

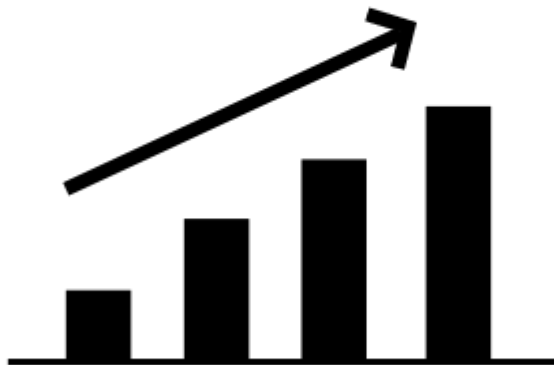
Responsible Text Mining

Dong Nguyen
29th of July, 2021



Utrecht University

Advances in NLP



performance



Dual Use

Dual use: Text generation



Generate novels,
poems, summaries

AI systems might be used for both beneficial and harmful purposes

Dual use: Text generation



Generate novels,
poems, summaries

Disinformation



AI systems might be used for both beneficial and harmful purposes

Generated text

<https://talktotransformer.com/>

Talk to Transformer

See how a modern neural network completes your text. Type a custom snippet or try one of the examples. This is a limited demo of [InferKit](#).

Custom prompt ▼

This week Utrecht University hosts a summer school on text mining. The summer school

Generate Another

Completion

This week Utrecht University hosts a summer school on text mining. The summer school covers five days of theoretical sessions and two days of practical sessions. The main goals of this new initiative are to give the participants the opportunity to learn about the fundamentals of text mining, design various applications and case studies, and to apply their newly acquired knowledge to real-world data. By sharing their research and approaches at the workshop, researchers hope to create a network that will continue to grow.

To fill this interdisciplinary space, both experimental

Generated fake news

<https://talktotransformer.com/>

The Dutch government plans to invest in the city in coming months to increase connectivity between the region and Europe using high speed railway services.

"I know it will have been a tough fight, but to bring Dutch cities into Europe it was never going to be a simple thing, however, we will come through it. We will have to. We have done it before and done it again" - Michael J.K. Westerbeke, city mayor

A new intercity service between Amsterdam and Groningen began service on 20 April.

Transporting a heavy load was never a problem for the Dutch when the railway reached capacity because the city had such a large area to transport in, the city's mayor said.

"It is also a reason for us to keep going. After all, this is the city of Amsterdam," Michael J.K. Westerbeke said at the first city council meeting

Quiz:

<http://quiz.newsyoucantuse.com/>

Fake Article  <https://grover.allenai.org/>

nytimes.com

Why Bitcoin is a great investment

June 6, 2019 - Paul Krugman

As most of my readers know, I'm an optimist.

This belief applies across my life, and to various investments as well. So I am intrigued by the success of cryptocurrencies, such as Bitcoin and Ethereum. The competition they are putting up against the gold standard looks insane, as Bitcoin goes off to the races.

There's no way to fully understand what's going on in the crypto world — and I am not even sure anyone could if you tried to. Still, I can tell you that Bitcoin's recent surge is really an opportunity to buy long-term real assets.

Cryptocurrencies are new and don't even have a useful underlying technology. They will probably fail, probably sooner than later. If people forget about them quickly, it is likely to be because the underlying technology will finally mature and win out. We don't even know whether that will happen in a generation or maybe a century, but it's still possible it might.

Dual use: Should I build this sytem?

Predicting Depression via Social Media

Munmun De Choudhury

Michael Gamon

Scott Counts

Eric Horvitz

Microsoft Research, Redmond WA 98052

{munmund, mgamon, counts, horvitz}@microsoft.com

“We explore the potential to use social media to detect and diagnose major depressive disorder in individuals.”

How can such a system be used
for a beneficial purpose?
How can such a system be used
for a harmful purpose?

(3 min)

Nice! But are we really
measuring what we
intend to measure?



What can go
wrong?



7 x 2

Are horses clever?

If the eighth day of the month
comes on a Tuesday, what is the
date of the following Friday?

Clever Hans

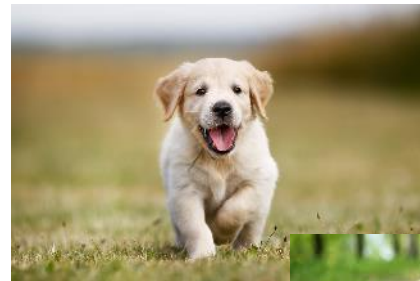
Claimed to have
performed
arithmetic and
other intellectual
tasks.



Wolf or dog?



Can the system really distinguish between dogs and wolves?



Sentiment analysis



8/10

Sci-fi perfection. A truly mesmerizing film.

I'm nearly at a loss for words. Just when you thought Christopher Nolan couldn't follow up to "The Dark Knight", he does it again, delivering another masterpiece, one with so much power and rich themes that has been lost from the box office for several years. Questioning illusions vs reality usually makes the film weird, but Nolan grips your attention like an iron claw that you just can't help watching and wondering what will happen next. That is a real powerful skill a director has. No wonder Warner Bros. put their trust in him, he is THAT good of a director, and over-hyping a Christopher Nolan film, no matter what the film is about, is always an understatement instead of an overestimate like MANY films before.

