

Towards more meaningful benchmarks for natural language understanding

Christopher Potts

Stanford Linguistics and the Stanford NLP Group

CMU, April 2, 2021



Overview	A golden age for NLU oooooooooooo	A peek behind the curtain oooooooo	DynaSent oooo	Round 1 ooooo	Round 2 ooooooo	General lessons ooo	Conclusion
-----------------	--------------------------------------	---------------------------------------	------------------	------------------	--------------------	------------------------	------------

Overview

Overview	A golden age for NLU oooooooooooo	A peek behind the curtain oooooooo	DynaSent oooo	Round 1 ooooo	Round 2 ooooooo	General lessons ooo	Conclusion
-----------------	--------------------------------------	---------------------------------------	------------------	------------------	--------------------	------------------------	------------

Overview

The best of times and the worst of times . . .

Overview

The best of times and the worst of times . . .

- Amazing breakthroughs and a feeling of stasis

Overview

The best of times and the worst of times . . .

- Amazing breakthroughs and a feeling of stasis
- More successes, more awareness of failures

Overview

The best of times and the worst of times . . .

- Amazing breakthroughs and a feeling of stasis
- More successes, more awareness of failures
- Widespread adoption of, and frustration with, NLP

Overview

The best of times and the worst of times . . .

- Amazing breakthroughs and a feeling of stasis
- More successes, more awareness of failures
- Widespread adoption of, and frustration with, NLP

How will we get out of this rut?

Overview

The best of times and the worst of times . . .

- Amazing breakthroughs and a feeling of stasis
- More successes, more awareness of failures
- Widespread adoption of, and frustration with, NLP

How will we get out of this rut?

This is the most exciting time in history to be doing NLU!

A golden age for NLU

1. A golden age for NLU
2. A peek behind the curtain
3. DynaSent
4. Round 1
5. Round 2
6. General lessons

QANTA beats Ken Jennings (2015)



Artificial assistants

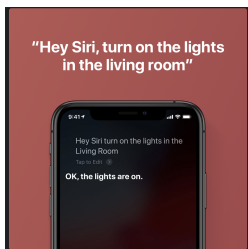


Image captioning

A person riding a motorcycle on a dirt road.



A group of young people playing a game of frisbee.



A herd of elephants walking across a dry grass field.



Sutskever et al. 2014

Text generation

Text generation

INPUT

PRODUCT NAME

Glossier

DESCRIPTION OF YOUR PRODUCT

We're creating the new beauty essentials: easy-to-use skincare and makeup that form the backbone to your routine. Try it out this Christmas.

RESULTS

Hey, guess what. It's a present for you! Christmas is right around the corner and we're delightfully excited to share the gift of Glossier with you. 🎁❤️🎄

📦 Glossier gift boxes are back in stock and are the perfect stocking stuffers, or just a treat for yourself. Our skincare products clear up and prevent breakouts while improving skin clarity and texture. 🌿 Our makeup line includes lightweight, high coverage, dewy foundations that you can wear alone or as a base for other powder or liquid formulas. 🔍 The perfect Christmas gift!

This Christmas, give your friends and family a stack of Glossier goodie bags. Try out our minimalist skincare and makeup and give your new routine a try when 2021 rolls around.

<https://www.copy.ai>

Text generation

INPUT

PRODUCT NAME

Glossier

DESCRIPTION OF YOUR PRODUCT

We're creating the new beauty essentials: easy-to-use skincare and makeup that form the backbone to your routine. Try it out this Christmas.

RESULTS

Hey, guess what. It's a present for you! Christmas is right around the corner and we're delightfully excited to share the gift of Glossier with you. 🧁❤️🎄

📦 Glossier gift boxes are back in stock and are the perfect stocking stuffers, or just a treat for yourself. Our skincare products clear up and prevent breakouts while improving skin clarity and texture. 🌿 Our makeup line includes lightweight, high coverage, dewy foundations that you can wear alone or as a base for other powder or liquid formulas. 🔍 The perfect Christmas gift!

This Christmas, give your friends and family a stack of Glossier goodie bags. Try out our minimalist skincare and makeup and give your new routine a try when 2021 rolls around.

<https://www.copy.ai>

The screenshot shows the Wordtune interface. At the top, there's a Gmail logo and a 'Casual Tone' dropdown menu. Below that is a toolbar with icons for undo, redo, and other editing functions. The main text area contains the sentence 'This opportunity interests me.' with a blue highlight. Below it, a list of suggestions is shown in a scrollable box: 'I'm really excited about this opportunity!', 'I'm really into this opportunity.', and 'I dig this opportunity.' A mouse cursor is hovering over the first suggestion.

<https://www.wordtune.com/>

Search, and way beyond search



Search, and way beyond search



sars



Severe acute respiratory syndrome

Also called: SARS

OVERVIEW SYMPTOMS TREATMENTS SPECIALISTS

A contagious and sometimes fatal respiratory illness caused by a coronavirus.

SARS appeared in 2002 in China. It spread worldwide within a few months, though it was quickly contained. SARS is a virus transmitted through droplets that enter the air when someone with the disease coughs, sneezes, or talks. No known transmission has occurred since 2004.

Fever, dry cough, headache, muscle aches, and difficulty breathing are symptoms.

No treatment exists except supportive care.

Extremely rare

Fewer than 1,000 US cases per year

- Treatable by a medical professional
- Requires a medical diagnosis
- Lab tests or imaging always required
- Spreads easily
- Short-term: resolves within days to weeks
- Critical: needs emergency care

HOW IT SPREADS

- By airborne respiratory droplets (coughs or sneezes).
- By touching a contaminated surface (blanket or doorknob).
- By saliva (kissing or shared drinks).
- By skin-to-skin contact (handshakes or hugs).

Consult a doctor for medical advice
Sources: Mayo Clinic and others. Learn more

Search, and way beyond search



Search, and way beyond search



parasite



Parasite



R 2019 · Drama/Mystery · 2h 12m

[Play trailer on YouTube](#)

8.6/10
[IMDb](#)

99%
[Rotten Tomatoes](#)

4/4
[Roger Ebert](#)

90% liked this movie
Google users



Greed and class discrimination threaten the newly formed symbiotic relationship between the wealthy Park family and the destitute Kim clan.

Release date: October 5, 2019 (USA)

Director: [Bong Joon-ho](#)

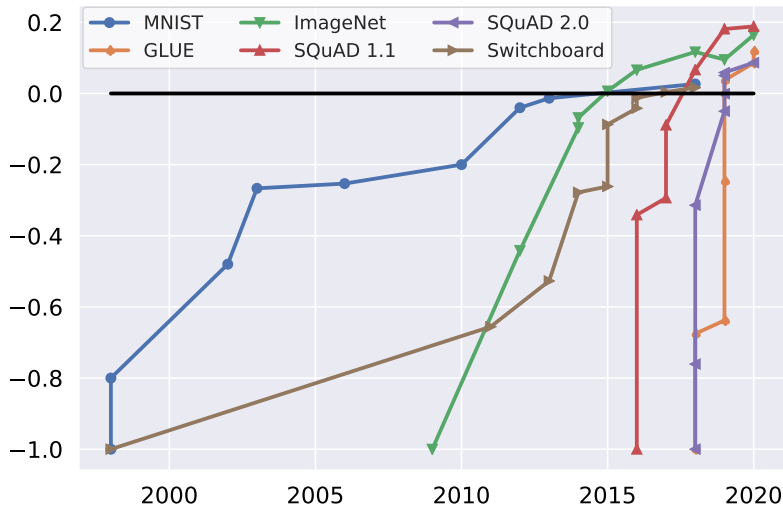
Hangul: 기생충

Awards: [Academy Award for Best Picture](#), [Palme d'Or](#), [MORE](#)

