Machine Learning Part 2: Evaluation

LIN 313 Language and Computers

UT Austin Fall 2025

Overview

- Bayes Theorem
- Naive Bayes again
- Evaluation
  - precision + recall

Beliefs and Evidence

Consider Steve:

(from an experiment by Daniel Kahneman and Amos Tversky)

Steve is very shy and withdrawn, invariably helpful but with very little interest in people or the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.

Beliefs and Evidence

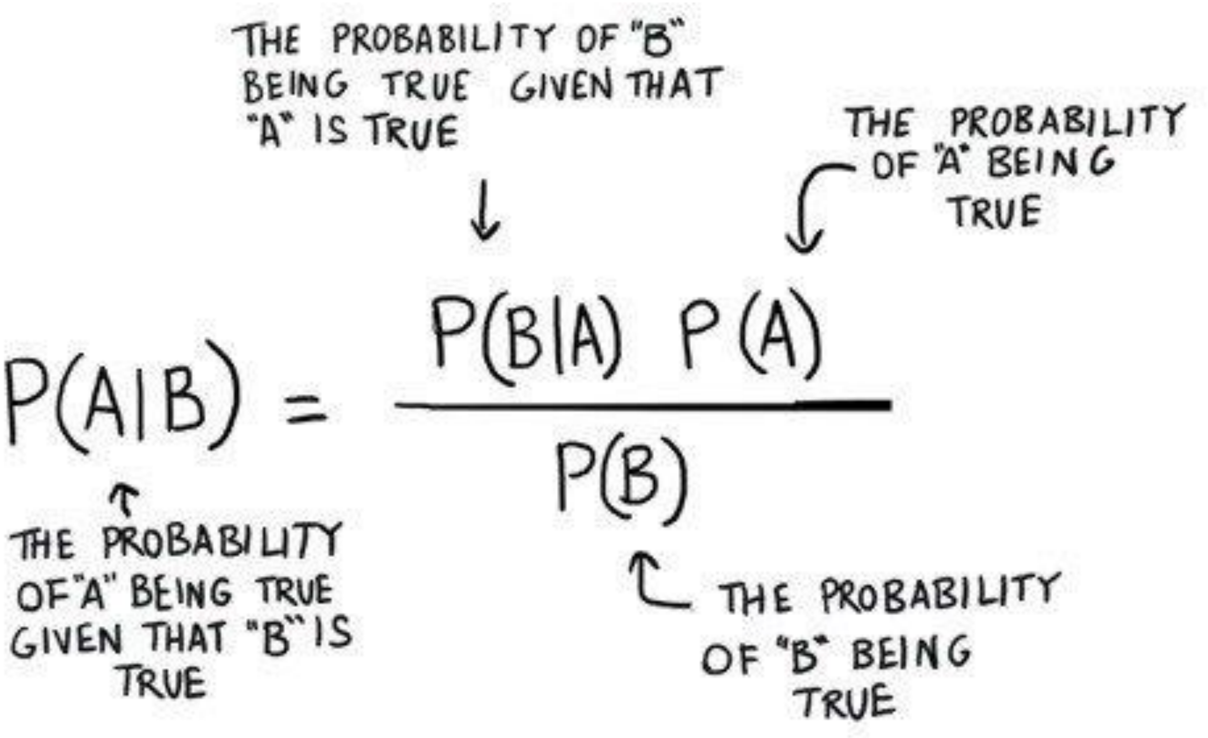
Consider Steve:

(from an experiment by Daniel Kahneman and Amos Tversky)

Steve is very shy and withdrawn, invariably helpful but with very little interest in people or the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.

Is Steve more likely to be a farmer or a librarian?

Bayes' Theorem to the Rescue



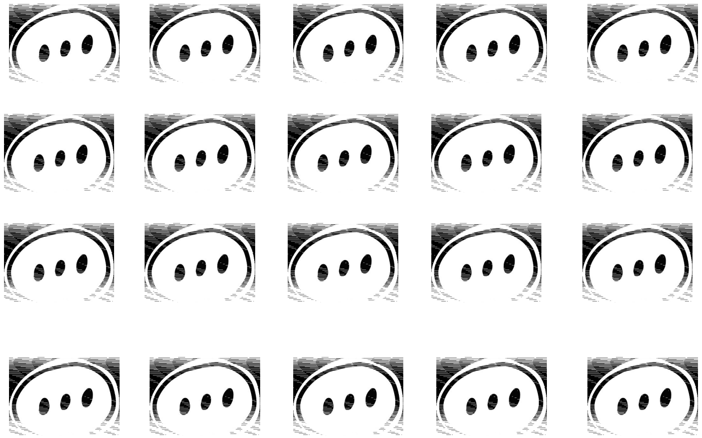
https://medium.com/@chonghankhai/bayesian-thinking-in-everyday-life-bf82fe2ab0af

