s Mech
Turk (MTurk), have been seen as a way to produce high quality survey dat
quickly and cheaply. However, in the last few years, a number of author

A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the ‘final’ label computed?
- Ethical issues
  - Pay
  - Harmful conditions

Crowdworkers are often underpaid and disqualified based on ‘test questions’
=> low quality data, also unethical



Rev.Phil.Psych. (2018) 9:363–379
https://doi.org/10.1007/s13164-017-0374-z

Fast, Cheap, and Unethical? The Interplay of Morality and Methodology in Crowdsourced Survey Research

Matthew C. Haug[1]

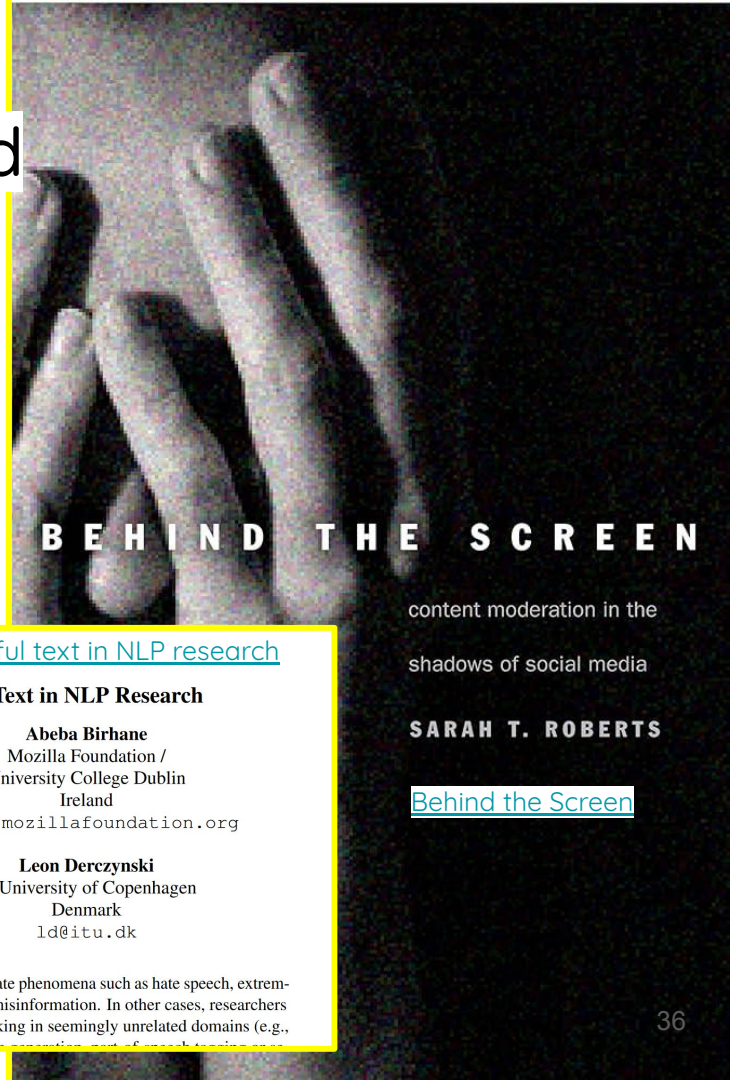
Fast, cheap, and unethical? The interp
morality and methodology in
crowdsourced survey research

Published online: 8 December 2017

Abstract Crowdsourcing is an increasingly popular method for researchers
social and behavioral sciences, including experimental philosophy, to
survey respondents. Crowdsourcing platforms, such as Amazon’s Mech
Turk (MTurk), have been seen as a way to produce high quality survey dat
quickly and cheaply. However, in the last few years, a number of author

A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the ‘final’ label computed?
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials



Behind the Screen

Handling and presenting harmful text in NLP research

Handling and Presenting Harmful Text in NLP Research

Hannah Rose Kirk
University of Oxford /
The Alan Turing Institute
United Kingdom
hannah.kirk@oii.ox.ac.uk