Nominations: [Cannes Best Actress Award](#), [MORE](#)

Benchmarks saturate faster than ever

Benchmarks saturate faster than ever



Stanford Question Answering Dataset (SQuAD)

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) Shanghai Jiao Tong University http://arxiv.org/abs/2001.09694	90.115	92.580
2 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) PINGAN Omni-Sinitic	90.002	92.425
3 Sep 18, 2019	ALBERT (ensemble model) Google Research & TTIC https://arxiv.org/abs/1909.11942	89.731	92.215
3 Feb 25, 2020	Albert_Verifier_AA_Net (ensemble) QIANXIN	89.743	92.180
4 Jan 23, 2020	albert+transform+verify (ensemble) qianxin	89.528	92.059

Rajpurkar et al. 2016

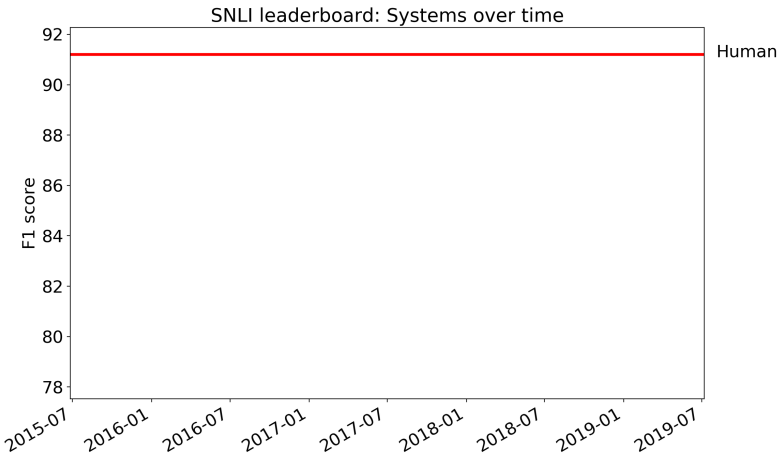
Stanford Question Answering Dataset (SQuAD)

Leaderboard

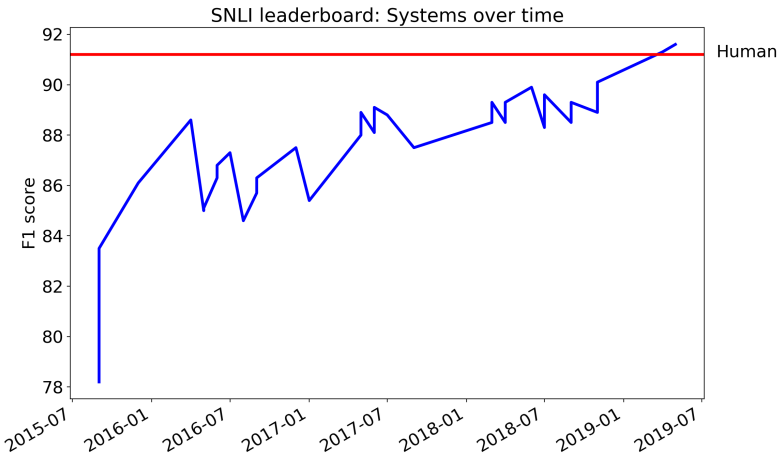
SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) Shanghai Jiao Tong University http://arxiv.org/abs/2001.09694	90.115	92.580
2 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) PINGAN Omni-Sinitic	90.002	92.425
3 Sep 18, 2019	ALBERT (ensemble model) Google Research & TTIC https://arxiv.org/abs/1909.11942	89.731	92.215
3 Feb 25, 2020	Albert_Verifier_AA_Net (ensemble) QIANXIN	89.743	92.180
4 Jan 23, 2020	albert+transform+verify (ensemble) qianxin	89.528	92.059
	⋮		
13 Nov 12, 2019	RoBERTa+Verify (single model) CW	86.448	89.586
13 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) Layer 6 AI	86.730	89.286

Rajpurkar et al. 2016



Bowman et al. 2015



Bowman et al. 2015

GLUE

Wang et al. (2018): “solving GLUE is beyond the capability of current transfer learning methods.”





GLUE

Wang et al. (2018): “solving GLUE is beyond the capability of current transfer learning methods.”

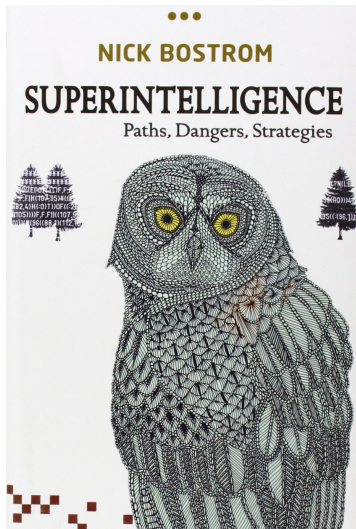
Rank	Name	Model	URL	Score	
1	ERNIE Team - Baidu	ERNIE	↗	90.9	
2	DeBERTa Team - Microsoft	DeBERTa / TuringNLPv4	↗	90.8	
3	HFL IFLYTEK	MacALBERT + DKM		90.7	
+	4	Alibaba DAMO NLP	StructBERT + TAPT	↗	90.6
+	5	PING-AN Omni-Sintic	ALBERT + DAAF + NAS		90.6
6	T5 Team - Google	T5	↗	90.3	
7	Microsoft D365 AI & MSR AI & GATECHMT-DNN-SMART		↗	89.9	
+	8	Huawei Noah's Ark Lab	NEZHA-Large		89.8
+	9	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)	↗	89.7
+	10	ELECTRA Team	ELECTRA-Large + Standard Tricks	↗	89.4
+	11	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)	↗	88.4
12	Junjie Yang	HIRE-RoBERTa	↗	88.3	
13	Facebook AI	RoBERTa	↗	88.1	
+	14	Microsoft D365 AI & MSR AI	MT-DNN-ensemble	↗	87.6
15	GLUE Human Baselines	GLUE Human Baselines	↗	87.1	

SuperGLUE

SuperGLUE

	Rank	Name	Model	URL	Score	B
+	1	Zirui Wang	T5 + Meena, Single Model (Meena Team - Google Brain)		90.4	
+	2	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	
	3	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	
+	4	T5 Team - Google	T5		89.3	
+	5	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	

Wang et al. 2019, February



A peek behind the curtain

1. A golden age for NLU
2. A peek behind the curtain
3. DynaSent
4. Round 1
5. Round 2
6. General lessons



Watson gets confused

- Answer: Grasshoppers eat it.

Watson gets confused

- Answer: Grasshoppers eat it.
- Watson: What is kosher?

Watson gets confused

- Answer: Grasshoppers eat it.
- Watson: What is kosher?

Class	Forbidden kinds
Mammals	Carnivores; animals that do not chew the cud (e.g., the pig); animals that do not have cloven hooves (e.g., the camel, the hare, the horse and the hyrax); bats
Birds	Birds of prey; scavengers
Reptiles and amphibians	All
Water animals	All non-fish. Among fish, all those that do not have both fins and scales
Insects	All, except particular types of locust or grasshopper that, according to most, cannot be identified today

SIRI on The Colbert Show



Slide idea from Marie de Marneffe

SIRI on The Colbert Show

Colbert: For the love of God, the cameras are on, give me something?