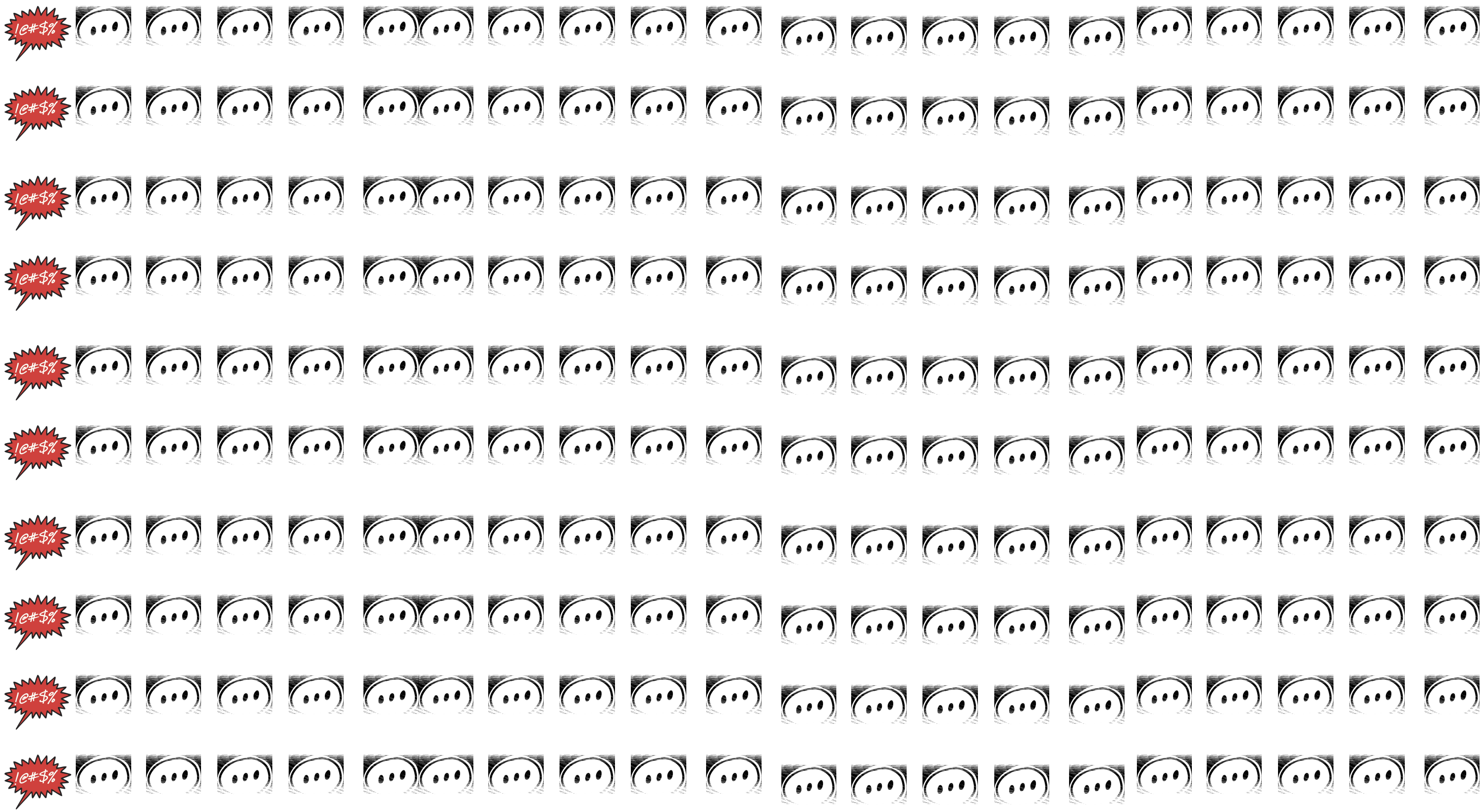
I have a tweet that uses the word "dummy". Based on this **evidence**, is it more likely to be "toxic" or normal text?

(https://www.youtube.com/watch?v=HZGCoVF3YvM goes through this same example but with farmers and librarians—great study video)

I have a tweet that uses the word "dummy". Based on this **evidence**, is it more likely to be "toxic" or normal text?

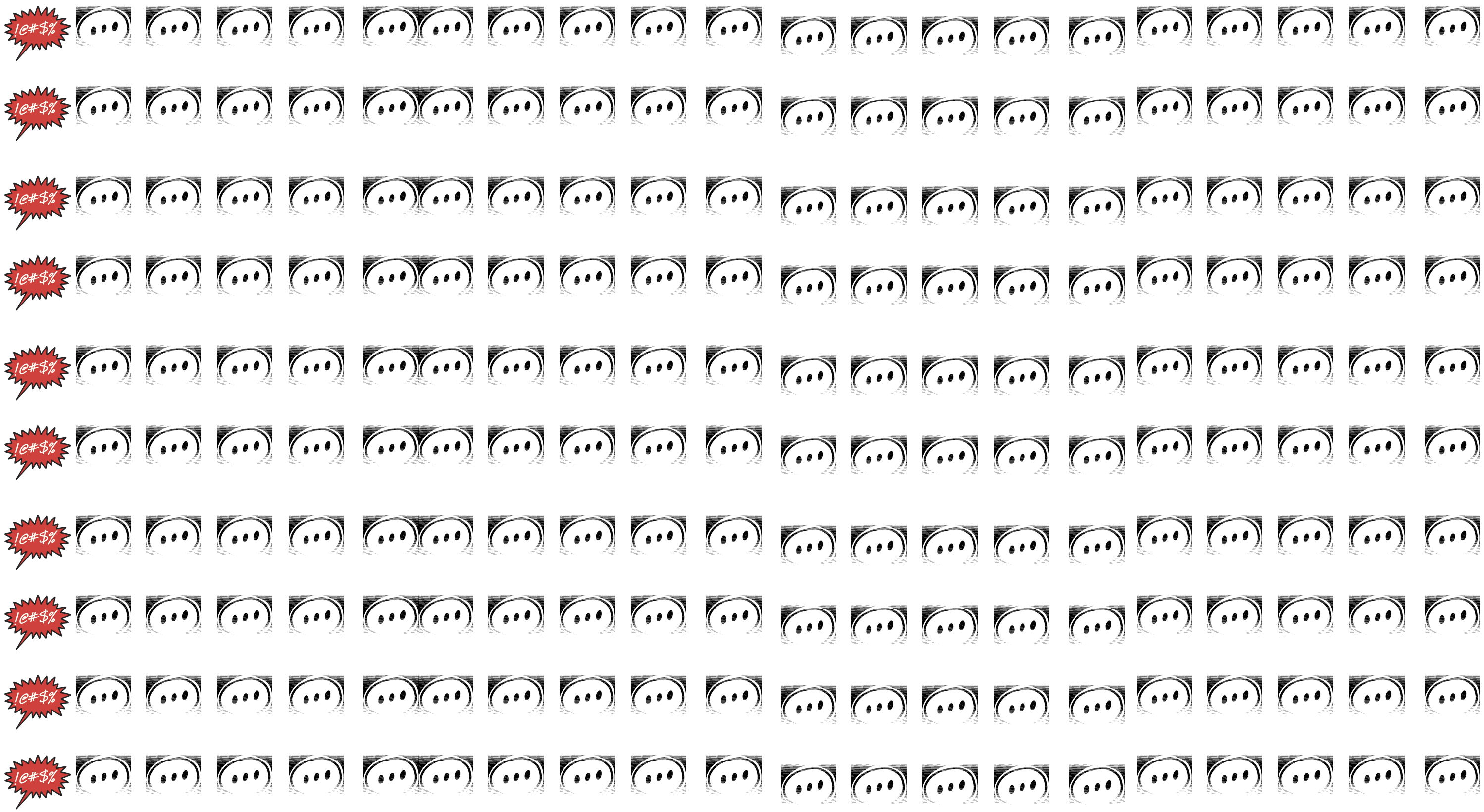
What if the ratio of normal to toxic is 20:1?





