Supervised sentiment analysis: DynaSent

Christopher Potts

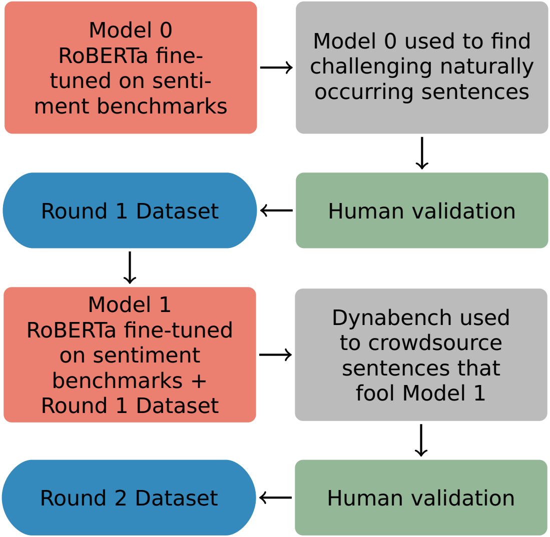
Stanford Linguistics

CS224u: Natural language understanding

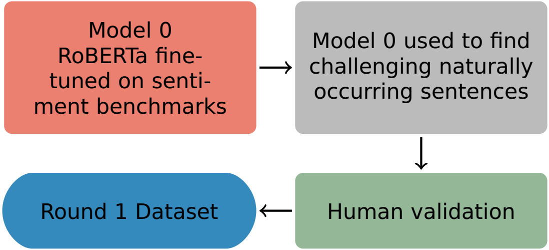
Project overview

- Data, code, and models:
  `https://github.com/cgpotts/dynasent`
- 121,634 sentences, across two rounds, each with 5 gold labels
- Paper: Potts et al. 2020
- Dynabench: `https://dynabench.org`

Dataset overview



Round 1


Model 0
RoBERTa fine-tuned on sentiment benchmarks
Model 0 used to find challenging naturally occurring sentences
Human validation
Round 1 Dataset

Model 0: RoBERTa-based classifier

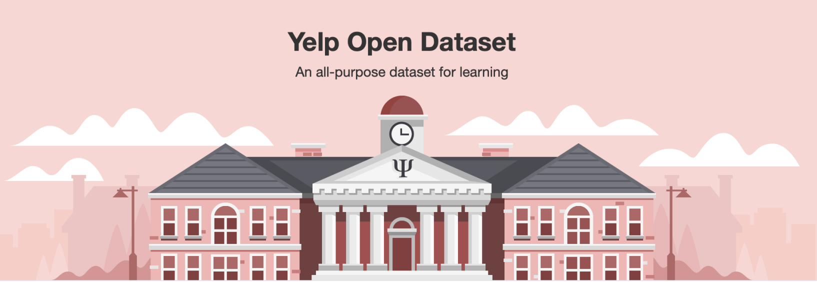
Training data

	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



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.**

Resulting dataset

	Dist	Majority Label		
	Train	Train	Dev	Test
Positive	130,045	21,391	1,200	1,200
Negative	86,486	14,021	1,200	1,200
Neutral	215,935	45,076	1,200	1,200
Mixed	39,829	3,900	0	0
No Majority	–	10,071	0	0
Total	472,295	94,459	3,600	3,600

47% adversarial examples

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

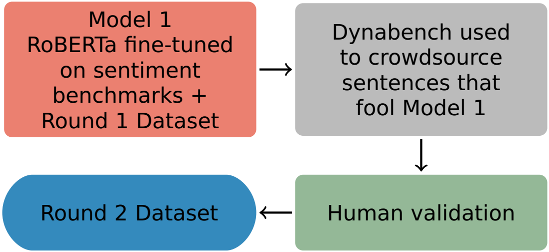
	Dev	Test
Positive	88.1	87.8
Negative	89.2	89.3
Neutral	86.6	86.9
Macro avg	88.0	88.0