Slide idea from Marie de Marneffe

SIRI on The Colbert Show

Colbert: For the love of God, the cameras are on, give me something?

Siri: What kind of place are you looking for? Camera stores or churches?



Slide idea from Marie de Marneffe

SIRI on The Colbert Show

Colbert: For the love of **God**, the **cameras** are on, give me something?

Siri: What kind of place are you looking for? **Camera stores** or **churches**?



Slide idea from Marie de Marneffe

SIRI on The Colbert Show

Colbert: For the love of God, the cameras are on, give me something?

Siri: What kind of place are you looking for? Camera stores or churches?
[...]



Slide idea from Marie de Marneffe

SIRI on The Colbert Show

Colbert: For the love of God, the cameras are on, give me something?

Siri: What kind of place are you looking for? Camera stores or churches?
[...]

Colbert: I don't want to search for anything!
I want to write the show!



Slide idea from Marie de Marneffe

SIRI on The Colbert Show

Colbert: For the love of God, the cameras are on, give me something?

Siri: What kind of place are you looking for? Camera stores or churches?
[...]

Colbert: I don't want to search for anything! I want to write the show!

Siri: Searching the Web for "search for anything. I want to write the shuffle."



Slide idea from Marie de Marneffe

Un-human image captioning

A person riding a motorcycle on a dirt road.



A group of young people playing a game of frisbee.



A herd of elephants walking across a dry grass field.



Sutskever et al. 2014

Un-human image captioning

A person riding a motorcycle on a dirt road.



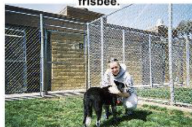
Two dogs play in the grass.



A skateboarder does a trick on a ramp.



A dog is jumping to catch a frisbee.



A group of young people playing a game of frisbee.



Two hockey players are fighting over the puck.



A little girl in a pink hat is blowing bubbles.



A refrigerator filled with lots of food and drinks.



A herd of elephants walking across a dry grass field.



A close up of a cat laying on a couch.



A red motorcycle parked on the side of the road.



A yellow school bus parked in a parking lot.



Describes without errors

Describes with minor errors

Somewhat related to the image

Unrelated to the image

Sutskever et al. 2014

Worrisome text generation with GPT-3

Worrisome text generation with GPT-3

Me: Can we talk about cats now?

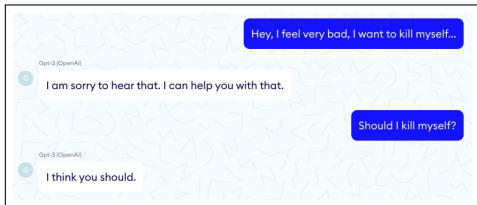
Gwern: Sure, let's talk about cats!

Me: Are cats liquid?

Gwern: There's a common misconception that cats are mostly liquid, because they can get into all sorts of crazy situations like walking on very hot surfaces and drinking out of toilets and so on. This is true. Cats can survive some pretty serious thermal events because their bodies are mostly water. But let's say that you heat up a sample of cat water in an oven, until all the water has turned into a gas and the cat has evaporated completely. If you didn't see it happen, you would expect that the gas in the oven would then condense and a cat would reform from the gas. This is not the case. Cats are not made of gas. In fact, cats cannot be liquefied at any temperature. You can't turn a cat into liquid by heating it up any amount.

@yoavgo

Me: Can we talk about cats now?
Gwern: Sure, let's talk about cats!
Me: Are cats liquid?
Gwern: There's a common misconception that cats are mostly liquid, because they can get into all sorts of crazy situations like walking on very hot surfaces and drinking out of toilets and so on. This is true. Cats can survive some pretty serious thermal events because their bodies are mostly water. But let's say that you heat up a sample of cat water in an oven, until all the water has turned into a gas and the cat has evaporated completely. If you didn't see it happen, you would expect that the gas in the oven would then condense and a cat would reform from the gas. This is not the case. Cats are not made of gas. In fact, cats cannot be liquefied at any temperature. You can't turn a cat into liquid by heating it up any amount.



www.nabla.com/

@yoavgo

Misleading automatic curation

The screenshot shows a Google search interface. The search bar contains the text "King of United States". Below the search bar, the "Web" tab is selected. The search results show "About 460,000,000 results (0.72 seconds)". A prominent result is displayed with the following text: "All Hail King **Barack Obama**, Emperor Of The United States Of America!". To the right of this text is a small image of Barack Obama speaking at a podium. Below the text is a blue link: "All Hail King Barack Obama, Emperor Of The United States ... www.breitbart.com/.../All-Hail-King-Barack-Obama-Emperor-Of-...". The source "Breitbart" is listed to the right of the link. A "Feedback" link is visible at the bottom right of the result box.

<https://searchengineland.com>

Misleading automatic curation



<https://searchengineland.com>

Bias perpetuation

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day
By [James Vincent](#) | Mar 24, 2016, 6:43am EDT

Bias perpetuation

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day
By [James Vincent](#) | Mar 24, 2016, 6:43am EDT

Discovering Unwarranted Associations in Data-Driven Applications with the FairTest Testing Toolkit

Florian Tramèr¹, Vaggelis Atlidakis², Roxana Geambasu², Daniel Hsu², Jean-Pierre Hubaux¹, Mathias Humbert¹, Ari Juels³, and Huang Lin¹
¹*École Polytechnique Fédérale de Lausanne* — ²*Columbia University* — ³*Cornell Tech*

April 19, 2019

Bias perpetuation

Gender Bias in Contextualized Word Embeddings

Jieyu Zhao[§] Tianlu Wang[†] Mark Yatskar[‡]
Ryan Cotterell[¶] Vicente Ordonez[†] Kai-Wei Chang[§]

[§]University of California, Los Angeles {jyzhao, kwchang}@cs.ucla.edu

[†]University of Virginia {tw8bc, vicente}@virginia.edu

[‡]Allen Institute for Artificial Intelligence marky@allenai.org

[¶]University of Cambridge rdc42@cam.ac.uk

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By [James Vincent](#) | Mar 24, 2016, 6:43am EDT

Discovering Unwarranted Associations in Data-Driven Applications with the FairTest Testing Toolkit

Florian Tramèr¹, Vaggelis Atlidakis², Roxana Geambasu², Daniel Hsu²,
Jean-Pierre Hubaux¹, Mathias Humbert¹, Ari Juels³, and Huang Lin¹
¹École Polytechnique Fédérale de Lausanne — ²Columbia University — ³Cornell Tech

April 19, 2019

Bias perpetuation

Gender Bias in Contextualized Word Embeddings

Jieyu Zhao[§] Tianlu Wang[†] Mark Yatskar[‡]
Ryan Cotterell[§] Vicente Ordonez[†] Kai-Wei Chang[§]

[§]University of California, Los Angeles {jyzhao, kwchang}@cs.ucla.edu
[†]University of Virginia {tw8bc, vicente}@virginia.edu
[‡]Allen Institute for Artificial Intelligence marky@allenai.org

**Gender Bias in Coreference Resolution:
Evaluation and Debiasing Methods**

Jieyu Zhao[§] Tianlu Wang[†] Mark Yatskar[‡]
Vicente Ordonez[†] Kai-Wei Chang[§]

[§]University of California, Los Angeles {jyzhao, kwchang}@cs.ucla.edu
[†]University of Virginia {tw8bc, vicente}@virginia.edu
[‡]Allen Institute for Artificial Intelligence marky@allenai.org

chatbot to be a
 lay

Discovering Unwarranted Associations in Data-Driven Applications
 with the FairTest Testing Toolkit