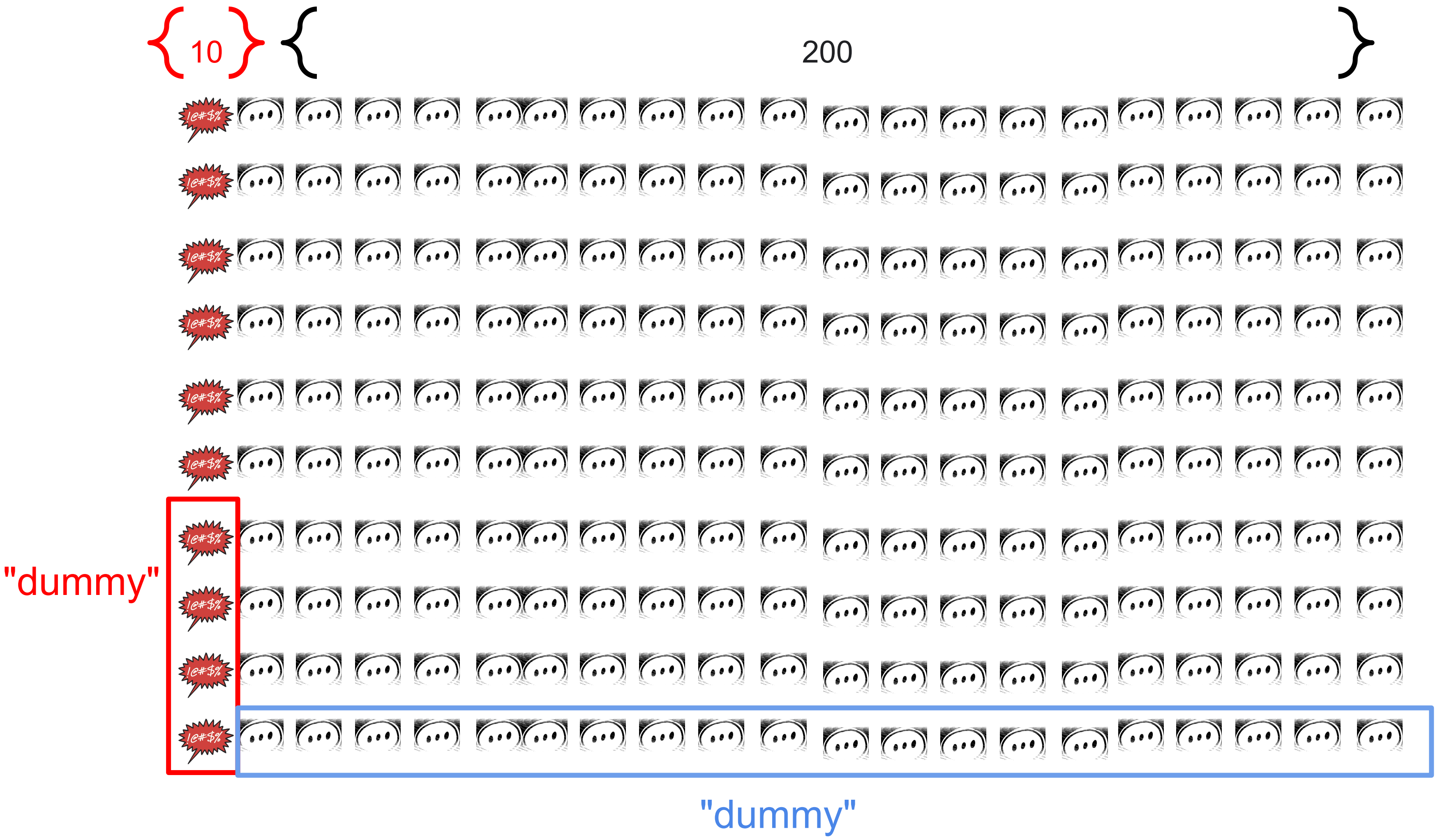
200 regular comments

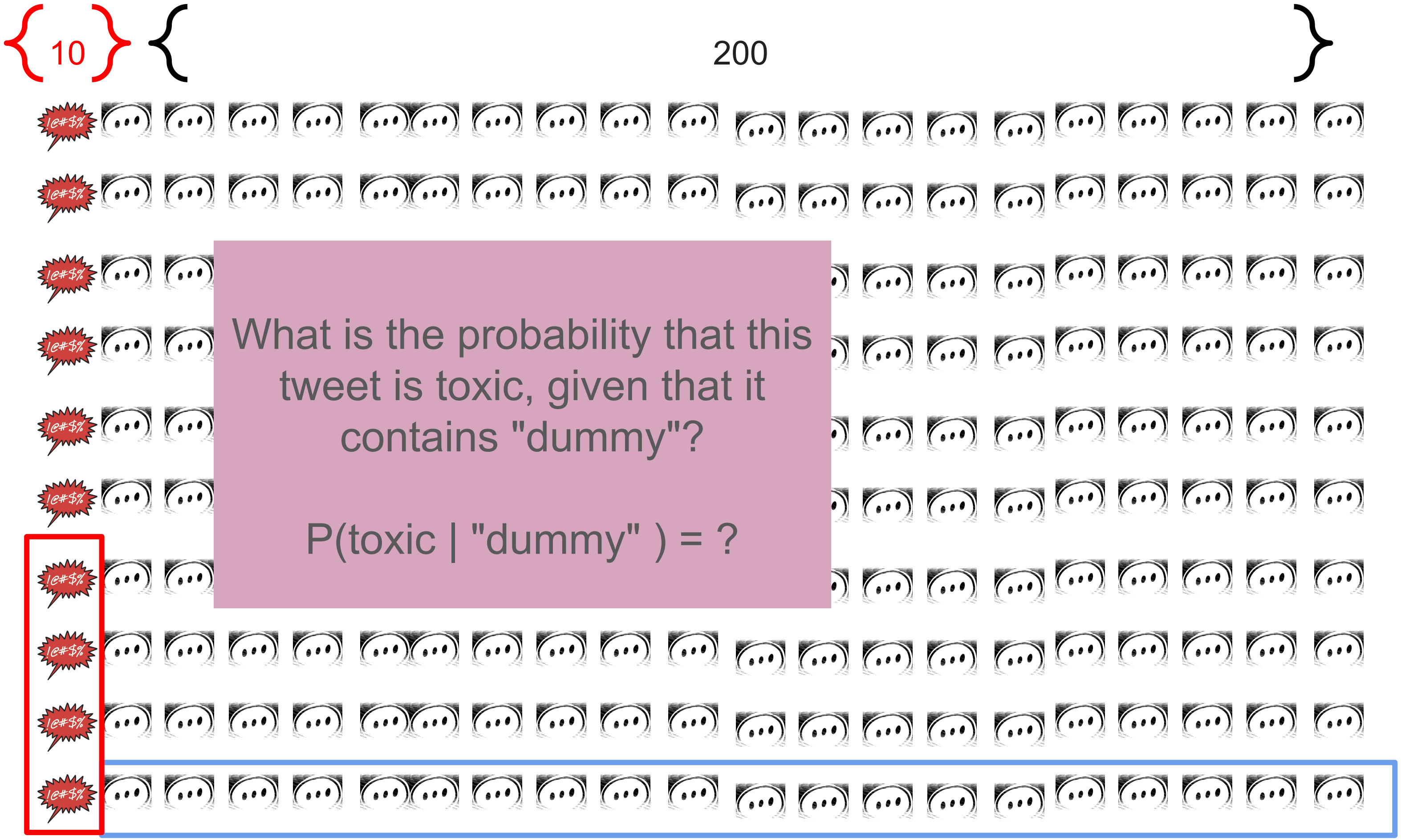
10 "toxic" comments



In our dataset,