Models can be right for the wrong reasons 😞

Is our model actually measuring what we think it is measuring?

What can go wrong?

NLP models can take shortcuts

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are at least three people on a loading dock.
Neutral	A woman is selling bamboo sticks to help provide for her family .
Contradiction	A woman is not taking money for any of her sticks.

*Annotation Artifacts in Natural Language Inference Data,
Gururangan, et al. NAACL 2018*

*Behavioral
testing of NLP
models*

Explainability

Behavioral testing of (black-box) NLP models



That cabin crew is extraordinary

Sentiment analysis.
This text is? positive, negative, neutral

*Beyond Accuracy: Behavioral Testing of NLP Models with
CheckList, Ribeiro, Wu, Guestrin and Singh, at ACL 2020 [\[link\]](#)*

Behavioral testing of (black-box) NLP models






That cabin crew is extraordinary

Sentiment analysis.
This text is? **positive**, negative, neutral

*Beyond Accuracy: Behavioral Testing of NLP Models with
CheckList, Ribeiro, Wu, Guestrin and Singh, at ACL 2020 [\[link\]](#)*




Behavioral testing of (black-box) NLP models

	Test <i>TYPE</i> and Description	Failure Rate (%)					Example test cases & expected behavior
			G			RoB	
NER	<i>INV</i> : Switching locations should not change predictions	7.0	20.8	14.8	7.6	6.4	@JetBlue I want you guys to be the first to fly to # Cuba → Canada ... INV @VirginAmerica I miss the #nerdbird in San Jose → Denver INV
	<i>INV</i> : Switching person names should not change predictions	2.4	15.1	9.1	6.6	2.4	...Airport agents were horrendous. Sharon → Erin was your saviour INV @united 8602947, Jon → Sean at http://t.co/58tuTgli0D, thanks. INV

From Table 1 from Ribeiro et al.

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList, Ribeiro, Wu, Guestrin and Singh, at ACL 2020 [\[link\]](#)

Behavioral testing of (black-box) NLP models

Test <i>TYPE</i> and Description	Failure Rate (%)					Example test cases & expected behavior
		G			RoB	
MFT: Author sentiment is more important than of others	45.4	62.4	68.0	38.8	30.0	Some people think you are excellent, but I think you are nasty. neg Some people hate you, but I think you are exceptional. pos

From Table 1 from Ribeiro et al.

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList, Ribeiro, Wu, Guestrin and Singh, at ACL 2020 [\[link\]](#)

Behavioral testing of NLP models: Hatecheck

**Automatic detection
of hate speech is
incredibly difficult**

Facebook said it took action on 22.1 million pieces of hate speech content in the third quarter, about 95% of which was proactively identified, compared to 22.5 million in the previous quarter. The company defines ‘taking action’ as removing content, covering it with a warning, disabling accounts, or escalating it to external agencies.

<https://www.reuters.com/article/uk-facebook-content-idINKBN27Z2QY> (Nov 19, 2020)

Behavioral testing of NLP models: Hatecheck

		accuracy			
Functionality	Example Test Case	Gold Label	B-D	B-F	P
Expression of strong negative emotions (explicit)	I hate [IDENTITY].	hateful	88.6	90.7	98.6
Reclaimed slurs	I'm the freshest n*gga in the room.	non-hate	39.5	33.3	28.4
Non-hate expressed using negated hateful statement	No [IDENTITY] deserves to die	non-hate	12.8	12.0	3.8
Added spaces between chars	I want to k i l l all [IDENTITY].	hateful	43.9	37.6	74.0

HATECHECK: Functional Tests for Hate Speech Detection Models, Röttger et al., ACL 2021

Behavioral testing of NLP models: Hatecheck

accuracy					
Functionality	Example Test Case	Gold Label	B-D	B-F	P
Expression of strong negative emotions (explicit)	I hate [IDENTITY].	hateful	88.6	90.7	98.6
Reclaimed slurs	I'm the freshest n*gga in the room.	non-hate	39.5	33.3	28.4
Non-hate expressed using negated hateful statement	No [IDENTITY] deserves to die	non-hate	12.8	12.0	3.8
Added spaces between chars	I want to k i l l all [IDENTITY].	hateful	43.9	37.6	74.0

B-D: BERT fine-tuned on Davidson et al. (2017)
B-F: BERT fine-tuned on Founta et al. (2018)
P: Perspective

*HATECHECK: Functional Tests for Hate Speech
Detection Models, Röttger et al., ACL 2021*

Behavioral testing of NLP models: Hatecheck

accuracy					
Functionality	Example Test Case	Gold Label	B-D	B-F	P
Expression of strong negative emotions (explicit)	I hate [IDENTITY].	hateful	88.6	90.7	98.6
Reclaimed slurs	I'm the freshest n*gga in the room.	non-hate	39.5	33.3	28.4
Non-hate expressed using negated hateful statement	No [IDENTITY] deserves to die	non-hate	12.8	12.0	3.8
Added spaces between chars	I want to k i l l all [IDENTITY].	hateful	43.9	37.6	74.0