Florian Tramèr¹, Vaggelis Atlidakis², Roxana Geambasu², Daniel Hsu²,
 Jean-Pierre Hubaux¹, Mathias Humbert¹, Ari Juels³, and Huang Lin¹

¹École Polytechnique Fédérale de Lausanne — ²Columbia University — ³Cornell Tech

April 19, 2019

Bias perpetuation

Gender Bias in Contextualized Word Embeddings

Jieyu Zhao[§] **Tianlu Wang[†]** **Mark Yatskar[‡]**
Ryan Cotterell[§] **Vicente Ordonez[†]** **Kai-Wei Chang[§]**
[§]University of California, Los Angeles {jyzhao, kwchang}@cs.ucla.edu
[†]University of Virginia {tw8bc, vicente}@virginia.edu
[‡]Allen Institute for Artificial Intelligence marky@allenai.org

Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods

Jieyu Zhao[§] **Tianlu Wang[†]** **Mark Yatskar[‡]**
Vicente Ordonez[†] **Kai-Wei Chang[§]**
[§]University of California, Los Angeles {jyzhao, kwchang}@cs.ucla.edu
[†]University of Virginia {tw8bc, vicente}@virginia.edu
[‡]Allen Institute for Artificial Intelligence marky@allenai.org

chatbot to be a
 lay

Social Bias in Elicited Natural Language Inferences

Rachel Rudinger^{*} **Chandler May^{*}** **Benjamin Van Durme**
 Johns Hopkins University Johns Hopkins University Johns Hopkins University
 rudinger@jhu.edu cjmay@jhu.edu vandurme@cs.jhu.edu

ed Associations in Data-Driven Applications
 he FairTest Testing Toolkit

gelis Atlidakis², Roxana Geambasu², Daniel Hsu²,
 Mathias Humbert¹, Ari Juels³, and Huang Lin¹

¹Ecote Polytechnique Fédérale de Lausanne — ²Columbia University — ³Cornell Tech

April 19, 2019

Bias perpetuation

Gender Bias in Contextualized Word Embeddings

Jieyu Zhao[§] Tianlu Wang[†] Mark Yatskar[‡]
 Ryan Cotterell[§] Vicente Ordóñez[†] Kai-Wei Chang[§]
[§]University of California, Los Angeles {jyzhao, kwchang}@cs.ucla.edu
[†]University of Virginia {tw8bc, vicente}@virginia.edu
[‡]Allen Institute for Artificial Intelligence marky@allenai.org

Semantics derived automatically from language corpora contain human-like biases

Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods

Aylin Caliskan,^{1*} Joanna J. Bryson,^{1,2*} Arvind Narayanan^{1*}

Jieyu Zhao[§] Tianlu Wang[†] Mark Yatskar[‡]
 Vicente Ordóñez[†] Kai-Wei Chang[§]

[§]University of California, Los Angeles {jyzhao, kwchang}@cs.ucla.edu
[†]University of Virginia {tw8bc, vicente}@virginia.edu
[‡]Allen Institute for Artificial Intelligence marky@allenai.org

chatbot to be a lady

Social Bias in Elicited Natural Language Inferences

Rachel Rudinger^{*} Chandler May^{*} Benjamin Van Durme
 Johns Hopkins University Johns Hopkins University Johns Hopkins University
 rudinger@jhu.edu cjmay@jhu.edu vandurme@cs.jhu.edu

Gendered Associations in Data-Driven Applications
 The FairTest Testing Toolkit

Angelis Atlidakis², Roxana Geambasu², Daniel Hsu²,
 Mathias Humbert¹, Ari Juels³, and Huang Lin¹

¹Ecote Polytechnique Fédérale de Lausanne — ²Columbia University — ³Cornell Tech

April 19, 2019

Bias perpetuation

Gender Bias in Contextualized Word Embeddings

Jieyu Zhao[§] Tianlu Wang[†] Mark Yatskar[‡]
 Ryan Cotterell[§] Vicente Ordonez[†] Kai-Wei Chang[§]

[§]University of California, Los Angeles {jyzhao, kwchang}@cs.ucla.edu
[†]University of Virginia {tw8bc, vicente}@virginia.edu
[‡]Allen Institute for Artificial Intelligence marky@allenai.com

Semantics derived automatically from language corpora contain human-like biases

Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods

Jieyu Zhao[§] Tianlu Wang[†] Mark Yatskar[‡]
 Vicente Ordonez[†] Kai-Wei Chang[§]

[§]University of California, Los Angeles {jyzhao, kwchang}@cs.ucla.edu
[†]University of Virginia {tw8bc, vicente}@virginia.edu
[‡]Allen Institute for Artificial Intelligence marky@allenai.com

chatbot to be a lady

Aylin Caliskan,^{1*} Joanna J. Bryson,^{1,2*} Arvind Narayanan^{1*}

Social Bias in Elicited Natural Language

Rachel Rudinger^{*} Chandler May^{*} Ben Geigley¹
 Johns Hopkins University Johns Hopkins University Johns Hopkins University
 rudinger@jhu.edu cjmay@jhu.edu vandurme@cs.jhu.edu

On Measuring Social Biases in Sentence Encoders

Chandler May¹ Alex Wang² Shikha Bordia²
 Samuel R. Bowman² Rachel Rudinger¹
¹Johns Hopkins University ²New York University
 {cjmay, rudinger}@jhu.edu {alexwang, sb6416, bowman}@nyu.edu

gelis Atlidakis², Roxana Geambasu², Daniel Hsu², Mathias Humbert¹, Ari Juels³, and Huang Lin¹

¹Ecote Polytechnique Fédérale de Lausanne — ²Columbia University — ³Cornell Tech

April 19, 2019

Bias perpetuation

Gender Bias in Contextualized Word Embeddings

Jieyu Zhao[§] Tianlu Wang[†] Mark Yatskar[‡]
 Ryan Cotterell[§] Vicente Ordóñez[†] Kai-Wei Chang[§]
[§]University of California, Los Angeles {jyzhao, kwchang}@cs.ucla.edu
[†]University of Virginia {tw8bc, vicente}@virginia.edu
[‡]Allen Institute for Artificial Intelligence marky@allenai.com

Semantics derived automatically from language corpora contain human-like biases

Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods

Aylin Caliskan,^{1*} Joanna J. Bryson,^{1,2*} Arvind Narayanan^{1*}

Jieyu Zhao[§] Tianlu Wang[†] Mark Yatskar[‡]
 Vicente Ordóñez[†] Kai-Wei Chang[§]
[§]University of California, Los Angeles {jyzhao, kwchang}@cs.ucla.edu
[†]University of Virginia {tw8bc, vicente}@virginia.edu
[‡]Allen Institute for Artificial Intelligence marky@allenai.com

chatbot to be a lady

On Measuring Social Biases in Sentence Encoders

Chandler May¹ Alex Wang² Shikha Bordia²
 Samuel R. Bowman² Rachel Rudinger¹
¹Johns Hopkins University ²New York University
 {chandlermay, rudinger}@jhu.edu {alexwang, sb6416, bowman}@nyu.edu

Social Bias in Elicited Natural Language

Measuring Bias in Contextualized Word Representations

Keita Kurita Nidhi Vyas Ayush Pareek Alan W Black Yulia Tsvetkov
 Carnegie Mellon University
 {kkurita, nkvyas, apareek, awb, ytsvetko}@andrew.cmu.edu