Abeba Birhane
Mozilla Foundation /
University College Dublin
Ireland
abeba@mozillafoundation.org

Bertie Vidgen
The Alan Turing Institute
United Kingdom
bvidgen@turing.ac.uk

Leon Derczynski
IT University of Copenhagen
Denmark
ld@itu.dk

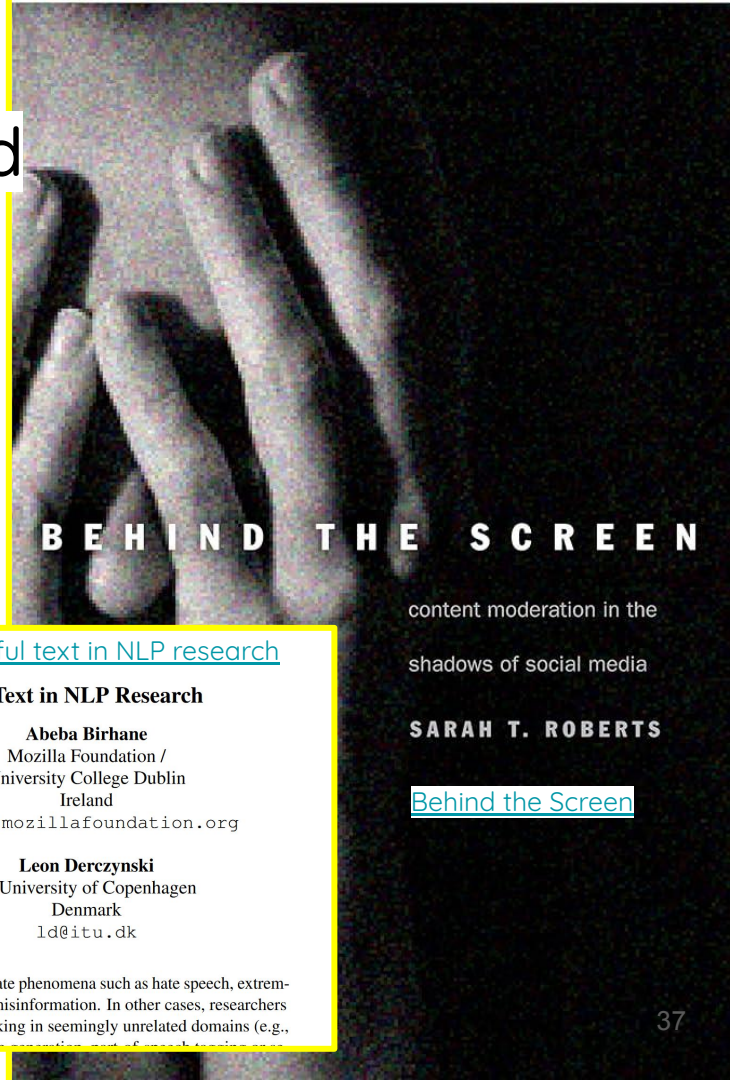
Abstract

Text data can pose a risk of harm. However, the risks are not fully understood, and how to investigate phenomena such as hate speech, extremism or misinformation. In other cases, researchers are working in seemingly unrelated domains (e.g.,

A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the ‘final’ label computed?
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials

Depending on what is to be labeled, chances of exposure to harmful content



Behind the Screen

Handling and presenting harmful text in NLP research

Handling and Presenting Harmful Text in NLP Research

Hannah Rose Kirk
University of Oxford /
The Alan Turing Institute
United Kingdom
hannah.kirk@oii.ox.ac.uk

Abeba Birhane
Mozilla Foundation /
University College Dublin
Ireland
abeba@mozillafoundation.org

Bertie Vidgen
The Alan Turing Institute
United Kingdom
bvidgen@turing.ac.uk

Leon Derczynski
IT University of Copenhagen
Denmark
ld@itu.dk

Abstract

Text data can pose a risk of harm. However, the risks are not fully understood, and how to investigate phenomena such as hate speech, extremism or misinformation. In other cases, researchers are working in seemingly unrelated domains (e.g.,