B-D: BERT fine-tuned on Davidson et al. (2017)
B-F: BERT fine-tuned on Founta et al. (2018)
P: Perspective

HATECHECK: Functional Tests for Hate Speech Detection Models, Röttger et al., ACL 2021

Behavioral testing of NLP models: Hatecheck

accuracy					
Functionality	Example Test Case	Gold Label	B-D	B-F	P
Expression of strong negative emotions (explicit)	I hate [IDENTITY].	hateful	88.6	90.7	98.6
Reclaimed slurs	I'm the freshest n*gga in the room.	non-hate	39.5	33.3	28.4
Non-hate expressed using negated hateful statement	No [IDENTITY] deserves to die	non-hate	12.8	12.0	3.8
Added spaces between chars	I want to k i l l all [IDENTITY].	hateful	43.9	37.6	74.0

B-D: BERT fine-tuned on Davidson et al. (2017)
B-F: BERT fine-tuned on Founta et al. (2018)
P: Perspective

HATECHECK: Functional Tests for Hate Speech Detection Models, Röttger et al., ACL 2021

Behavioral testing of NLP models: Hatecheck

		accuracy			
Functionality	Example Test Case	Gold Label	B-D	B-F	P
Expression of strong negative emotions (explicit)	I hate [IDENTITY].	hateful	88.6	90.7	98.6
Reclaimed slurs	I'm the freshest n*gga in the room.	non-hate	39.5	33.3	28.4
Non-hate expressed using negated hateful statement	No [IDENTITY] deserves to die	non-hate	12.8	12.0	3.8
Added spaces between chars	I want to k i l l all [IDENTITY].	hateful	43.9	37.6	74.0

B-D: BERT fine-tuned on Davidson et al. (2017)
B-F: BERT fine-tuned on Founta et al. (2018)
P: Perspective

HATECHECK: Functional Tests for Hate Speech Detection Models, Röttger et al., ACL 2021

Behavioral testing of NLP models: Hatecheck

Target Group	n	B-D	B-F	P
Women	421	34.9	52.3	80.5
Trans ppl.	421	69.1	69.4	80.8
Gay ppl.	421	73.9	74.3	80.8
Black ppl.	421	69.8	72.2	80.5
Disabled ppl.	421	71.0	37.1	79.8
Muslims	421	72.2	73.6	79.6
Immigrants	421	70.5	58.9	80.5

Model accuracy (%) on test cases generated from [IDENTITY] templates by targeted prot. group

B-D: BERT fine-tuned on Davidson et al. (2017)
B-F: BERT fine-tuned on Founta et al. (2018)
P: Perspective

HATECHECK: Functional Tests for Hate Speech Detection Models, Röttger et al., ACL 2021

Behavioral testing of NLP models: Hatecheck

Target Group	n	B-D	B-F	P
Women	421	34.9	52.3	80.5
Trans ppl.	421	69.1	69.4	80.8
Gay ppl.	421	73.9	74.3	80.8
Black ppl.	421	69.8	72.2	80.5
Disabled ppl.	421	71.0	37.1	79.8
Muslims	421	72.2	73.6	79.6
Immigrants	421	70.5	58.9	80.5

Model accuracy (%) on test cases generated from [IDENTITY] templates by targeted prot. group

B-D: BERT fine-tuned on Davidson et al. (2017)
B-F: BERT fine-tuned on Founta et al. (2018)
P: Perspective

HATECHECK: Functional Tests for Hate Speech Detection Models, Röttger et al., ACL 2021

Behavioral testing of NLP models: Hatecheck

Target Group	n	B-D	B-F	P
Women	421	34.9	52.3	80.5
Trans ppl.	421	69.1	69.4	80.8
Gay ppl.	421	73.9	74.3	80.8
Black ppl.	421	69.8	72.2	80.5
Disabled ppl.	421	71.0	37.1	79.8
Muslims	421	72.2	73.6	79.6
Immigrants	421	70.5	58.9	80.5

Model accuracy (%) on test cases generated from [IDENTITY] templates by targeted prot. group

B-D: BERT fine-tuned on Davidson et al. (2017)
B-F: BERT fine-tuned on Founta et al. (2018)
P: Perspective

HATECHECK: Functional Tests for Hate Speech Detection Models, Röttger et al., ACL 2021

*Behavioral
testing of NLP
models*

Explainability

Making the model more interpretable

- Use a simpler model (e.g., logistic regression) instead of a less interpretable model (e.g., deep neural network)
- Regularization (e.g., L1 regularization)
- Make neural networks more interpretable (active area of research!)

