

Machine Learning Part 3: Evaluation (for real)

LIN 313 Language and Computers
UT Austin Fall 2025
Instructor: Gabriella Chronis

Overview 9/26

- Machine Learning Evaluation
 - precision + recall
- kinds of classification tasks
 - binary + multi-class
- classification examples
- classification scheme brainstorming activity

How do we evaluate performance on the test set?

1. We build a 'toxicity' classifier
2. We run the classifier on the test set to get predictions
3. What metric can we use to evaluate performance?

UserName	ScreenName	Location	TweetAt	OriginalTweet	Toxicity
3799	48751	London	16-03-2020	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Normal
3800	48752	UK	16-03-2020	advice Talk to your neighbours family to excha...	Normal
3801	48753	Vagabonds	16-03-2020	Coronavirus Australia: Woolworths to give elde...	Normal
3802	48754	NaN	16-03-2020	My food stock is not the only one which is emp...	Toxic
3803	48755	NaN	16-03-2020	Me, ready to go at supermarket during the #COV... Extremely	Toxic

How do we evaluate performance on the test set?

1. We build a 'toxicity' classifier
2. We run the classifier on the test set to get predictions
3. What metric can we use to evaluate performance?


UserName	ScreenName	Location	TweetAt	OriginalTweet	Toxicity
3799	48751	London	16-03-2020	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Normal
3800	48752	UK	16-03-2020	advice Talk to your neighbours family to excha...	Normal
3801	48753	Vagabonds	16-03-2020	Coronavirus Australia: Woolworths to give elde...	Normal
3802	48754	NaN	16-03-2020	My food stock is not the only one which is emp...	Toxic
3803	48755	NaN	16-03-2020	Me, ready to go at supermarket during the #COV... Extremely	Toxic



How do we evaluate performance on the test set?

1. We build a 'toxicity' classifier
2. We run the classifier on the test set to get predictions
3. What metric can we use to evaluate performance?

UserName	ScreenName	Location	TweetAt	OriginalTweet	Toxicity
3799	48751	London	16-03-2020	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Normal
3800	48752	UK	16-03-2020	advice Talk to your neighbours family to excha...	Normal
3801	48753	Vagabonds	16-03-2020	Coronavirus Australia: Woolworths to give elde...	Normal
3802	48754	NaN	16-03-2020	My food stock is not the only one which is emp...	Toxic
3803	48755	NaN	16-03-2020	Me, ready to go at supermarket during the #COV... Extremely	Toxic



Accuracy Of a Hate Speech Detector

Our test set contains 100 examples.

The model was right on 91 of them.

Accuracy Of a Hate Speech Detector

$$\textit{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Number of total predictions}}$$

Our test set contains 100 examples.

The model was right on 91 of them.

$$\begin{aligned}\textit{Acc} &= 91 / 100 \\ &= 91\%\end{aligned}$$

Dataset imbalance

Accuracy is not always the best indicator of performance.

In real life the data is extremely skewed or *class imbalanced*.

Maybe 0.05% of comments actually contain hate speech.

Dataset imbalance

Accuracy is not always the best indicator of performance.

In real life the data is extremely skewed or *class imbalanced*.

Maybe 0.05% of real life comments actually contain hate speech.

Imagine a classifier that predicts "normal text" for every input comment.

If the test set reflects real life data distribution, how accurate would the model be?

Dataset imbalance

Accuracy is not always the best indicator of performance.

In real life the data is extremely skewed or *class imbalanced*.

Maybe 0.05% of real life comments actually contain hate speech.

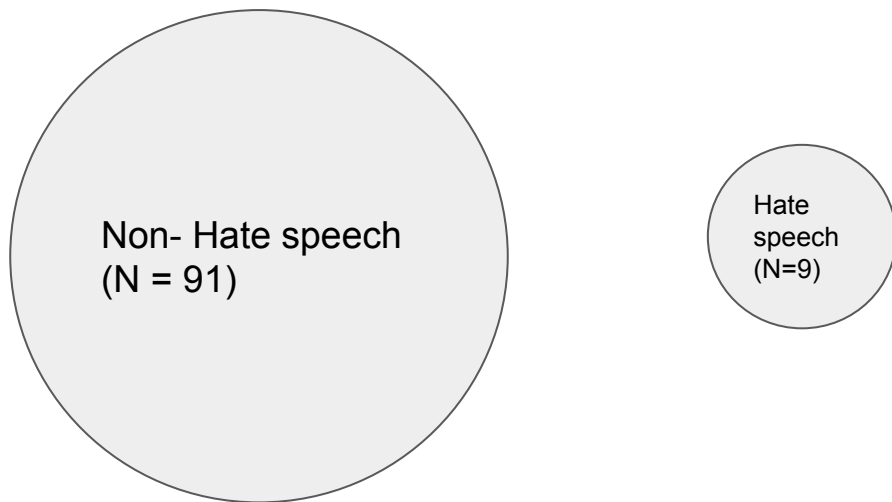
Imagine a classifier that predicts "normal text" for every input comment.