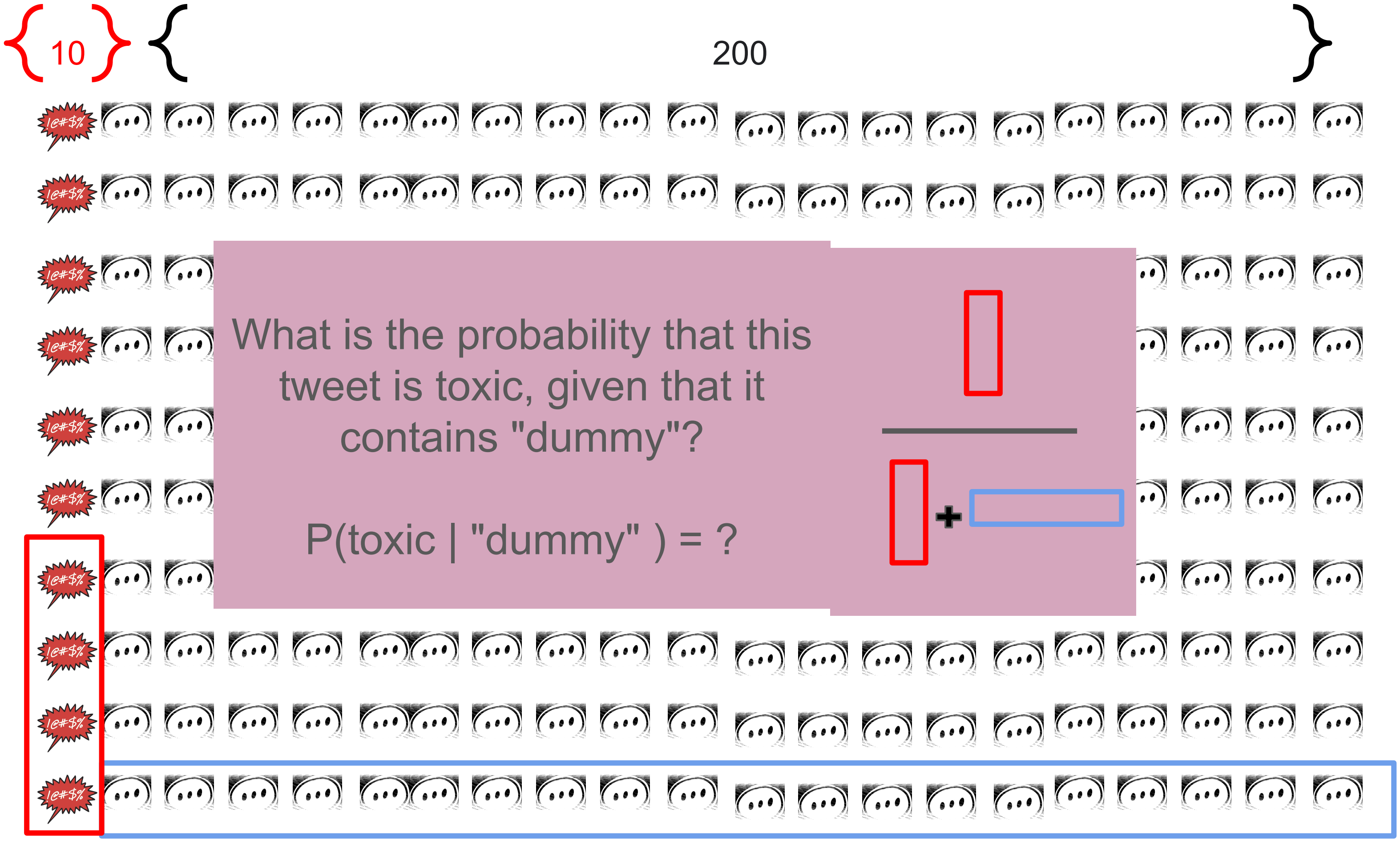
- **40%** of toxic tweets contain "**dummy**"
- **10%** of regular tweets contain dummy