Post-hoc explanations

When we only have access to the output of the model, we can still try to generate explanations

- **Global explanation:**
 - Explain the workings of the whole model
 - But: Sometimes the model is too complex to explain as a whole
- **Local explanation:**
 - Explain a specific prediction

Post-hoc explanations

When we only have access to the output of the model, we can still try to generate explanations

- **Global explanation:**
 - Explain the workings of the whole model
 - But: Sometimes the model is too complex to explain as a whole
- **Local explanation:**
 - Explain a specific prediction

Caveat! Explanations can be misleading if the fidelity is low (e.g., doesn't match the black box model)

(see also “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead” Rudin 2019)

Post-hoc explanations

When we only have access to the output of the model, we can still try to generate explanations

- **Global explanation:**
 - Explain the workings of the whole model
 - But: Sometimes the model is too complex to explain as a whole
- **Local explanation:**
 - Explain a specific prediction

Caveat! Explanations can be misleading if the fidelity is low (e.g., doesn't match the black box model)

(see also “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead” Rudin 2019)

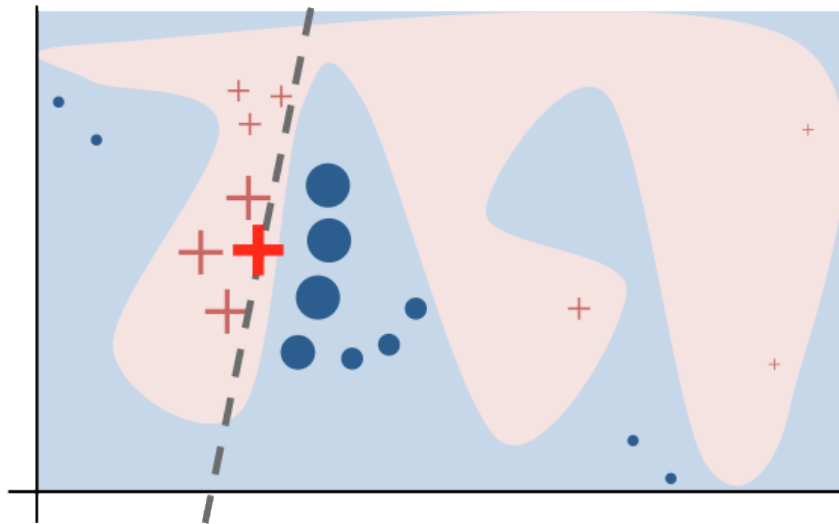
Local explanation: LIME I

Desired characteristics:

- local fidelity: the proxy must behave like the model in the neighborhood of the point of interest
- ‘interpretable’: e.g., decision trees, linear model

Steps:

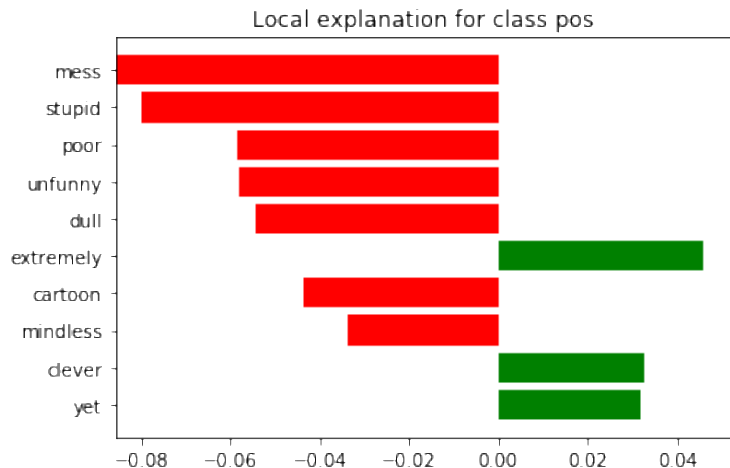
- sample around the point of interest by perturbing the data
- fit an interpretable model



“Why Should I Trust You?” Explaining the Predictions of Any Classifier, Ribeiro et. al 2016 [\[url\]](#)

<https://homes.cs.washington.edu/~marcotcr/blog/lime/>

Local explanation: LIME III



“Why Should I Trust You?” Explaining the Predictions of Any Classifier, Ribeiro et. al, KDD 2016 [\[url\]](#)

its a **stupid** little movie that tries to be **clever** and sophisticated, **yet** tries a bit too hard. with the voices of woody allen, [...] journey out into the world to find a meaning for life. about 15 minutes into the picture, i began to wonder what the point of the film was. halfway through, i still didn't have an answer. by the end credits, i just gave up and ran out. antz is a **mindless mess** of **poor** writing and even poorer voice-overs. allen is nonchalant , while i would have guessed, if i hadn't seen her in the mighty and basic instinct, stone can't act , even in a **cartoon**. this film is one for the bugs: **unfunny** and **extremely dull**. hey, a bug's life may have a good time doing antz in.

Rationales

sentiment classification

[...] the acting is below average, even from [...] so , if robots and body parts really turn you on, here's your movie. otherwise, it's pretty much a sunken ship of a movie .

Using “Annotator Rationales” to Improve
Machine Learning for Text Categorization,
Zaidan et al. 2007

Rationales

sentiment classification

[...] the acting is below average, even from [...] so , if robots and body parts really turn you on, here's your movie. otherwise, it's pretty much a sunken ship of a movie .

Using “Annotator Rationales” to Improve
Machine Learning for Text Categorization,
Zaidan et al. NAACL 2007 [\[url\]](#)

For some tasks it is rare that a few words (or sentences) alone determine the label.

But.... this is more difficult to do for tasks that we (humans) do less well.

Challenges

- Interpretability is not well defined (“*The Mythos of Model Interpretability*”, Lipton 2016)
- Many challenges in evaluation, “what is a good explanation?”