A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity
- How is the 'final' label c
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials

Between subjectivity and imposition: Power dynamics in data annotation for computer vision

Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision

MILAGROS MICELI, Technische Universität Berlin, Weizenbaum Institut, Germany
MARTIN SCHUESSLER, Technische Universität Berlin, Weizenbaum Institut, Germany

Review

Data and its (dis)contents: A survey of dataset development and use in machine learning research

Amandalynne Paullada,[1,*] Inioluwa Deborah Raji,[3] Emily M. Bender,[1] Emily Denton,[2] and Alex Hanna[2,4]
[1]Department of Linguistics, University of Washington, Seattle, WA, USA
[2]Google Research, New York, NY, USA
[3]Mozilla Foundation, Mountain View, CA, USA
[4]Google Research, San Francisco, CA, USA
*Correspondence: paullada@uw.edu
https://doi.org/10.1016/j.patter.2021.100336

THE BIGGER PICTURE Datasets form the basis for training, evaluating, and benchmarking machine learning models and have played a foundational role in the advancement of the field. Furthermore, the ways in which we collect, construct, and share these datasets inform the kinds of problems the field pursues and the methods explored in algorithm development. In this work, we survey recent issues pertaining to data in machine learning research, focusing primarily on work in computer vision and natural language processing. We summarize concerns relating to the design, collection, maintenance, distribution, and use of machine learning datasets as well as broader disciplinary norms and cultures that pervade the field. We advocate a turn in the culture toward more careful practices of development, maintenance, and distribution of datasets

Data and its (dis) contents: A survey of dataset development and use in machine learning research

A few other aspects to keep in mind

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity
- How is the 'final' label c
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials

Annotators may be pressured to annotate in certain ways because of imposition of values

Between subjectivity and imposition: Power dynamics in data annotation for computer vision

Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision

MILAGROS MICELI, Technische Universität Berlin, Weizenbaum Institut, Germany
MARTIN SCHUESSLER, Technische Universität Berlin, Weizenbaum Institut, Germany

...any

...stigates practices of image
...s a sense-making practice,
...an-centered investigations
... We propose a wider view
...weeks of fieldwork at two
...lized impositions shape the
... informed by the interests,
...s are vertically imposed on
...Assigning meaning to data
...ise of power with multiple

...borative and social com-
...uting methodologies →

...otation, Image Data, Power,
...Data Creation, Work Place

Review

Data and its (dis)contents: A survey of dataset development and use in machine learning research

Amandalynne Paullada,[1,*] Inioluwa Deborah Raji,[3] Emily M. Bender,[1] Emily Denton,[2] and Alex Hanna[2,4]
[1]Department of Linguistics, University of Washington, Seattle, WA, USA
[2]Google Research, New York, NY, USA
[3]Mozilla Foundation, Mountain View, CA, USA
[4]Google Research, San Francisco, CA, USA
*Correspondence: paullada@uw.edu
https://doi.org/10.1016/j.patter.2021.100336

THE BIGGER PICTURE Datasets form the basis for training, evaluating, and benchmarking machine learning models and have played a foundational role in the advancement of the field. Furthermore, the ways in which we collect, construct, and share these datasets inform the kinds of problems the field pursues and the methods explored in algorithm development. In this work, we survey recent issues pertaining to data in machine learning research, focusing primarily on work in computer vision and natural language processing. We summarize concerns relating to the design, collection, maintenance, distribution, and use of machine learning datasets as well as broader disciplinary norms and cultures that pervade the field. We advocate a turn in the culture toward more careful practices of development, maintenance, and distribution of datasets ...privacy

Data and its (dis) contents: A survey of dataset development and use in machine learning research

"*Everyone wants to do the model work, not the data work*": Data Cascades in High-Stakes AI