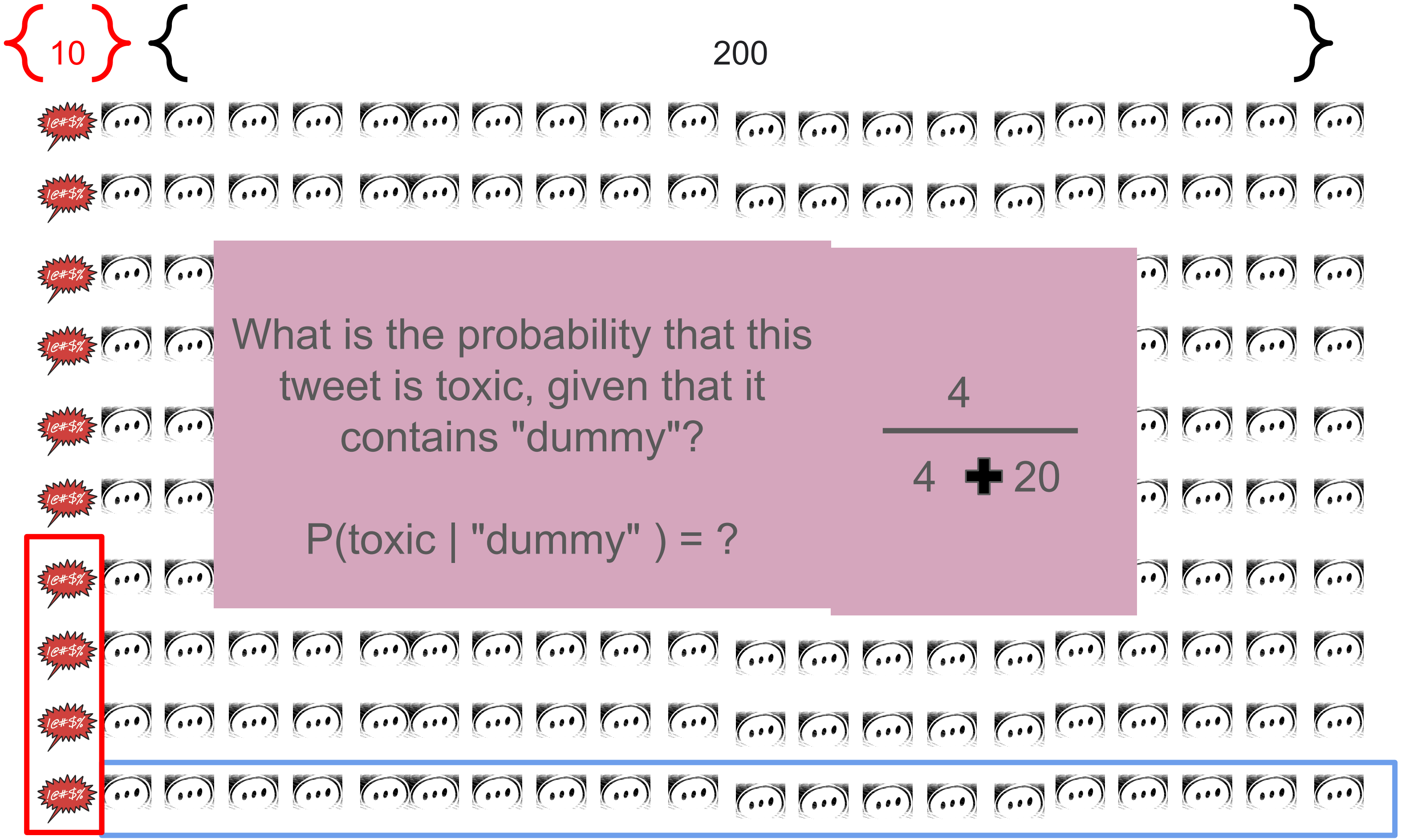

10
200
"dummy"
"dummy"



What is the probability that this tweet is toxic, given that it contains "dummy"?