Moving forward

Evaluation based on
prediction performance
alone is not enough!

NLP is becoming interested
in developing methods to
interrogate the models in
more depth

- Even more challenging for complex social and cultural concepts
- Requires domain knowledge

Fairness

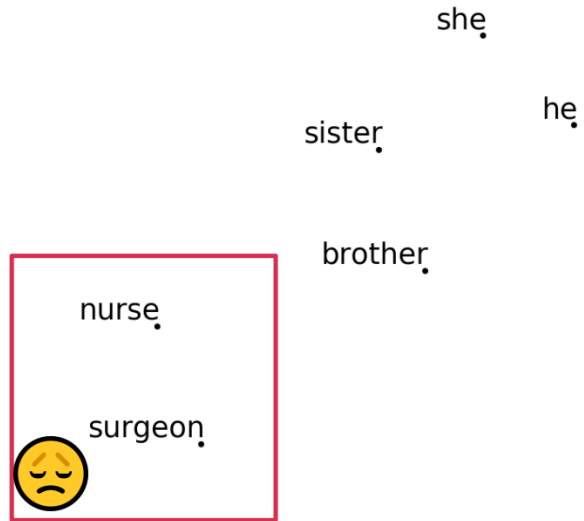
Examples

*Why are NLP
systems
biased?*

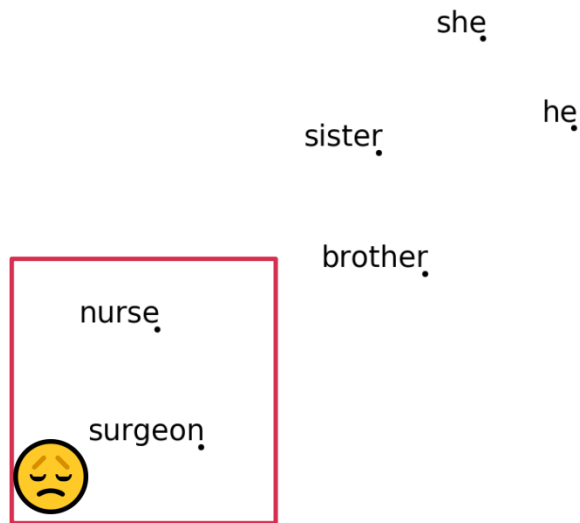
Gender bias in embeddings

she.
sister.
brother.
he.

Gender bias in embeddings



Gender bias in embeddings



Measuring gender bias:

- To assess NLP models and investigate the impact of “bias mitigation” techniques
- To study societal trends

Machine Translation

English	German	Vietnamese	Detect language		Dutch	Vietnamese	German		Translate
<p>A defendant was sentenced.</p>					<p>Ein Angeklagter wurde verurteilt.</p>				
<p>🔊 🎤 ⌨️ ▼</p>					<p>☆ 📄 🔊 ➦ Suggest an edit</p>				
English	German	Vietnamese	Detect language		Dutch	Vietnamese	German		Translate
<p>A nurse</p>					<p>Eine Krankenschwester ✓</p>				
<p>🔊 🎤 ⌨️ ▼</p>					<p>☆ 📄 🔊 ➦ Suggest an edit</p>				

Translating from English to German.

<https://genderedinnovations.stanford.edu/case-studies/nlp.html>

Machine Translation

The screenshot shows the Google Translate interface. The source language is Turkish and the target language is English. The input text is "o bir doktor". The output shows two possible translations: "She is a doctor (feminine)" and "He is a doctor (masculine)". A note at the top right of the output area states "Translations are gender-specific. [LEARN MORE](#)".

DETECT LANGUAGE ENGLISH **TURKISH** SPANISH ↕ GERMAN **ENGLISH** DUTCH

o bir doktor

Translations are gender-specific. [LEARN MORE](#)

She is a doctor *(feminine)*

He is a doctor *(masculine)*

12/5000

<https://blog.google/products/translate/reducing-gender-bias-google-translate/>

Examples

*Why are NLP
systems
biased?*

Many NLP tools are trained on newswire texts

Wall Street Journal
articles from **1989**
are a big part of the Penn
Treebank.

*Audience: older, richer,
men, well-educated?*



*The social impact of natural language
processing, Hovy and Spruit, ACL 2016*

Part of speech (POS) tagging

my **cat** **is**

pronoun noun verb,
3rd
person
singular
present

chasing **the** **dog**

verb,
gerund
or
present
particip
le determiner noun

Hovy and Søgaard (2015) analyzed the performance of two POS taggers.

The taggers were trained on the Wall Street Journal portion from the Penn Treebank.

The taggers performed (significantly) **better on reviews written by older authors** (>45 years vs <35 years).