If the test set reflects real life data distribution, how accurate would the model be?

99.5% accurate!

Accuracy. . .

. . . alone doesn't tell the full story when you're working with a **class-imbalanced data set**, like this one, where there is a significant disparity between the number of positive and negative labels



Types of Error

What are the different ways that a model can be wrong?

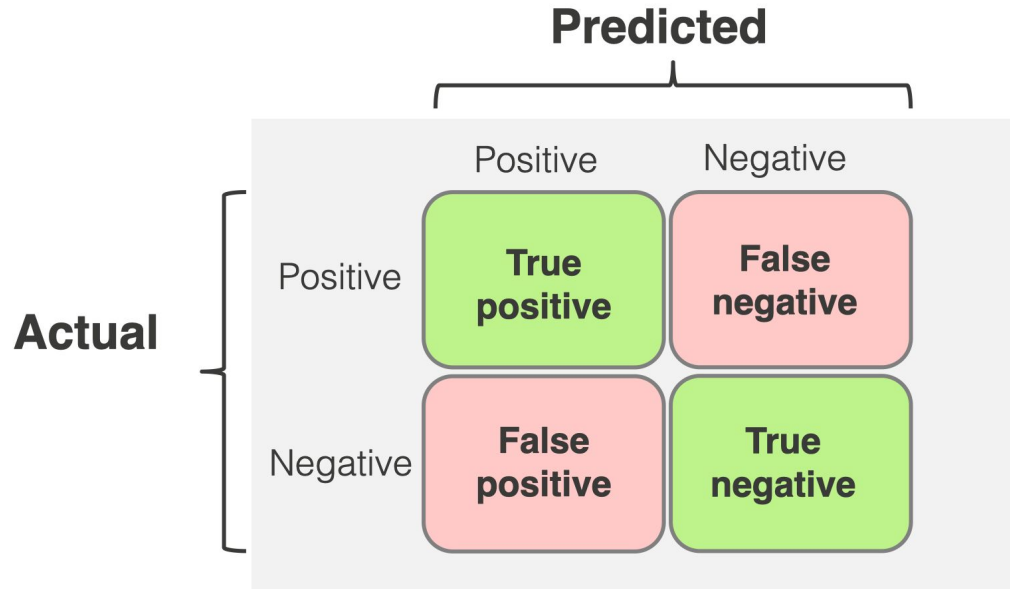
Types of Error

What are the different ways that a model can be wrong?

HINT: what are the possible relationships between predicted value and the actual value?

Types of error

We can visualize types of error like this. (called a **confusion matrix**)



Types of error

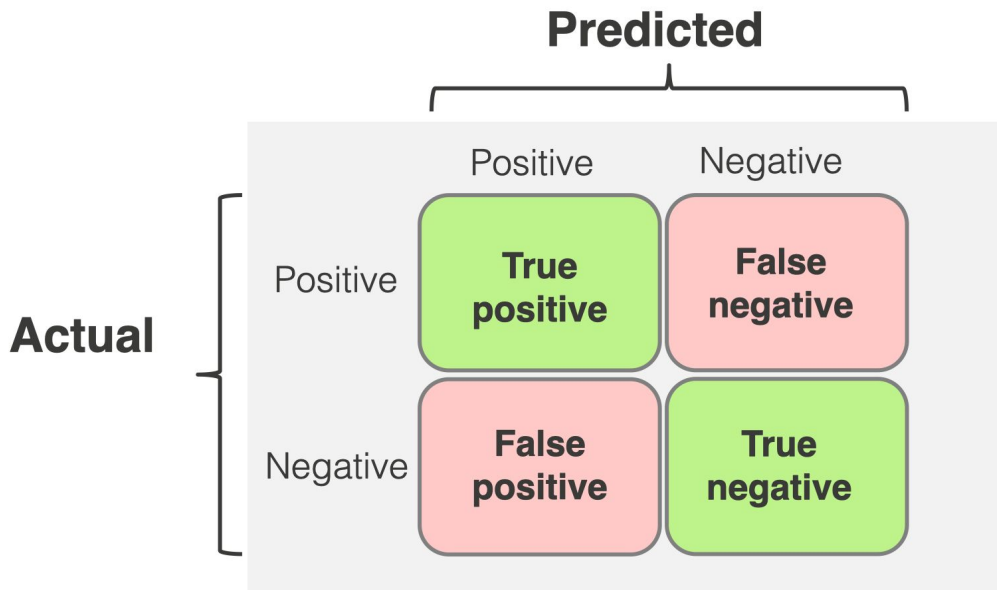
We can visualize types of error like this. (called a **confusion matrix**)