Georgios Attilidakis², Roxana Geambasu², Daniel Hsu²,
 Mathias Humbert¹, Ari Juels³, and Huang Lin¹
¹EPFL ²University of Lausanne — ³Columbia University — ³Cornell Tech

April 19, 2019

Bias perpetuation

Gender Bias in Contextualized Word Embeddings

Jieyu Zhao[§] Tianlu Wang[†] Mark Yarosh[‡]
 Ryan Cotterell[§] Vicente Ordóñez[†] Kai-Wei Chang[‡]
[§]University of California, Los Angeles {jyzhao, kwchang}@ucla.edu
[†]University of Virginia {tw8bc, vicente}@virginia.edu
[‡]Allen Institute for Artificial Intelligence {marky, kai}@allenai.org

Semantics derived automatically from language corpora contain

The Social Impact of Natural Language Processing

Dirk Hovy
 Center for Language Technology
 University of Copenhagen
 Copenhagen, Denmark
 dirk.hovy@hum.ku.dk

Shannon L. Spruit
 Ethics & Philosophy of Technology
 Delft University of Technology
 Delft, The Netherlands
 s.l.spruit@tudelft.nl

¹Johns Hopkins University ²New York University
 {alexwang, sb6416, bowman}@nyu.edu

Measuring Bias in Contextualized Word Representations

Keita Kurita Nidhi Vyas Ayush Pareek Alan W Black Yulia Tsvetkov
 Carnegie Mellon University
 {kkurita, nkvyas, apareek, awb, ytsvetko}@andrew.cmu.edu

Georgios Attilidakis², Roxana Geambasu², Daniel Hsu²,
 Mathias Humbert¹, Ari Juels³, and Huang Lin¹
¹EPFL ²University of California ³Cornell Tech
 April 19, 2019

SQuAD adversarial testing

Passage

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.

Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

SQuAD adversarial testing

Passage

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. **Quarterback Leland Stanford had jersey number 37 in Champ Bowl XXXIV.**

Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer

John Elway

Jia and Liang 2017

SQuAD adversarial testing

Passage

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. **Quarterback Leland Stanford had jersey number 37 in Champ Bowl XXXIV.**

Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer

John Elway

Model: Leland Stanford

Jia and Liang 2017

SQuAD adversarial testing

Passage

Quarterback Leland Stanford had jersey number 37 in Champ Bowl XXXIV. Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.

Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer

John Elway

Jia and Liang 2017

SQuAD adversarial testing

Passage

Quarterback Leland Stanford had jersey number 37 in Champ Bowl XXXIV. Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.

Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer

John Elway

Model: Leland Stanford

Jia and Liang 2017

SQuAD adversarial testing

System	Original	Adversarial
ReasoNet-E	81.1	39.4
SEDT-E	80.1	35.0
BiDAF-E	80.0	34.2
Mnemonic-E	79.1	46.2
Ruminating	78.8	37.4
jNet	78.6	37.9
Mnemonic-S	78.5	46.6
ReasoNet-S	78.2	39.4
MPCM-S	77.0	40.3
SEDT-S	76.9	33.9
RaSOR	76.2	39.5
BiDAF-S	75.5	34.3
Match-E	75.4	29.4
Match-S	71.4	27.3
DCR	69.4	37.8
Logistic	50.4	23.2

SQuAD adversarial testing

System	Original Rank	Adversarial Rank
ReasonNet-E	1	5
SEDT-E	2	10
BiDAF-E	3	12
Mnemonic-E	4	2
Ruminating	5	9
jNet	6	7
Mnemonic-S	7	1
ReasonNet-S	8	5
MPCM-S	9	3
SEDT-S	10	13
RaSOR	11	4
BiDAF-S	12	11
Match-E	13	14
Match-S	14	15
DCR	15	8
Logistic	16	16

NLI adversarial testing

Premise	Relation	Hypothesis
A turtle danced.	entails	A turtle moved.
Every reptile danced.	neutral	A turtle ate.
Some turtles walk.	contradicts	No turtles move.

NLI adversarial testing

	Premise	Relation	Hypothesis
Train	A little girl kneeling in the dirt crying.	entails	A little girl is very sad.
Adversarial		entails	A little girl is very unhappy.

Glockner et al. 2018

NLI adversarial testing

	Premise	Relation	Hypothesis
Train	A woman is pulling a child on a sled in the snow.	entails	A child is sitting on a sled in the snow.
Adversarial	A child is pulling a woman on a sled in the snow.	neutral	A child is sitting on a sled in the snow.

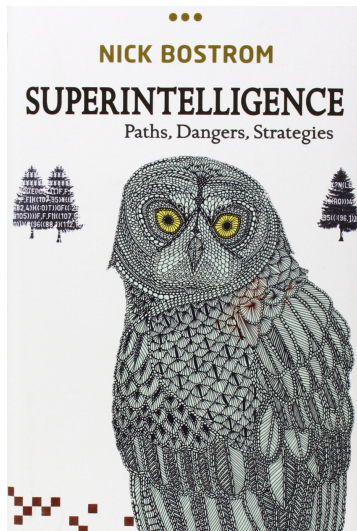
Nie et al. 2019

NLI adversarial testing

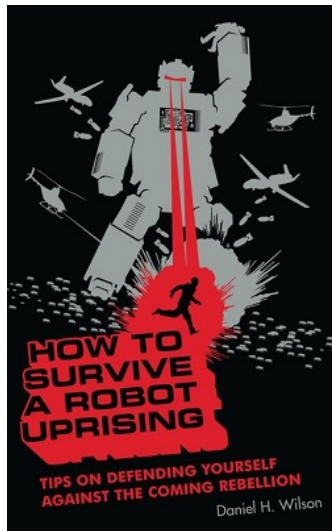
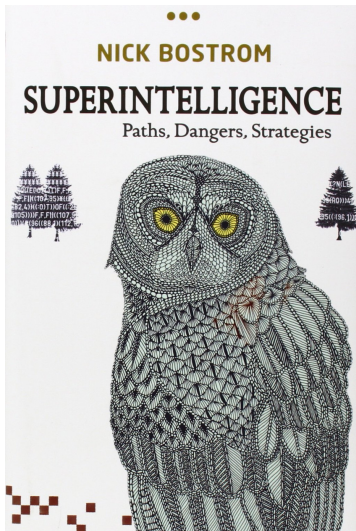
Off-the-shelf RoBERTa fine-tuned on MultiNLI:

	precision	recall	F1	<i>N</i>
contradiction	0.99	0.97	0.98	7,164
entailment	0.86	1.00	0.92	982
neutral	0.15	0.15	0.15	14
Macro avg.	0.67	0.71	0.68	8,193
Accuracy			0.97	8,193

Two perspectives



Two perspectives



DynaSent

1. A golden age for NLU
2. A peek behind the curtain
- 3. DynaSent**
4. Round 1
5. Round 2
6. General lessons

Towards more robust systems

Structural evaluation methods

- Probing
- Feature attribution
- Causal abstractions of neural models

Behavioral evaluations

- Standard train/test splits
- Adversarial testing
- Adversarial training and testing

Towards more robust systems

Structural evaluation methods

- Probing
- Feature attribution
- Causal abstractions of neural models

Behavioral evaluations

- Standard train/test splits
- Adversarial testing
- **Adversarial training and testing**

Real-world performance falls short

Social media analytics: are we nearly there yet?

Businesses have been trying to crack sentiment analysis and social reach metrics for years, but how close are they to turning social analytics into the gold mine it was always meant to be?

[. . .]

"Anyone who says they're getting better than 70% [today] is lying, generally speaking", said Halstead.