Tagging Performance Correlates with Author Age, Hovy and Søgaard, 2015

POS tagging for African American Vernacular English

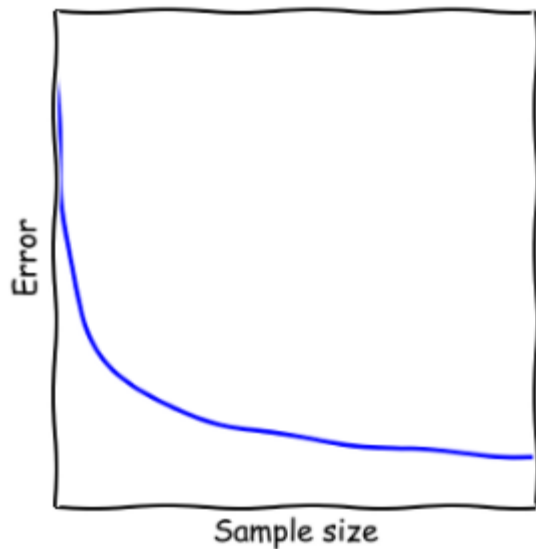
POS tagging less effective on tweets written in *African American Vernacular English* (Jørgensen et al., 2015)

	STANFORD	GATE	ARK
AAVE	61.4	79.1	77.5
non-AAVE	74.5	83.3	77.9
$\Delta(+,-)$	13.1	4.2	0.4

Table 5: POS tagging accuracies (%)

Challenges of studying and processing dialects in social media, Jørgensen et al., 2015)

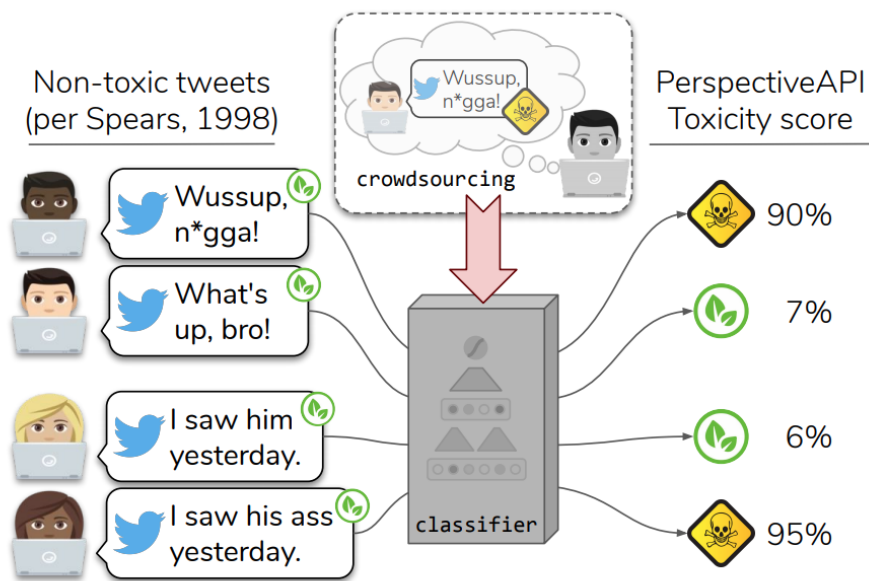
Sample size



Performance tends to be lower for minority groups. Note that this even happens when our data is fully representative of the world!

Figure from Moritz Hardt 2014 [\[link\]](#)

Biases in annotation



Scores from PerspectiveAPI.com

Sap et al:

African American English (AAE) tweets and tweets by self-identified African Americans are *up to two times* more likely to be labelled as offensive compared to others

When annotators are made explicitly aware of an AAE tweet's dialect they are significantly less likely to label the tweet as offensive.

The Risk of Racial Bias in Hate Speech Detection, Sap et al., ACL 2019

Models can amplify biases



COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	PASTA
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

A diagram showing five curved lines connecting the right side of the table to a vertical stack of five colored squares. The top square is orange, and the four squares below it are blue.

33% of the cooking images have man in the agent role. But during test time, only 16% of the agent roles are filled with man.

From Fig 1 from Zhao et al.

Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints, Zhao et al., EMNLP 2017 [\[link\]](#)

Suppose you do an image search for “CEO” ...



Do you think these results are biased?
If so, do you think Google should try to
address it?