True positives: model guesses toxic, actually toxic

False positives: model guesses toxic, actually normal

True negatives: model guesses normal, actually normal

False negatives: model guesses normal, actually toxic



Accuracy Of a Hate Speech Detector

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Number of total predictions}}$$

Actual

		Predicted	
		Positive	Negative
Actual	Positive	True Positives Reality: toxic Prediction: "toxic" N = 1	False Negatives Reality: toxic Prediction: "normal" N = 8
	Negative	False Positives Actual: normal Prediction: "toxic" N = 1	True Negatives Reality: normal Prediction: normal N = 90

Accuracy Of a Hate Speech Detector

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Number of total predictions}}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Actual

Predicted

	Predicted	
	Positive	Negative
Positive	True Positives Reality: toxic Prediction: "toxic" N = 1	False Negatives Reality: toxic Prediction: "normal" N = 8
Negative	False Positives Actual: normal Prediction: "toxic" N = 1	True Negatives Reality: normal Prediction: normal N = 90

Accuracy Of a Hate Speech Detector

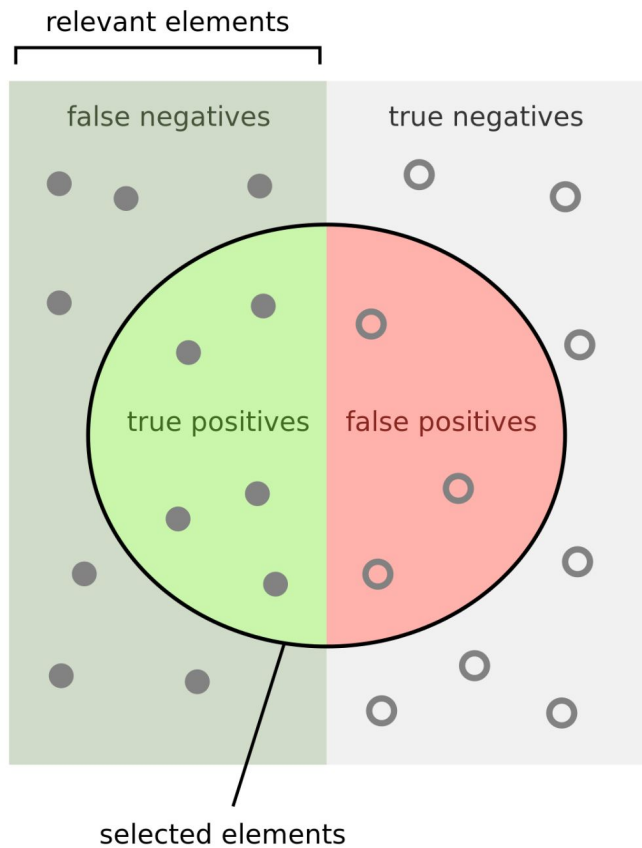
$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Number of total predictions}}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$= \frac{1 + 90}{1 + 90 + 1 + 8} = 0.91$$

Actual

		Predicted	
		Positive	Negative
Actual	Positive	True Positives Reality: toxic Prediction: "toxic" N = 1	False Negatives Reality: toxic Prediction: "normal" N = 8
	Negative	False Positives Actual: normal Prediction: "toxic" N = 1	True Negatives Reality: normal Prediction: normal N = 90

Precision/Recall



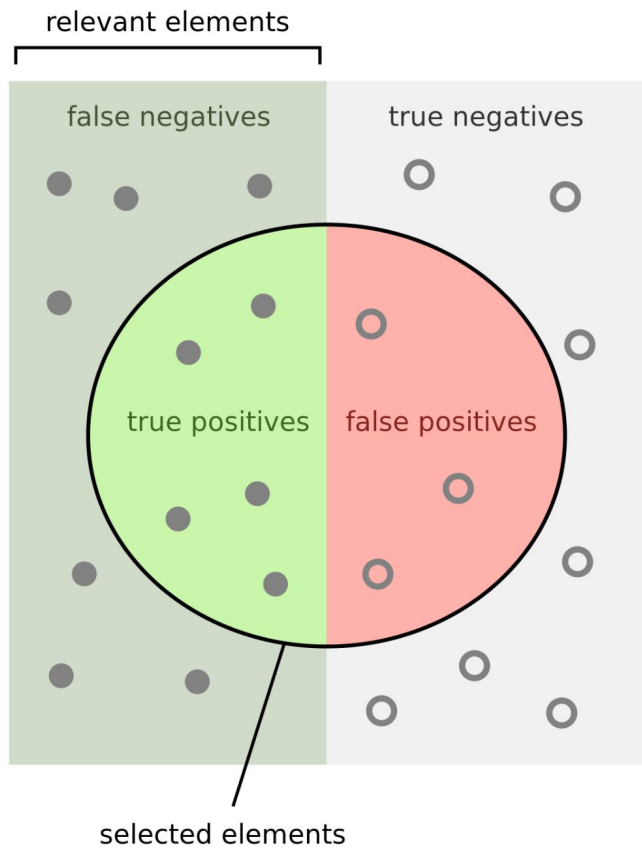
Precision:

- Of all the tweets that you predicted to be POS, what proportion was correct?
- *how many of the tweets that you said were toxic actually were?*

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Precision = $\frac{\text{TP}}{\text{TP} + \text{FP}}$

Precision/Recall



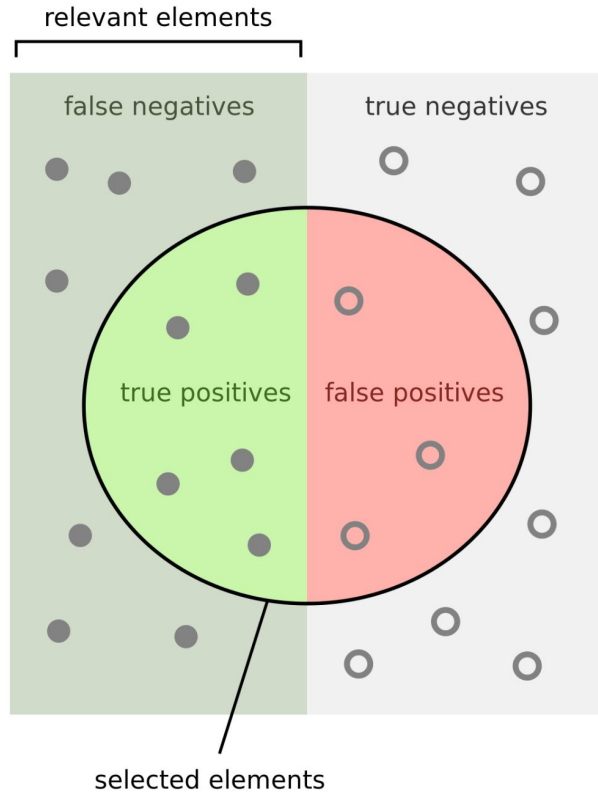
Recall

- Of all the POS tweets in the test set, how many did you recall correctly?
- *how many of the actually toxic tweets did you catch?*

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

In terms of the relevant category



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Precision & Recall

$$Precision = \frac{TP}{TP + FP}$$

		Predicted	
		Positive	Negative
Actual	Positive	True Positives Reality: toxic Prediction: "toxic" N = 1	False Negatives Reality: hate speech Prediction: "normal" N = 8
	Negative	False Positives Actual: normal Prediction: "hate speech" N = 1	True Negatives Reality: Normal Prediction: Benign N = 90

Precision & Recall

$$\textit{Precision} = \frac{TP}{TP + FP}$$
$$= \frac{1}{1 + 1} = \frac{1}{2} = .5$$

		Predicted	
		Positive	Negative
Actual	Positive	True Positives Reality: toxic Prediction: "toxic" N = 1	False Negatives Reality: hate speech Prediction: "toxic" N = 8
	Negative	False Positives Actual: normal Prediction: "hate speech" N = 1	True Negatives Reality: Normal Prediction: Benign N = 90

Precision & Recall

$$\begin{aligned} \textit{Precision} &= \frac{TP}{TP + FP} \\ &= \frac{1}{1 + 1} = \frac{1}{2} = .5 \end{aligned}$$

$$\textit{Recall} = \frac{TP}{TP + FN}$$

Actual

Predicted

		Predicted	
		Positive	Negative
Actual	Positive	True Positives Reality: toxic Prediction: "toxic" N = 1	False Negatives Reality: hate speech Prediction: "toxic" N = 8
	Negative	False Positives Actual: normal Prediction: "hate speech" N = 1	True Negatives Reality: Normal Prediction: Benign N = 90

Precision & Recall

$$\begin{aligned} \textit{Precision} &= \frac{TP}{TP + FP} \\ &= \frac{1}{1 + 1} = \frac{1}{2} = .5 \end{aligned}$$

$$\begin{aligned} \textit{Recall} &= \frac{TP}{TP + FN} \\ &= \frac{1}{1 + 8} = \frac{1}{9} = .11 \end{aligned}$$

Actual

		Predicted	
		Positive	Negative
Actual	Positive	True Positives Reality: toxic Prediction: "toxic" N = 1	False Negatives Reality: hate speech Prediction: "normal" N = 8
	Negative	False Positives Actual: normal Prediction: "hate speech" N = 1	True Negatives Reality: Normal Prediction: Benign N = 90

Precision / Recall Tradeoff

The Naive Bayes classifier gives us a probability between 0 and 1:

$$0 < P(\text{"toxic"}) < 1$$

When should we label a tweet toxic?