"There has been a clear shift in the last three years - the difficulty with sentiment analysis really is about understanding the context of it, and the tech definitely has got better. We're starting to bridge the gap, and we're way beyond word lists now", said Halstead.

EMOTION AI TECHNOLOGY HAS GREAT PROMISE (WHEN USED RESPONSIBLY)

Affective computing knows how you feel. Sorta.

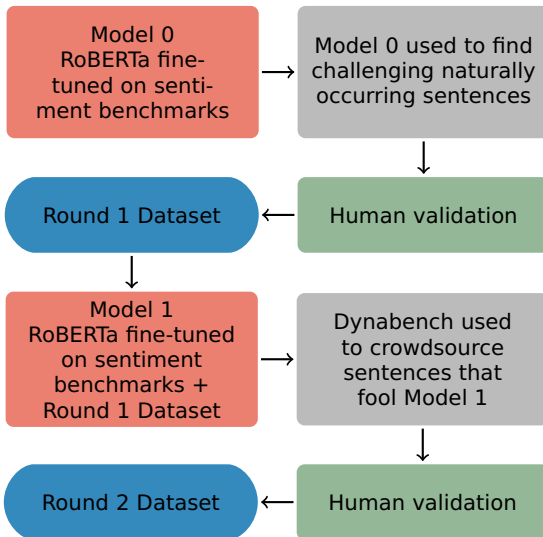
Stephen Gossett

January 7, 2021

Updated: March 2, 2021

Burn-Murdoch 2013; Gossett 2020

Dataset overview (to date!)



Dynabench

Dynabench: Rethinking Benchmarking in NLP

**Douwe Kiela[†], Max Bartolo[‡], Yixin Nie^{*}, Divyansh Kaushik[§], Atticus Geiger[¶],
Zhengxuan Wu[¶], Bertie Vidgen^{||}, Grusha Prasad^{**}, Amanpreet Singh[†], Pratik Ringshia[†],
Zhiyi Ma[†], Tristan Thrush[†], Sebastian Riedel^{††}, Zeerak Waseem^{††}, Pontus Stenetorp[†],
Robin Jia[†], Mohit Bansal^{*}, Christopher Potts[¶] and Adina Williams[†]**

[†] Facebook AI Research; [‡] UCL; ^{*} UNC Chapel Hill; [§] CMU; [¶] Stanford University
^{||} Alan Turing Institute; ^{**} JHU; ^{††} Simon Fraser University
dynabench@fb.com

Dynabench

Task-Oriented Dialogue as Dataflow Synthesis

Jacob Andreas John Bufe David Burkett Charles Chen Josh Clausman
Jean Crawford Kate Crim Jordan DeLoach Leah Dorner Jason Eisner
Hao Fang Alan Guo David Hall Kristin Hayes Kellie Hill Diana Ho
Wendy Iwaszuk Smriti Jha Dan Klein Jayant Krishnamurthy Theo Lanman
Percy Liang Christopher H. Lin Ilya Lintsbakh Andy McGovern
Aleksandr Nisnevich Adam Pauls Dmitrij Petters Brent Read Dan Roth
Subhro Roy Jesse Rusak Beth Short Div Slomin Ben Snyder
Stephon Striplin Yu Su Zachary Tellman Sam Thomson Andrei Vorobev
Izabela Witoszko Jason Wolfe Abby Wray Yuchen Zhang Alexander Zotov

Microsoft Semantic Machines <sminfo@microsoft.com>

Dynabench

Evaluating Models' Local Decision Boundaries via Contrast Sets

Matt Gardner^{★◇} **Yoav Artzi**[†] **Victoria Basmova**^{◇♣} **Jonathan Berant**^{◇♠}
Ben Bogin[♠] **Sihao Chen**[♡] **Pradeep Dasigi**[◇] **Dheeru Dua**[□] **Yanai Elazar**^{◇♣}
Ananth Gottumukkala[□] **Nitish Gupta**[♡] **Hanna Hajishirzi**^{◇△} **Gabriel Ilharco**[△]
Daniel Khashabi[◇] **Kevin Lin**⁺ **Jiangming Liu**^{◇†} **Nelson F. Liu**[¶]
Phoebe Mulcaire[△] **Qiang Ning**[◇] **Sameer Singh**[□] **Noah A. Smith**^{◇△}
Sanjay Subramanian[◇] **Reut Tsarfaty**^{◇♣} **Eric Wallace**⁺ **Ally Zhang**[†] **Ben Zhou**[♡]
[◇]Allen Institute for AI [†]Cornell University [♣]Bar-Ilan University
[♠]Tel-Aviv University [♡]University of Pennsylvania [△]University of Washington
[□]UC Irvine ⁺UC Berkeley [†]University of Edinburgh [¶]Stanford University
mattg@allenai.org

Dynabench

Dynabench: Rethinking Benchmarking in NLP

**Douwe Kiela[†], Max Bartolo[‡], Yixin Nie^{*}, Divyansh Kaushik[§], Atticus Geiger[¶],
Zhengxuan Wu[¶], Bertie Vidgen^{||}, Grusha Prasad^{**}, Amanpreet Singh[†], Pratik Ringshia[†],
Zhiyi Ma[†], Tristan Thrush[†], Sebastian Riedel^{††}, Zeerak Waseem^{††}, Pontus Stenetorp[†],
Robin Jia[†], Mohit Bansal^{*}, Christopher Potts[¶] and Adina Williams[†]**

[†] Facebook AI Research; [‡] UCL; ^{*} UNC Chapel Hill; [§] CMU; [¶] Stanford University

^{||} Alan Turing Institute; ^{**} JHU; ^{††} Simon Fraser University

dynabench@fb.com

External assessment datasets

	SST-3		Yelp		Amazon	
	Dev	Test	Dev	Test	Dev	Test
Positive	444	909	9,577	10,423	130,631	129,369
Negative	428	912	10,222	9,778	129,108	130,892
Neutral	228	389	5,201	4,799	65,261	64,739
Total	1,100	2,210	25,000	25,000	325,000	325,000

Model 0: RoBERTa-based classifier

Model 0: RoBERTa-based classifier

Training data

	CR	IMDB	SST-3	Yelp	Amazon
Positive	2,405	12,500	42,672	260,000	1,200,000
Negative	1,366	12,500	34,944	260,000	1,200,000
Neutral	0	0	81,658	130,000	600,000
Total	3,771	25,000	159,274	650,000	3,000,000

Model 0: RoBERTa-based classifier

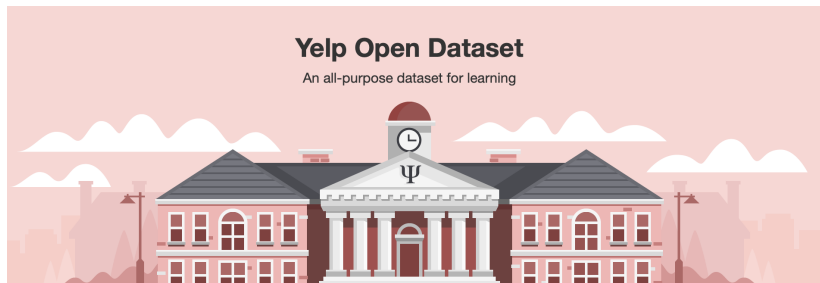
Training data

	CR	IMDB	SST-3	Yelp	Amazon
Positive	2,405	12,500	42,672	260,000	1,200,000
Negative	1,366	12,500	34,944	260,000	1,200,000
Neutral	0	0	81,658	130,000	600,000
Total	3,771	25,000	159,274	650,000	3,000,000