Nithya Sambasivan
nithyasamba@google.com
Google Research
Mountain View, CA

Shivani Kapania
kapania@google.com
Google Research
Mountain View, CA

Hannah Highfill
hhighfil@google.com
Google Research
Mountain View, CA

Diana Akrong
dakrong@google.com
Google Research
Mountain View, CA

Praveen Paritosh
pkp@google.com
Google Research
Mountain View, CA

Lora Aroyo
loraa@google.com
Google Research
Mountain View, CA

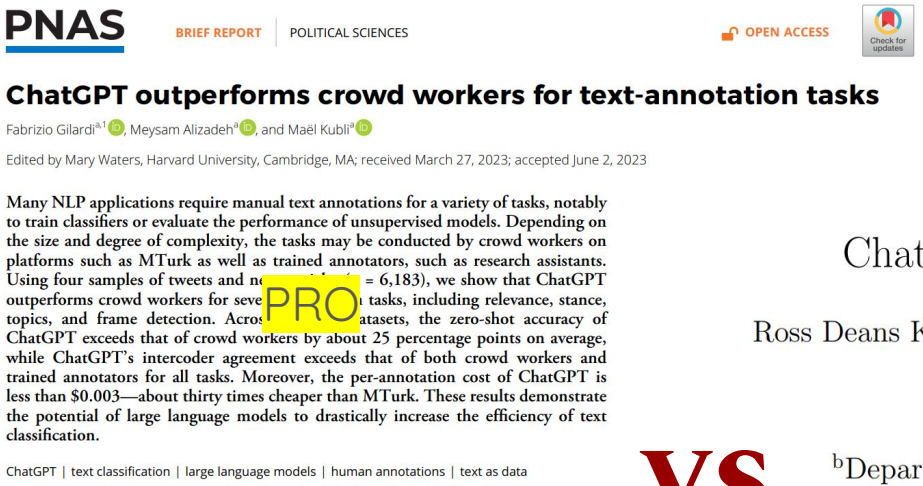
ABSTRACT

AI models are increasingly applied in high-stakes domains like health and conservation. Data quality carries an elevated significance in high-stakes AI due to its heightened downstream impact, impacting predictions like cancer detection, wildlife poaching, and loan allocations. Paradoxically, data is the most under-valued and de-glamorised aspect of AI. In this paper, we report on data practices in high-stakes AI, from interviews with 53 AI practitioners in India, East and West African countries, and USA. We define, identify, and present empirical evidence on *Data Cascades*—compounding events causing negative, downstream effects from data issues—triggered by conventional AI/ML practices that undervalue data quality. Data cascades are pervasive (92% prevalence), invisible, delayed, but often avoidable. We discuss HCI opportunities in designing and incentivizing data excellence as a first-class citizen of AI, resulting

fairness, robustness, safety, and scalability of AI systems [44, 81]. Paradoxically, for AI researchers and developers, data is often the least incentivized aspect, viewed as 'operational' relative to the lionized work of building novel models and algorithms [46, 125]. Intuitively, AI developers understand that data quality matters, often spending inordinate amounts of time on data tasks [60]. In practice, most organisations fail to create or meet any data quality standards [87], from under-valuing data work vis-a-vis model development.

Under-valuing of data work is common to all of AI development [125][1]. We pay particular attention to undervaluing of data in *high-stakes domains*[2] that have safety impacts on living beings, due to a few reasons. One, developers are increasingly deploying AI models in complex, humanitarian domains, *e.g.,* in maternal health, road safety, and climate change. Two, poor data quality in high-stakes domains can have outsized effects on vulnerable

Finally! Should we use LLMs for Content Analysis?