We could set the threshold higher or lower than 0.5

- What happens if we set it to 0.95?
- What happens if we set it to 0.2?

Selecting Metrics

There is not always one right way to evaluate a machine learning classifier. Depending on the situation, we might be more interested in a very precise model, or one with a high recall.

Selecting Metrics

[From Cassie Kozyrkov](#)

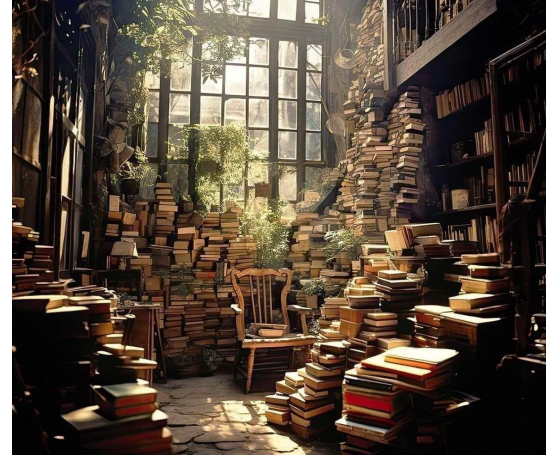


Selecting Metrics

From Cassie Kozyrkov

Precision vs. Recall Mindset

A system with high **precision** might leave some things out, but what it returns is of high quality.



A system of high **recall** might give you a lot of duds, but it also returns most of the good items



Extending Classification

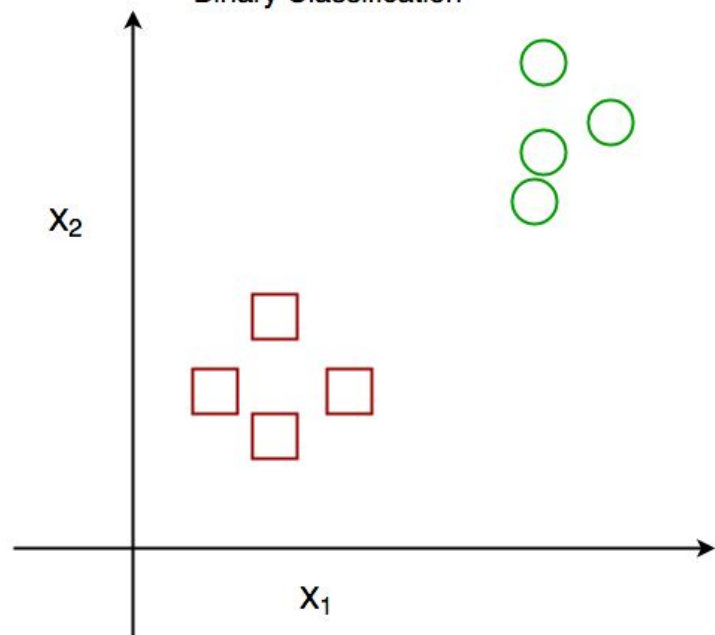
We've talked about binary classification;

- sentiment analysis,
- toxicity/hate speech
- spam

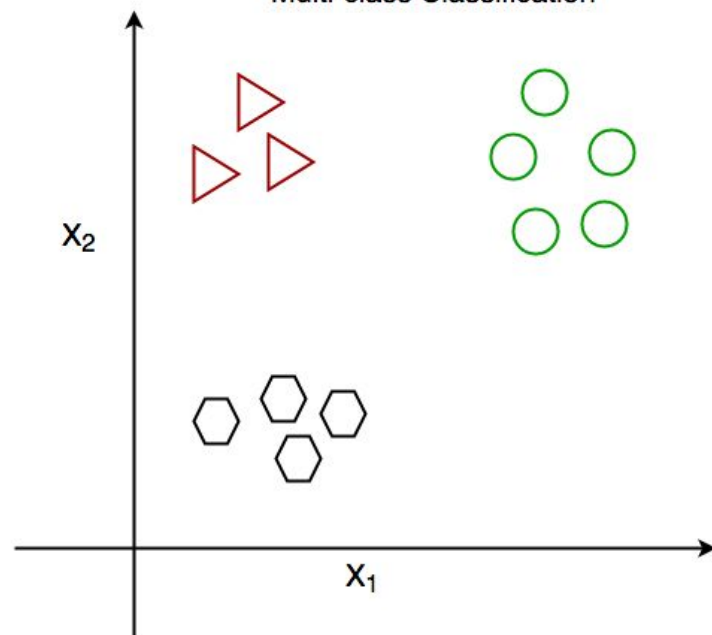
Sometimes there's more than one category

- news categorization
- email labeling
- named entities recognition
- intent recognition
 - Search type recognition (Calculation? Location? Song?)
 - Customer support queries
- emotion detection

Binary Classification



Multi-class Classification



Named Entity Recognition - ELMo

Model Output

Share

Entities

When I told John
PER that I wanted to move to Alaska
LOC , he warned me that I 'd have trouble
finding a Starbucks
ORG there .