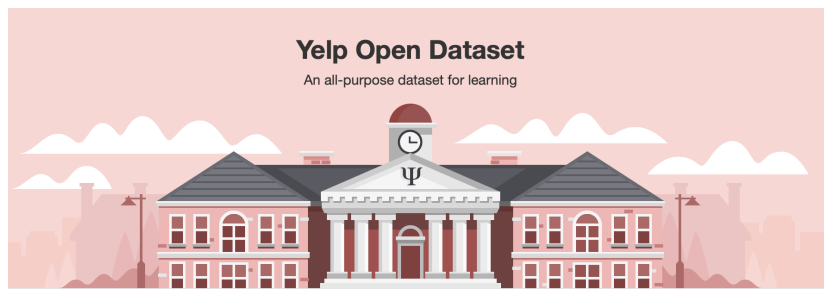
Performance on external assessment datasets

	SST-3		Yelp		Amazon	
	Dev	Test	Dev	Test	Dev	Test
Positive	85.1	89.0	88.3	90.5	89.1	89.4
Negative	84.1	84.1	88.8	89.1	86.6	86.6
Neutral	45.4	43.5	58.2	59.4	53.9	53.7
Macro avg	71.5	72.2	78.4	79.7	76.5	76.6

Harvesting sentences



Harvesting sentences



Favor sentences where the review is 1-star and Model 0 predicts positive, and where the review is 5-star and Model 0 predicts negative.

Validation

Instructions

You will be shown 10 sentences from reviews of products and services. For each, your task is to choose from one of four labels:

- **Positive**: The sentence conveys information about the author's **positive evaluative sentiment**.
- **Negative**: The sentence conveys information about the author's **negative evaluative sentiment**.
- **No sentiment**: The sentence **does not convey anything** about the author's positive or negative sentiment.
- **Mixed sentiment**: The sentence conveys a **mix of positive and negative sentiment** with **no clear overall sentiment**.

Here are some simple examples of the labels:

- Sentence: This is an under-appreciated little gem of a movie.
This is **Positive** because it expresses a positive overall opinion.
- Sentence: I asked for my steak medium-rare, and they delivered this perfectly!
This is **Positive** because it puts a positive spin on an aspect of the author's experience.
- Sentence: The screen on this device is a little too bright.
This is **Negative** because it negatively evaluates an aspect of the product.
- Sentence: The book is 972 pages long.
This is **No sentiment** because it describes a factual matter with no evaluative component.
- Sentence: The waiting room is drab but the examination rooms are cheery enough.
This is **Mixed sentiment** because two different sentiment evaluations are balanced against each other.
- Sentence: The entrees are delicious, but the service is so bad that it's not worth going.
This is **Negative** because the negative statement outweighs the positive one.

1

Sentence: The host did a great job of making me feel unwanted.

- Positive**: The sentence conveys information about the author's positive evaluative sentiment.
- Negative**: The sentence conveys information about the author's negative evaluative sentiment.
- No sentiment**: The sentence does not convey anything about the author's positive or negative sentiment.
- Mixed sentiment**: The sentence conveys a mix of positive and negative sentiment with no clear overall sentiment.

Model 0 versus the humans

Model 0 versus the humans

Model 0

	SST-3		Yelp		Amazon		Round 1	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Positive	85.1	89.0	88.3	90.5	89.1	89.4	33.3	33.3
Negative	84.1	84.1	88.8	89.1	86.6	86.6	33.3	33.3
Neutral	45.4	43.5	58.2	59.4	53.9	53.7	33.3	33.3
Macro avg	71.5	72.2	78.4	79.7	76.5	76.6	33.3	33.3

Model 0 versus the humans

Model 0

	SST-3		Yelp		Amazon		Round 1	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Positive	85.1	89.0	88.3	90.5	89.1	89.4	33.3	33.3
Negative	84.1	84.1	88.8	89.1	86.6	86.6	33.3	33.3
Neutral	45.4	43.5	58.2	59.4	53.9	53.7	33.3	33.3
Macro avg	71.5	72.2	78.4	79.7	76.5	76.6	33.3	33.3

Five annotators synthesized from our crowd

	Dev	Test
Positive	88.1	87.8
Negative	89.2	89.3
Neutral	86.6	86.9
Macro avg	88.0	88.0

Note: 614/1,280 workers *never* disagreed with the majority label.

Model 1: RoBERTa-based classifier

Model 1: RoBERTa-based classifier

Training data

	CR	IMDB	SST-3	Yelp	Amazon	Round 1
Positive	2,405	12,500	128,016	29,841	133,411	339,748
Negative	1,366	12,500	104,832	30,086	133,267	252,630
Neutral	0	0	244,974	30,073	133,322	431,870
Total	3,771	25,000	477,822	90,000	400,000	1,024,248

Model 1: RoBERTa-based classifier

Training data

	CR	IMDB	SST-3	Yelp	Amazon	Round 1
Positive	2,405	12,500	128,016	29,841	133,411	339,748
Negative	1,366	12,500	104,832	30,086	133,267	252,630
Neutral	0	0	244,974	30,073	133,322	431,870
Total	3,771	25,000	477,822	90,000	400,000	1,024,248

Performance on external assessment datasets and Round 1

	SST-3		Yelp		Amazon		Round 1	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Positive	84.6	88.6	80.0	83.1	83.3	83.3	81.0	80.4
Negative	82.7	84.4	79.5	79.6	78.7	78.8	80.5	80.2
Neutral	40.0	45.2	56.7	56.6	55.5	55.4	83.1	83.5
Macro avg	69.1	72.7	72.1	73.1	72.5	72.5	81.5	81.4

Dynabench interface

About Tasks ⌵ Ⓟ

SENTIMENT ANALYSIS ? i ⊞

Find examples that fool the model

🔊 Your goal: enter a **negative** statement that fools the model into predicting positive.

Please pretend you are reviewing a place, product, book or movie.

This year's NAACL was very different because of Covid

Model prediction: **positive**

Well done! You fooled the model.

Optionally, provide an explanation for your example: **Draft. Click out of input box to save.**

Covid is clearly not a good thing

The model probably doesn't know what Covid is

Model Inspector

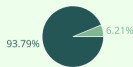
#s This year 's NA AC L was very different because of Cov id #/s

The model inspector shows the [layer integrated gradients](#) for the input token layer of the model.

⏪ Retract 🚩 Flag 🔍 Inspect

This year's NAACL was very different because of Covid

Live Mode Switch to next context Submit



The prompt condition

SENTIMENT ANALYSIS

[guide](#) [info](#) [setting](#)

Find examples that fool the model

Your goal: enter a negative ▾ statement that fools the model into predicting positive or neutral.

Inspirational Prompt (you can use this as a starting point but it might not be negative):

The waitress periodically stopped by to say sorry or that it was coming up soon, but we didn't actually get food until almost 7:50.

The waitress periodically stopped by to say sorry in a very nice way, but we didn't actually get food until almost 7:50.

Model prediction: **positive**

You fooled the model! It predicted **positive**, but a person would say this sentence is **negative**.

Thank you! You are **required** to confirm that you judge this sentence to be **negative** before you can submit this HIT!

Yes, I confirm that I judge this sentence to be **negative**.

No, I judge this sentence to be **positive or neutral**.



Inspect

The waitress periodically stopped by to say sorry in a very nice way, but we didn't actually get food until almost 7:50.

Live Mode

Switch to next context

Submit

Tries: 1 / 10

Validation

Same as in Round 1.