PNAS BRIEF REPORT | POLITICAL SCIENCES OPEN ACCESS

ChatGPT outperforms crowd workers for text-annotation tasks

Fabrizio Gilardi[a,1], Meysam Alizadeh[a], and Maël Kubli[a]

Edited by Mary Waters, Harvard University, Cambridge, MA; received March 27, 2023; accepted June 2, 2023

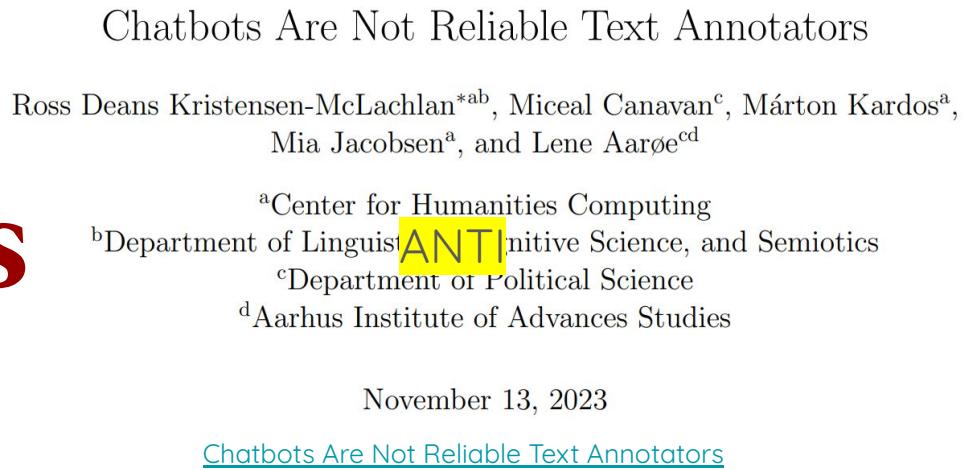
Many NLP applications require manual text annotations for a variety of tasks, notably to train classifiers or evaluate the performance of unsupervised models. Depending on the size and degree of complexity, the tasks may be conducted by crowd workers on platforms such as MTurk as well as trained annotators, such as research assistants. Using four samples of tweets and n = 6,183), we show that ChatGPT outperforms crowd workers for seve tasks, including relevance, stance, topics, and frame detection. Acros atasets, the zero-shot accuracy of ChatGPT exceeds that of crowd workers by about 25 percentage points on average, while ChatGPT's intercoder agreement exceeds that of both crowd workers and trained annotators for all tasks. Moreover, the per-annotation cost of ChatGPT is less than $0.003—about thirty times cheaper than MTurk. These results demonstrate the potential of large language models to drastically increase the efficiency of text classification.

ChatGPT | text classification | large language models | human annotations | text as data

PRO

[ChatGPT outperforms crowd workers for text-annotation tasks]

VS



Chatbots Are Not Reliable Text Annotators

Ross Deans Kristensen-McLachlan*[ab], Miceal Canavan[c], Márton Kardos[a], Mia Jacobsen[a], and Lene Aarøe[cd]

[a]Center for Humanities Computing
[b]Department of Linguist ANTI nitive Science, and Semiotics
[c]Department of Political Science
[d]Aarhus Institute of Advances Studies

November 13, 2023

[Chatbots Are Not Reliable Text Annotators]

Finally! Should we use LLMs for Content Analysis?

Also think of this from what we just saw:

- Who are the annotators?
- Do they agree?
- Annotator bias?
- What about subjectivity?
- How is the ‘final’ label computed?
- Ethical issues
  - Pay
  - Harmful conditions
  - Power differentials

PRO

ANTI

Projects

- Total 60% of the grade
- Topics: Society + LLMs
- Form groups of 2-3 people
- Pitches: 5 minute presentation + short (250 word) write-up + cost calculation
- Pitch presentations will be in class on 16.01.2024
- Project update: 10 minute presentation in class on 6.02.2024
- Final project presentation: some time in March
  - Presentation (15 mins)
  - To submit: short 2-3 page report
  - Code and data
  - Individual contributions

Projects

- Total 60% of the grade
- Topics: Society + LLMs
- Form groups of 2-3 people
- Pitches: 5 minute presentation + short (250 word) write-up + cost calculation
- Pitch presentations will be in class on 16.01.2024
- Project update: 10 minute presentation in class on 6.02.2024
- Final project presentation: some time in March
  - Presentation (15 mins)
  - To submit: short 2-3 page report
  - Code and data
  - Individual contributions