Named Entity Recognition - ELMo

Model Output

Share

Entities

When I told Qwerty that I wanted to move to Alnoranalia , he warned me that I 'd have
trouble finding a Starbricks there .

Visual Question Answering - ViLBERT

Image



Question













what are the people doing?

Run Model

<https://demo.allennlp.org/visual-question-answering>

Visual Question Answering

Model Output

Score	Answer
 61.9%	traveling
 1.1%	riding bikes
 1.1%	waiting
 1.1%	2
 0.7%	working
 0.6%	4
 0.6%	watching
 0.5%	skateboarding
 0.5%	riding
 0.4%	walking
 0.4%	getting on bus
 0.4%	standing

Our decisions

1. What text will we use for the dataset?
 - a. our corpus? goosebumps? something else?
2. What classification scheme will we use?
 - a. binary? multi-class?
 - b. what are the labels?
3. What level of analysis?
 - a. are we annotating sentences? words
4. what does "annotation" look like?
 - a. yes/no? scale of 1-10?

Our restrictions

Our dataset will be relatively small. This isn't a problem for state of the art classifiers. But if we want the model to learn a pattern that means there has to be a pattern. So we will limit ourselves

- we need a fixed number of categories
 - binary is simplest
 - multi-class is more ambitious
- the categories should be pretty broad
- they need to be something we all understand (or think we understand)

R.L. STINE

Goosebumps



EGG MONSTERS FROM MARS

SCHOLASTIC

6

Thump, thump, THUMP.

I had to see what was happening in my dresser drawer.

Had the egg hatched? Was the turtle bumping up against the sides of the drawer, trying to climb out?

Was it a turtle?

Or was it something weird?

Suddenly I felt very afraid of it.

I took a deep breath and rose to my feet. My legs felt rubbery and weak as I made my way across the room. My mouth was suddenly as dry as cotton.

Thump, THUMP, thump.

I clicked on the light. Blinked several times, struggling to force my eyes to focus.

The steady thuds grew louder as I approached the dresser.

Heartbeats, I told myself.

Heartbeats of the creature inside the egg.

I grabbed the drawer handles with both hands. Took another deep breath.

Dana, this is your last chance to run away, I warned myself.

This is your last chance to leave the drawer safely closed.

Thump, thump, thump, thump, thump.

I tugged open the drawer and peered inside.

I stared in, amazed that nothing had changed. The egg sat exactly where I had left it. The blue-and-purple veins along the shell pulsed as before.

Feeling a little calmer, I picked it up.

"Ouch!"

I nearly dropped it. The shell was burning hot.

I cupped it in my hands and blew on it. "This is so totally weird," I murmured to myself.

Mom and Dad have to see it, I decided. Right now. Maybe they can tell me what it is.

They were still awake. I could hear them talking in their room down the hall.

I carried the egg carefully, cradling it in both hands. I had to knock on their door with my elbow. "It's me," I said.

"Dana, what is it?" Dad demanded grumpily. "It's been a long day. We're all very tired."

I pushed open their door a crack. "I have an egg I want to show you," I started.

"No eggs!" they both cried at once.

"Haven't we seen enough eggs for one day?" Mom griped.

SPOOKY



SCARY

R.L. STINE

Goosebumps



EGG MONSTERS FROM MARS

SCHOLASTIC

6

Thump, thump, THUMP.

I had to see what was happening in my dresser drawer.

Had the egg hatched? Was the turtle bumping up against the sides of the drawer, trying to climb out?

Was it a turtle?

Or was it something weird?

Suddenly I felt very afraid of it.

I took a deep breath and rose to my feet. My legs felt rubbery and weak as I made my way across the room. My mouth was suddenly as dry as cotton.

Thump, THUMP, thump.

I clicked on the light. Blinked several times, struggling to force my eyes to focus.

The steady thuds grew louder as I approached the dresser.

Heartbeats, I told myself.

Heartbeats of the creature inside the egg.

I grabbed the drawer handles with both hands. Took another deep breath.

Dana, this is your last chance to run away, I warned myself.

This is your last chance to leave the drawer safely closed.

Thump, thump, thump, thump, thump.

I tugged open the drawer and peered inside.