P(toxic | "dummy") = ?

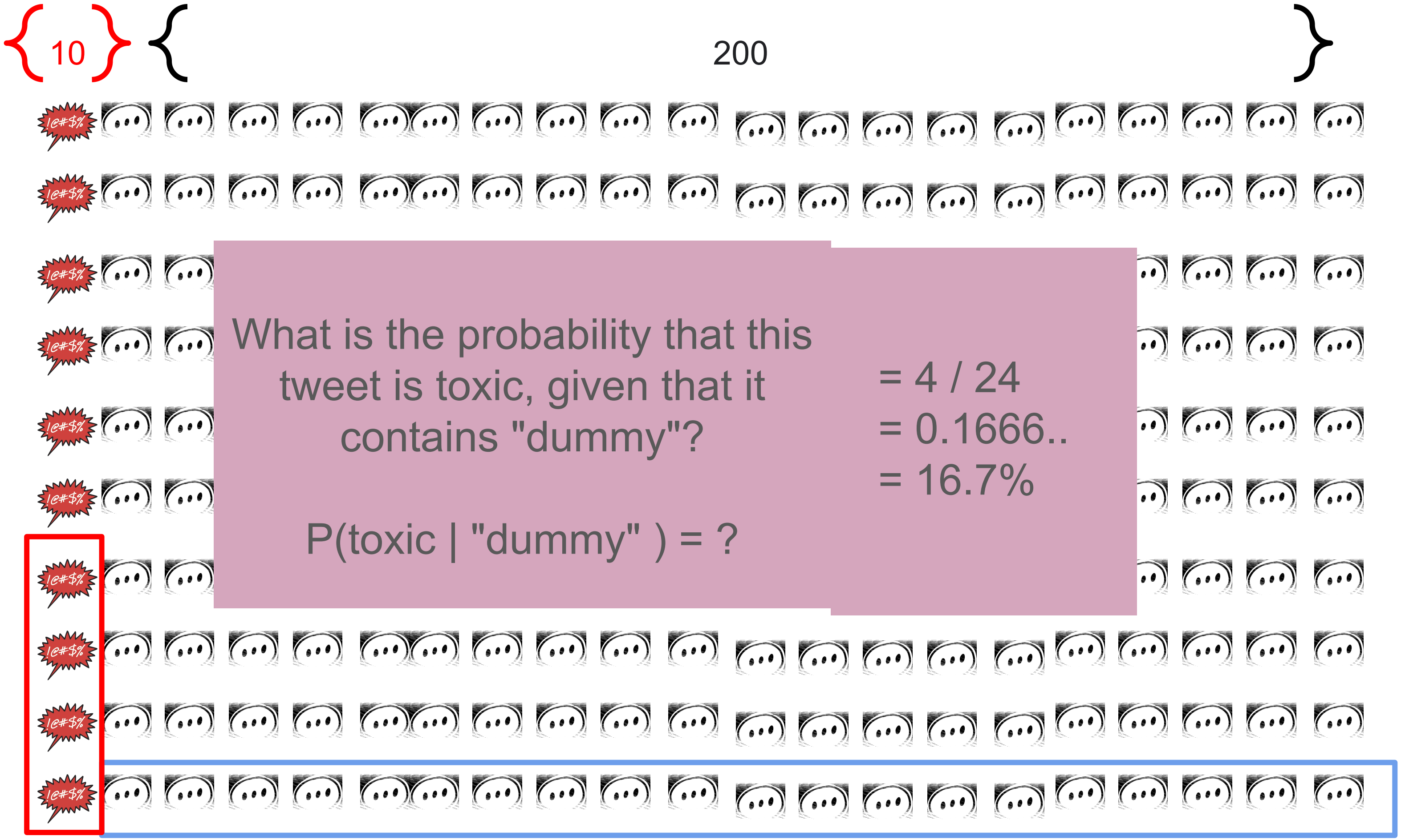

10
200
What is the probability that this tweet is toxic, given that it contains "dummy"?
P(toxic | "dummy") = ?



What is the probability that this tweet is toxic, given that it contains "dummy"?

P(toxic | "dummy") = ?