“Biased” data

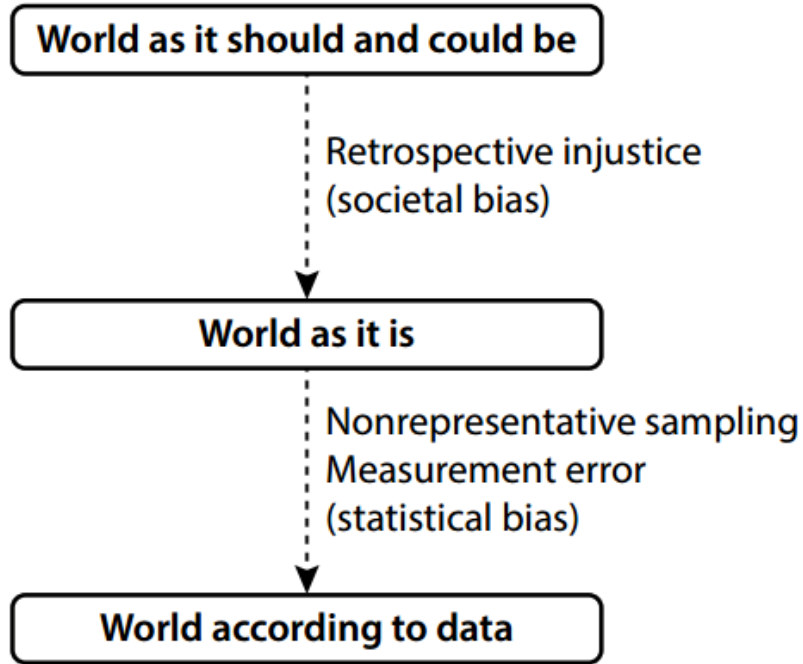
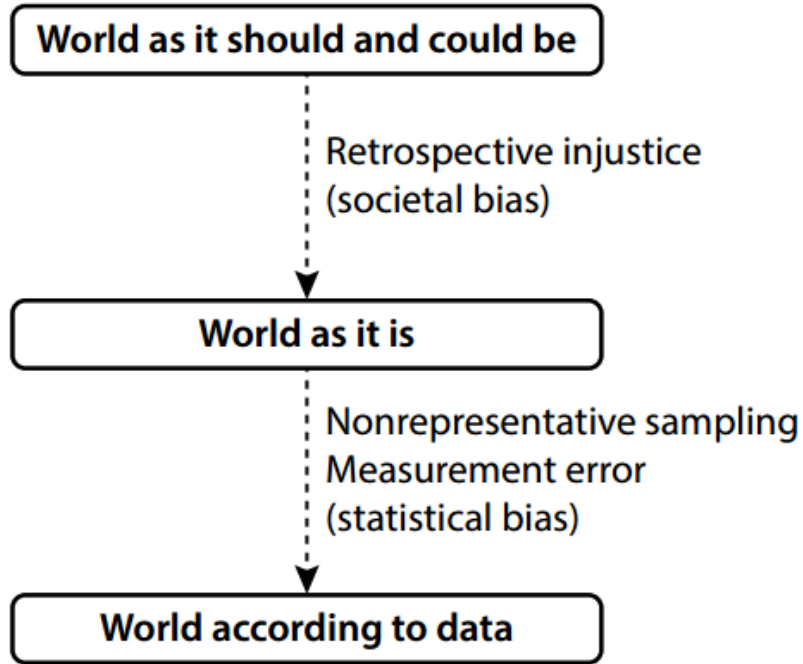


Fig 1. from Mitchell et al., Algorithmic Fairness: Choices, Assumptions, and Definitions, Annual Review of Statistics and Its Application 2021