I stared in, amazed that nothing had changed. The egg sat exactly where I had left it. The blue-and-purple veins along the shell pulsed as before.

Feeling a little calmer, I picked it up.

"Ouch!"

I nearly dropped it. The shell was burning hot.

I cupped it in my hands and blew on it. "This is so totally weird," I murmured to myself.

Mom and Dad have to see it, I decided. Right now. Maybe they can tell me what it is.

They were still awake. I could hear them talking in their room down the hall.

I carried the egg carefully, cradling it in both hands. I had to knock on their door with my elbow. "It's me," I said.

"Dana, what is it?" Dad demanded grumpily. "It's been a long day. We're all very tired."

I pushed open their door a crack. "I have an egg I want to show you," I started.

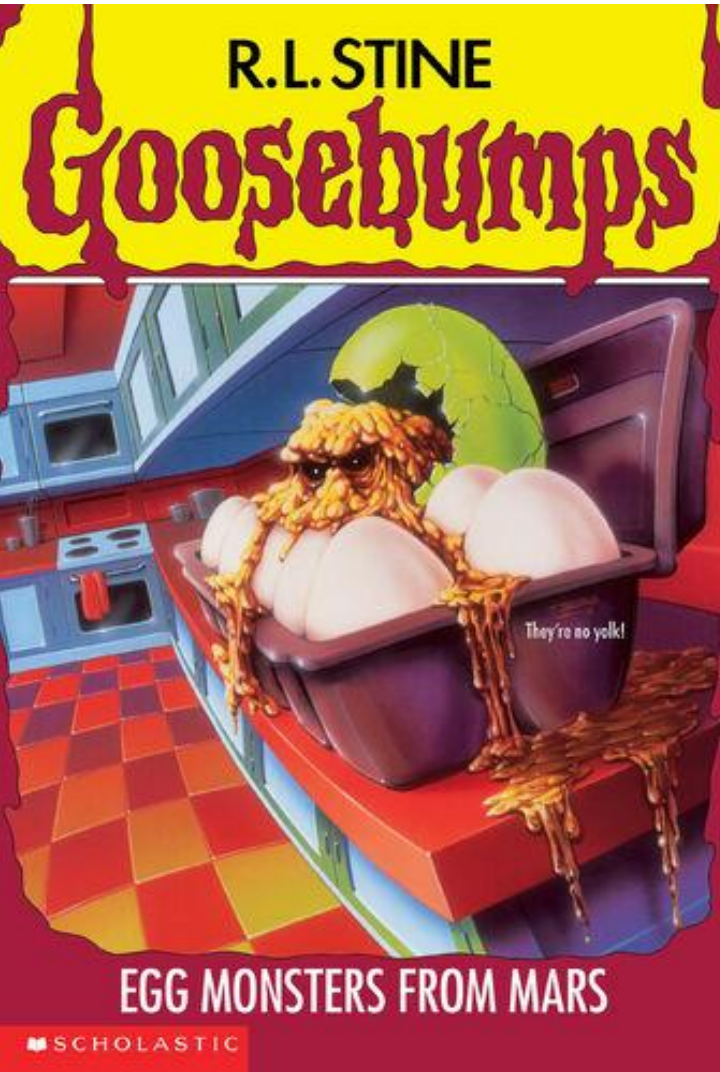
"No eggs!" they both cried at once.

"Haven't we seen enough eggs for one day?" Mom griped.

SPOOKY



SCARY



Thump, thump, THUMP.

I had to see what was happening in my dresser drawer.

Had the egg hatched? Was the turtle bumping up against the sides of the drawer, trying to climb out?

Was it a turtle?

Or was it something weird?

Suddenly I felt very afraid of it.

I took a deep breath and rose to my feet. My legs felt rubbery and weak as I made my way across the room. My mouth was suddenly as dry as cotton.

Thump, THUMP, thump.

I clicked on the light. Blinked several times, struggling to force my eyes to focus.

The steady thuds grew louder as I approached the dresser.

Heartbeats, I told myself.

Heartbeats of the creature inside the egg.

I grabbed the drawer handles with both hands. Took another deep breath.

Dana, this is your last chance to run away, I warned myself.

This is your last chance to leave the drawer safely closed.

Thump, thump, thump, thump, thump.

I tugged open the drawer and peered inside.

I stared in, amazed that nothing had changed. The egg sat exactly where I had left it. The blue-and-purple veins along the shell pulsed as before.

Feeling a little calmer, I picked it up.

"Ouch!"

I nearly dropped it. The shell was burning hot.

I cupped it in my hands and blew on it. "This is so totally weird," I murmured to myself.

Mom and Dad have to see it, I decided. Right now. Maybe they can tell me what it is.

They were still awake. I could hear them talking in their room down the hall.