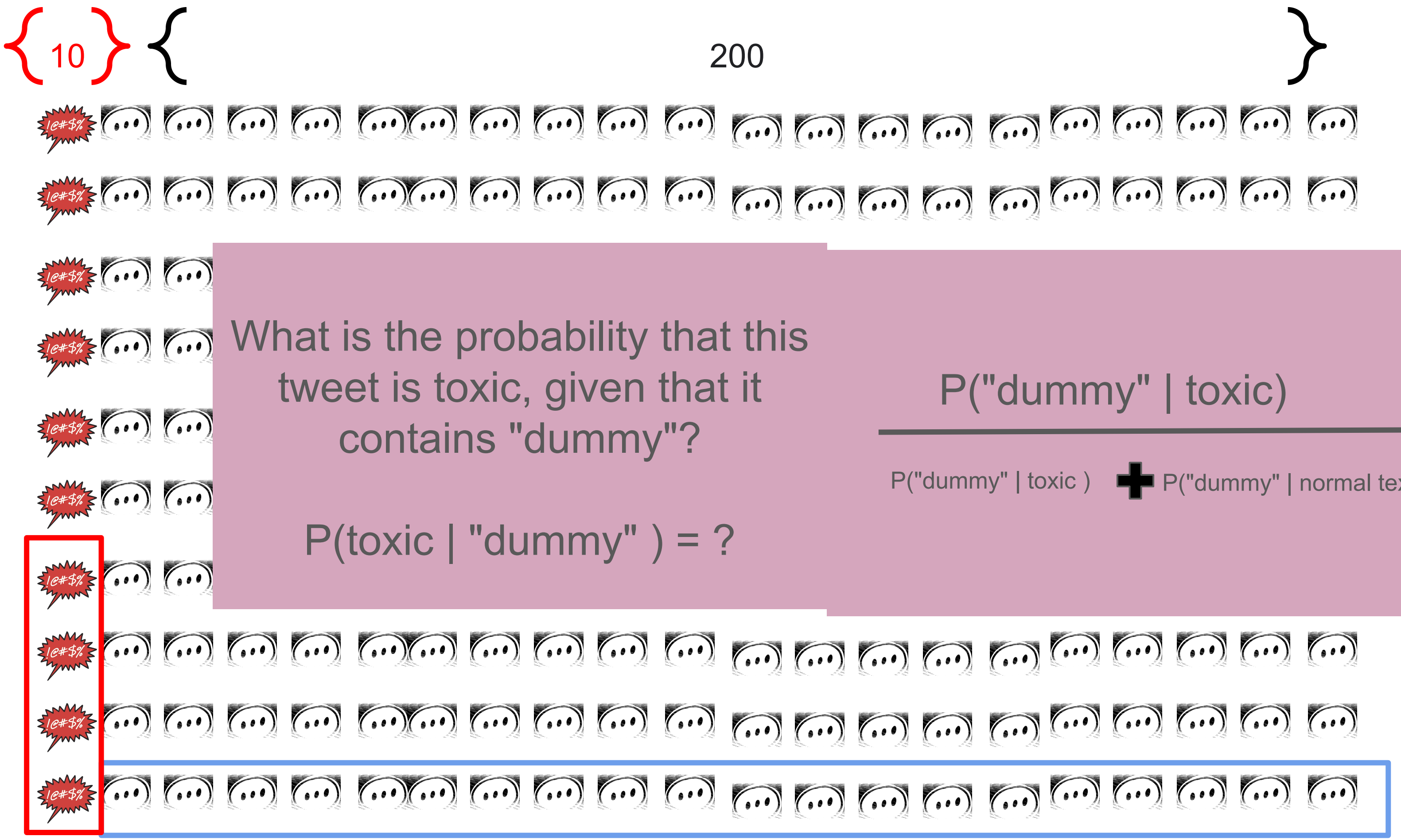
$$\frac{4}{4 + 20}$$



What is the probability that this tweet is toxic, given that it contains "dummy"?

P(toxic | "dummy") = ?

= 4 / 24
= 0.1666..
= 16.7%



What is the probability that this tweet is toxic, given that it contains "dummy"?

P(toxic | "dummy") = ?

$$\frac{P(\text{"dummy"} \mid \text{toxic})}{P(\text{"dummy"} \mid \text{toxic}) + P(\text{"dummy"} \mid \text{normal text})}$$

Updating beliefs

Prior: basic belief that a tweet is toxic = ?

Posterior : basic belief that a tweet is toxic after seeing the word "dummy" = ?

Updating beliefs

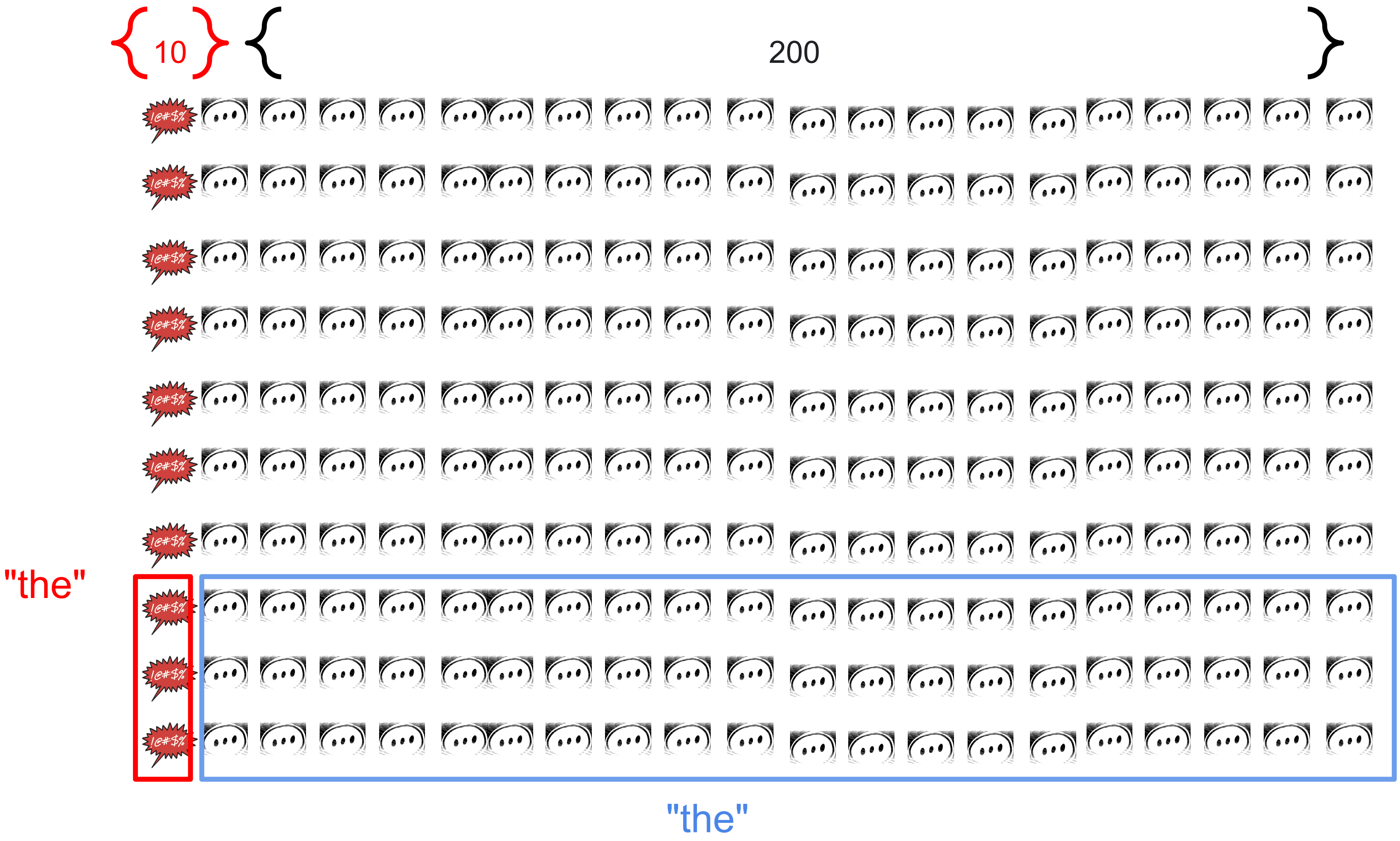
Prior: basic belief that a tweet is toxic