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

Randomly sampled (short) examples

Sentence	Model 0	Responses
Good food nasty attitude by hostesses .	neg	**mix, mix, mix**, neg, neg
Not much of a cocktail menu that I saw.	neg	**neg, neg, neg, neg, neg**
I scheduled the work for 3 weeks later.	neg	**neu, neu, neu, neu**, pos
I was very mistaken, it was much more!	neg	neg, **pos, pos, pos, pos**
It is a gimmick, but when in Rome, I get it.	neu	**mix, mix, mix**, neu, neu
Probably a little pricey for lunch.	neu	mix, **neg, neg, neg, neg**
But this is strictly just my opinion.	neu	**neu, neu, neu, neu**, pos
The price was okay, not too pricey.	neu	mix, neu, **pos, pos, pos**
The only downside was service was a little slow.	pos	**mix, mix, mix**, neg, neg
However there is a 2 hr seating time limit.	pos	mix, **neg, neg, neg**, neu
With Alex, I never got that feeling.	pos	**neu, neu, neu, neu**, pos
Its ran very well by management.	pos	**pos, pos, pos, pos, pos**

Round 2


Model 1
RoBERTa fine-tuned
on sentiment
benchmarks +
Round 1 Dataset
Dynabench used
to crowdsource
sentences that
fool Model 1
Human validation
Round 2 Dataset

Model 1: RoBERTa-based classifier

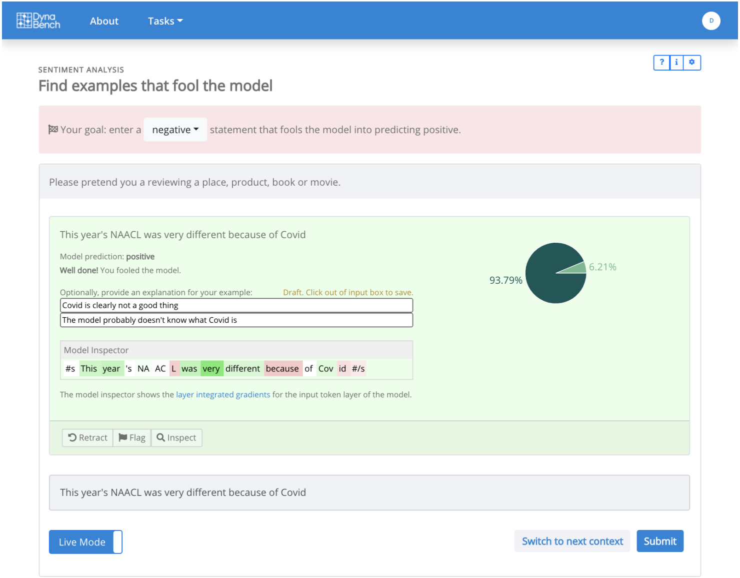
Training data

	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
Model 0	71.5	72.2	78.4	79.7	76.5	76.6	33.3	33.3

Dynabench interface



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.

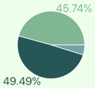
Model prediction: **positive**

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.

Resulting dataset

	Dist	Majority Label		
	Train	Train	Dev	Test
Positive	32,551	6,038	240	240
Negative	24,994	4,579	240	240
Neutral	16,365	2,448	240	240
Mixed	18,765	3,334	0	0
No Majority	–	2,136	0	0
Total	92,675	18,535	720	720

19% adversarial examples

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

Five annotators synthesized from our crowd

	Dev	Test
Positive	91.0	90.9
Negative	91.2	91.0
Neutral	88.9	88.2
Macro avg	90.4	90.0

Note: 116/244 workers *never* disagreed with the majority label.

Randomly sampled (short) examples

Sentence	Model 1	Responses
The place was somewhat good and not well	neg	**mix, mix, mix, mix**, neg
I bought a new car and met with an accident.	neg	**neg, neg, neg, neg, neg**
The retail store is closed for now at least.	neg	**neu, neu, neu, neu, neu**
Prices are basically like garage sale prices.	neg	neg, neu, **pos, pos, pos**
That book was good. I need to get rid of it.	neu	**mix, mix, mix**, neg, pos
I REALLY wanted to like this place	neu	mix, **neg, neg, neg**, pos
I'm going to leave my money for the next vet.	neu	neg, **neu, neu, neu, neu**
once the model made a super decision.	neu	**pos, pos, pos, pos, pos**
I cook my caribbean food and it was okay	pos	**mix, mix, mix**, pos, pos
This concept is really cool in name only.	pos	mix, **neg, neg, neg**, neu
Wow, it'd be super cool if you could join us	pos	**neu, neu, neu, neu**, pos
Knife cut thru it like butter! It was great.	pos	**pos, pos, pos, pos, pos**

References I

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2020. DynaSent: A dynamic benchmark for sentiment analysis. *arXiv preprint arXiv:2012.15349*.