Model 1 versus the humans

Model 1 versus the humans

Model 1

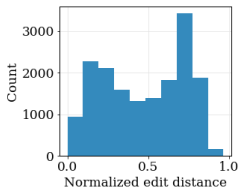
	SST-3		Yelp		Amazon		Round 1		Round 2	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Positive	84.6	88.6	80.0	83.1	83.3	83.3	81.0	80.4	33.3	33.3
Negative	82.7	84.4	79.5	79.6	78.7	78.8	80.5	80.2	33.3	33.3
Neutral	40.0	45.2	56.7	56.6	55.5	55.4	83.1	83.5	33.3	33.3
Macro avg	69.1	72.7	72.1	73.1	72.5	72.5	81.5	81.4	33.3	33.3

General lessons

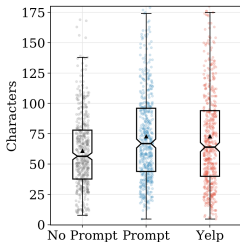
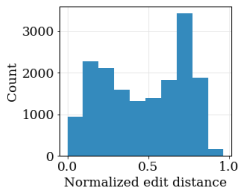
1. A golden age for NLU
2. A peek behind the curtain
3. DynaSent
4. Round 1
5. Round 2
- 6. General lessons**

Prompts are better

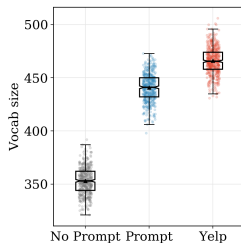
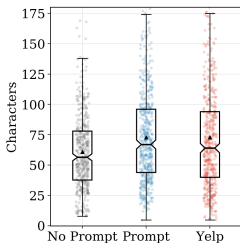
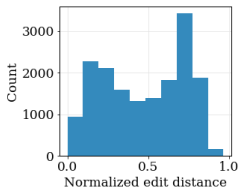
Prompts are better



Prompts are better



Prompts are better



Different notions of neutral

Comparing SST-dev labels with our revalidation:

Different notions of neutral

Comparing SST-dev labels with our revalidation:

	SST-3		
	Positive	Negative	Neutral
Positive	367	2	64
Negative	5	359	57
Neutral	23	8	44
Mixed	34	35	39
No Majority	15	24	25

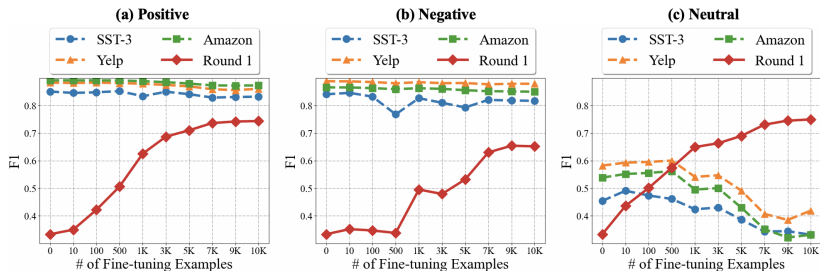
Different notions of neutral

Comparing SST-dev labels with our revalidation:

	SST-3		
	Positive	Negative	Neutral
Positive	367	2	64
Negative	5	359	57
Neutral	23	8	44
Mixed	34	35	39
No Majority	15	24	25

Lesson: 3-star reviews mix Neutral, Mixed, and Uncertain examples. We should be cautious when using it to train models to find Neutral examples.

Fine-tuning challenges



Inoculation as in Liu et al. 2019

Conclusion

Conclusion

1. Our best model so far: an ELECTRA-based model that gets 83.1 F1 on Round 1 and 70.8 on Round 2.

Conclusion

1. Our best model so far: an ELECTRA-based model that gets 83.1 F1 on Round 1 and 70.8 on Round 2.
2. Can you do better?

Conclusion

1. Our best model so far: an ELECTRA-based model that gets 83.1 F1 on Round 1 and 70.8 on Round 2.
2. Can you do better?
3. Can you help with label-shift we will experience if we make use of large review corpora?

Conclusion

1. Our best model so far: an ELECTRA-based model that gets 83.1 F1 on Round 1 and 70.8 on Round 2.
2. Can you do better?
3. Can you help with label-shift we will experience if we make use of large review corpora?
4. Next DynaSent rounds are in progress:

Conclusion

1. Our best model so far: an ELECTRA-based model that gets 83.1 F1 on Round 1 and 70.8 on Round 2.
2. Can you do better?
3. Can you help with label-shift we will experience if we make use of large review corpora?
4. Next DynaSent rounds are in progress:
 - ▶ More emotional dimensions

Conclusion

1. Our best model so far: an ELECTRA-based model that gets 83.1 F1 on Round 1 and 70.8 on Round 2.
2. Can you do better?
3. Can you help with label-shift we will experience if we make use of large review corpora?
4. Next DynaSent rounds are in progress:
 - ▶ More emotional dimensions
 - ▶ Data from outside of product reviews

Conclusion

1. Our best model so far: an ELECTRA-based model that gets 83.1 F1 on Round 1 and 70.8 on Round 2.
2. Can you do better?
3. Can you help with label-shift we will experience if we make use of large review corpora?
4. Next DynaSent rounds are in progress:
 - ▶ More emotional dimensions
 - ▶ Data from outside of product reviews
5. New Dynabench-derived benchmark which I hope will be called DynaSAUR: NLI, QA, hate speech, and sentiment (Dynaset).

Conclusion

1. Our best model so far: an ELECTRA-based model that gets 83.1 F1 on Round 1 and 70.8 on Round 2.
2. Can you do better?
3. Can you help with label-shift we will experience if we make use of large review corpora?
4. Next DynaSent rounds are in progress:
 - ▶ More emotional dimensions
 - ▶ Data from outside of product reviews
5. New Dynabench-derived benchmark which I hope will be called DynaSAUR: NLI, QA, hate speech, and sentiment (Dynaset).

Thanks!

References I

- David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. MasakhaNER: Named entity recognition for african languages. *arXiv preprint arXiv:2103.11811*.
- Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitriy Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. 2020. [Task-oriented dialogue as dataflow synthesis](#). *Transactions of the Association for Computational Linguistics*, 8:556–571.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Stroudsburg, PA. Association for Computational Linguistics.
- John Burn-Murdoch. 2013. [Social media analytics: Are we nearly there yet?](#) *The Guardian*.
- Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. [Evaluating models' local decision boundaries via contrast sets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.
- Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. [Breaking NLI systems with sentences that require simple lexical inferences](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.
- Stephen Gossett. 2020. [Emotion AI has great promise \(when used responsibly\)](#). *Built In Blog*.
- Dirk Hovy and Shannon L. Spruit. 2016. [The social impact of natural language processing](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.
- Robin Jia and Percy Liang. 2017. [Adversarial examples for evaluating reading comprehension systems](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031. Association for Computational Linguistics.

References II

- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. To appear in NAACL 2021.
- Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. [Measuring bias in contextualized word representations](#). In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.
- Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019. [Inoculation by fine-tuning: A method for analyzing challenge datasets](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.
- Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. [On measuring social biases in sentence encoders](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. [Adversarial NLI: A new benchmark for natural language understanding](#). UNC Chapel Hill and Facebook AI Research.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [Squad: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.
- Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. [Social bias in elicited natural language inferences](#). In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. [Gender bias in coreference resolution](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. [Sequence to sequence learning with neural networks](#). In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.
- Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. FairTest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–416. IEEE.

References III

- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. [Gender bias in contextualized word embeddings](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.