I carried the egg carefully, cradling it in both hands. I had to knock on their door with my elbow. "It's me," I said.

"Dana, what is it?" Dad demanded grumpily. "It's been a long day. We're all very tired."

I pushed open their door a crack. "I have an egg I want to show you," I started.

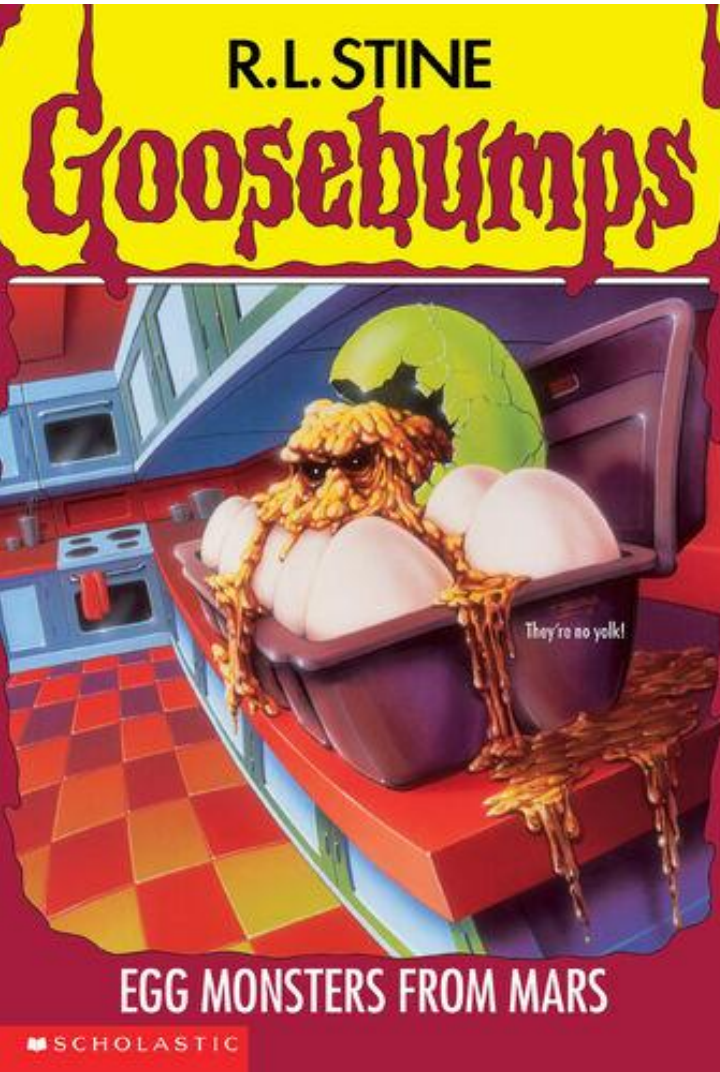
"No eggs!" they both cried at once.

"Haven't we seen enough eggs for one day?" Mom griped.

SPOOKY

creepy?

SCARY



6

Thump, thump, THUMP.

I had to see what was happening in my dresser drawer.

Had the egg hatched? Was the turtle bumping up against the sides of the drawer, trying to climb out?

Was it a turtle?

Or was it something weird?

Suddenly I felt very afraid of it.

I took a deep breath and rose to my feet. My legs felt rubbery and weak as I made my way across the room. My mouth was suddenly as dry as cotton.

Thump, THUMP, thump.

I clicked on the light. Blinked several times, struggling to force my eyes to focus.

The steady thuds grew louder as I approached the dresser.

Heartbeats, I told myself.

Heartbeats of the creature inside the egg.

I grabbed the drawer handles with both hands. Took another deep breath.

Dana, this is your last chance to run away, I warned myself.

This is your last chance to leave the drawer safely closed.

Thump, thump, thump, thump, thump.

I tugged open the drawer and peered inside.

I stared in, amazed that nothing had changed. The egg sat exactly where I had left it. The blue-and-purple veins along the shell pulsed as before.

Feeling a little calmer, I picked it up.

"Ouch!"

I nearly dropped it. The shell was burning hot.

I cupped it in my hands and blew on it. "This is so totally weird," I murmured to myself.

Mom and Dad have to see it, I decided. Right now. Maybe they can tell me what it is.

They were still awake. I could hear them talking in their room down the hall.

I carried the egg carefully, cradling it in both hands. I had to knock on their door with my elbow. "It's me," I said.

"Dana, what is it?" Dad demanded grumpily. "It's been a long day. We're all very tired."

I pushed open their door a crack. "I have an egg I want to show you," I started.

"No eggs!" they both cried at once.

"Haven't we seen enough eggs for one day?" Mom griped.

Things
your mom
would text



Things your
best friend
would text