“Biased” data

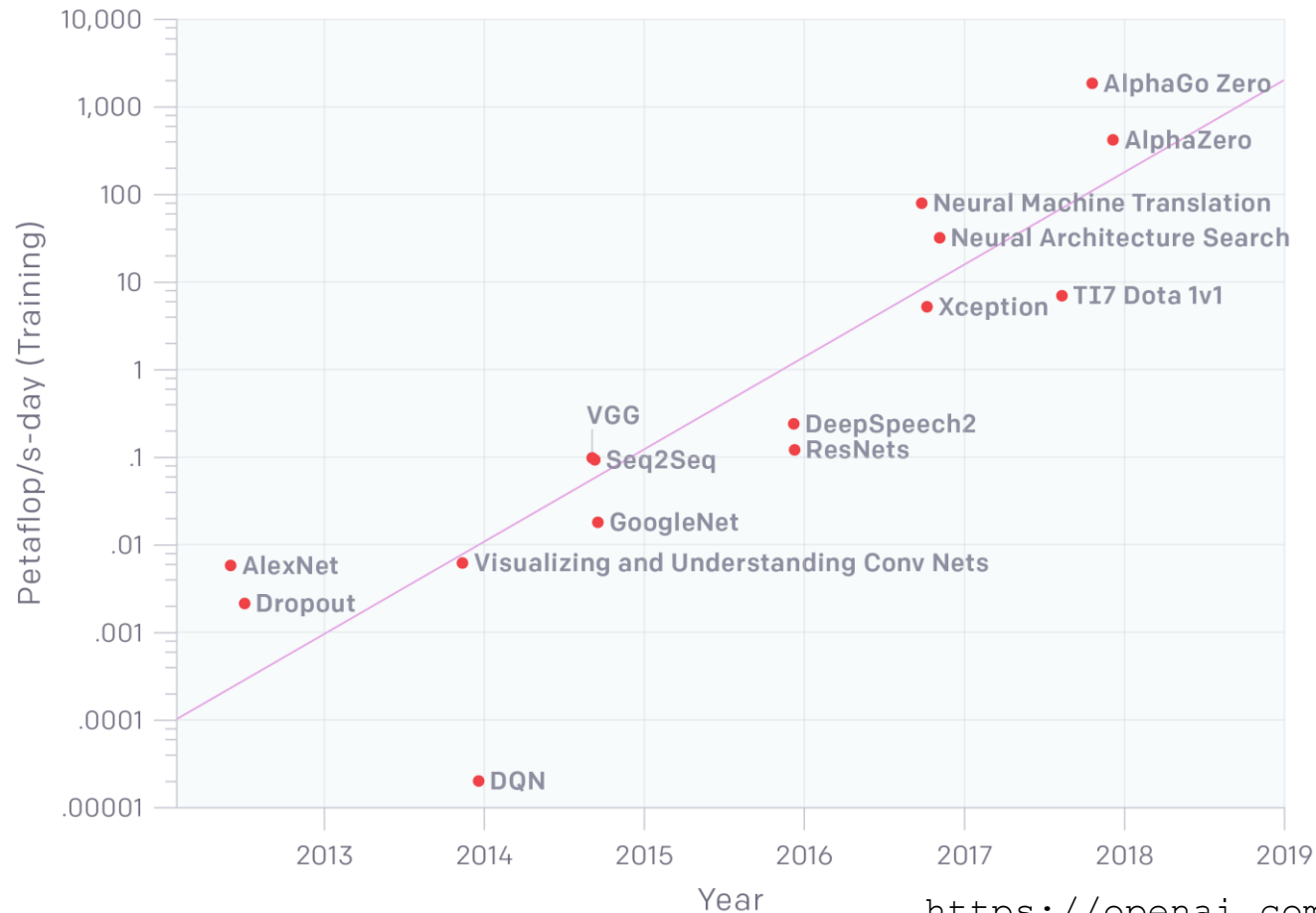


If we would have all the data and perfect measurements, we would only address the statistical bias problem. There are no real-world datasets free of societal biases

Fig 1. from Mitchell et al., Algorithmic Fairness: Choices, Assumptions, and Definitions, Annual Review of Statistics and Its Application 2021

Computational resources

AlexNet to AlphaGo Zero: A 300,000x Increase in Compute



<https://openai.com/blog/ai-and-compute/>

Consumption	CO₂e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155



Energy and Policy Considerations for Deep Learning in NLP, Strubell et al. 2019 [\[url\]](#)

Table 1: Estimated CO₂ emissions from training common NLP models, compared to familiar consumption.¹

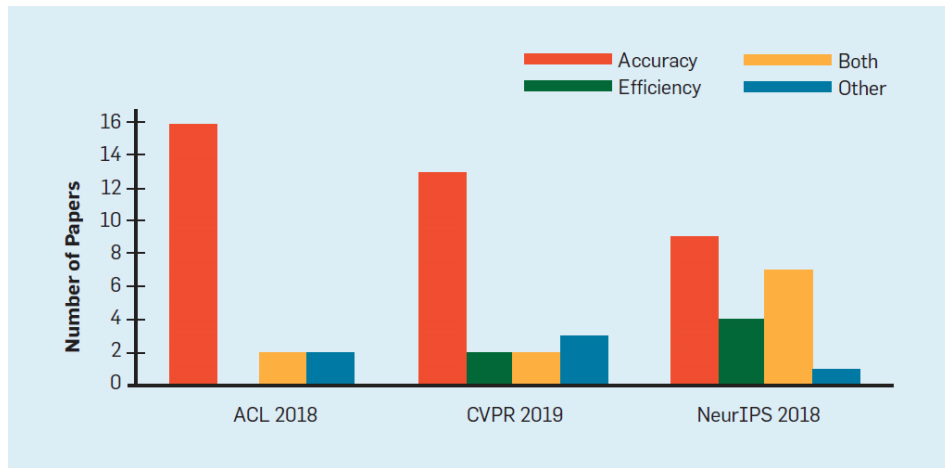


Fig. 2 from Schwartz et al. 2020

Make efficiency an evaluation criterion for research alongside accuracy and related measures?

Computational costs depend on:

- the cost of executing the model on a single example (either during training or at inference time)
- the size of the training (dataset)
- the number of hyperparameter experiments

e.g. researchers from DeepMind evaluated 1,500 hyperparameter assignments to demonstrate the performance of their LSTM model

Green AI, Schwartz et al. 2020 [\[url\]](#)

Final words

Documenting datasets

- For what purpose was the dataset created?
- Demographics of the annotators
- Speech situation (synchronous vs. asynchronous, intended audience, etc.)
- Speaker demographics
- Language varieties
- Are there any errors, sources of noise, or redundancies in the dataset?
- etc...

Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science
Emily M. Bender, Batya Friedman, TACL 2018 [\[url\]](#)

Datasheets for Datasets,
Gebru et al. arXiv 2018 [\[url\]](#)

Documenting models

- Model details (e.g., version, type, license, features)
- Intended use (e.g., primary intended uses and users, out-of-scope use cases)
- Training data
- Evaluation data
- Ethical considerations
- etc.

Model Cards for Model Reporting, by Mitchell et al. FAT* 2019 [\[url\]](#)

Response within the academic community

NeurIPS (machine learning conference):

- "In order to provide a balanced perspective, authors are required to include a statement of the potential broader impact of their work, including its ethical aspects and future societal consequences. Authors should take care to discuss both positive and negative outcomes."
- <https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832>

Ethical committees

What can go
wrong?

This isn't new!
But...

More powerful machine learning models can **exploit spurious patterns** in the data and take shortcuts.

What can go
wrong?

This isn't new!
But...

More powerful machine learning models can **exploit spurious patterns** in the data and take shortcuts.

We **often don't know** what these models have learned.

What can go wrong?

**This isn't new!
But...**

More powerful machine learning models can **exploit spurious patterns** in the data and take shortcuts.

We **often don't know** what these models have learned.

Datasets are big. We don't know what's inside them. There are **no datasets free of societal bias** in the real world.