$$P(\text{toxic}) = 10 / 210 = 0.047 = 4.7\ \%$$

Posterior : basic belief that a tweet is toxic after seeing the word "dummy" = ?

$$P(\text{toxic} \mid \text{"dummy"}) = 4 / 24 = 0.16666 = 16.7\%$$

I've updated my beliefs about the hypothesis given new evidence!!!


10
200
"the"
"the"

$$P(H) = \text{Probability a hypothesis is true (before any evidence)}$$

$$P(E|H) = \text{Probability of seeing the evidence if the hypothesis is true}$$

$$P(E) = \text{Probability of seeing the evidence}$$

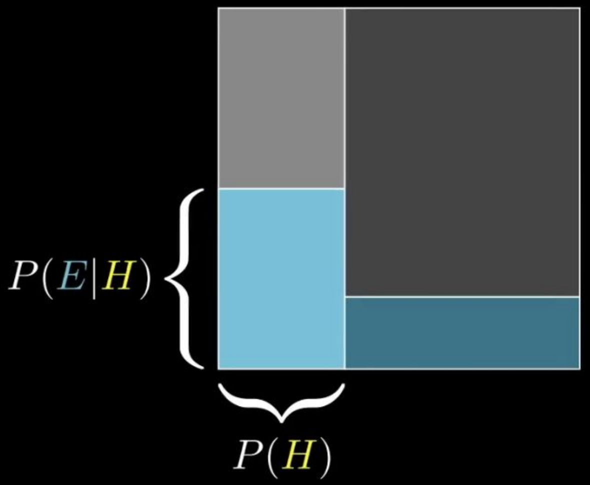
$$P(H|E) = \text{Probability a hypothesis is true given some evidence}$$

$$P(H) =$$ Probability a hypothesis is true (before any evidence)

$$P(E|H) =$$ Probability of seeing the evidence if the hypothesis is true

$$P(E) =$$ Probability of seeing the evidence

$$P(H|E) =$$ Probability a hypothesis is true given some evidence

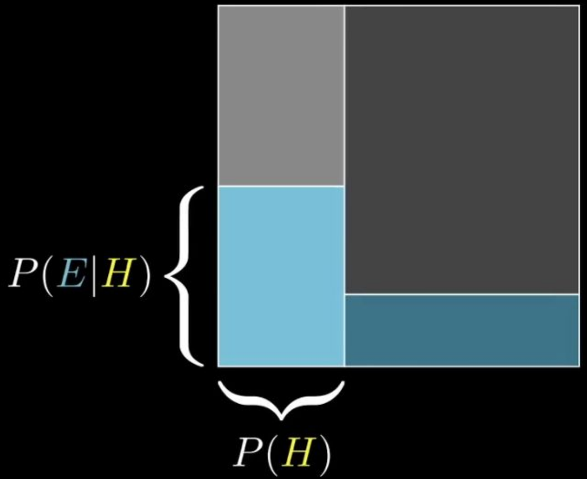


$$P(H) =$$ Probability a hypothesis is true (before any evidence)

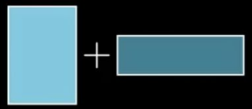
$$P(E|H) =$$ Probability of seeing the evidence if the hypothesis is true

$$P(E) =$$ Probability of seeing the evidence

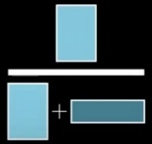
$$P(H|E) =$$ Probability a hypothesis is true given some evidence



$$P(E) =$$



$$P(H|E) =$$



$$P(H|E) = \frac{P(H) \cdot P(E|H)}{P(E)}$$

$$P(H|E) = \frac{P(H) \cdot P(E|H)}{P(E)}$$

What is the probability that this tweet is toxic, given that it contains "dummy"?

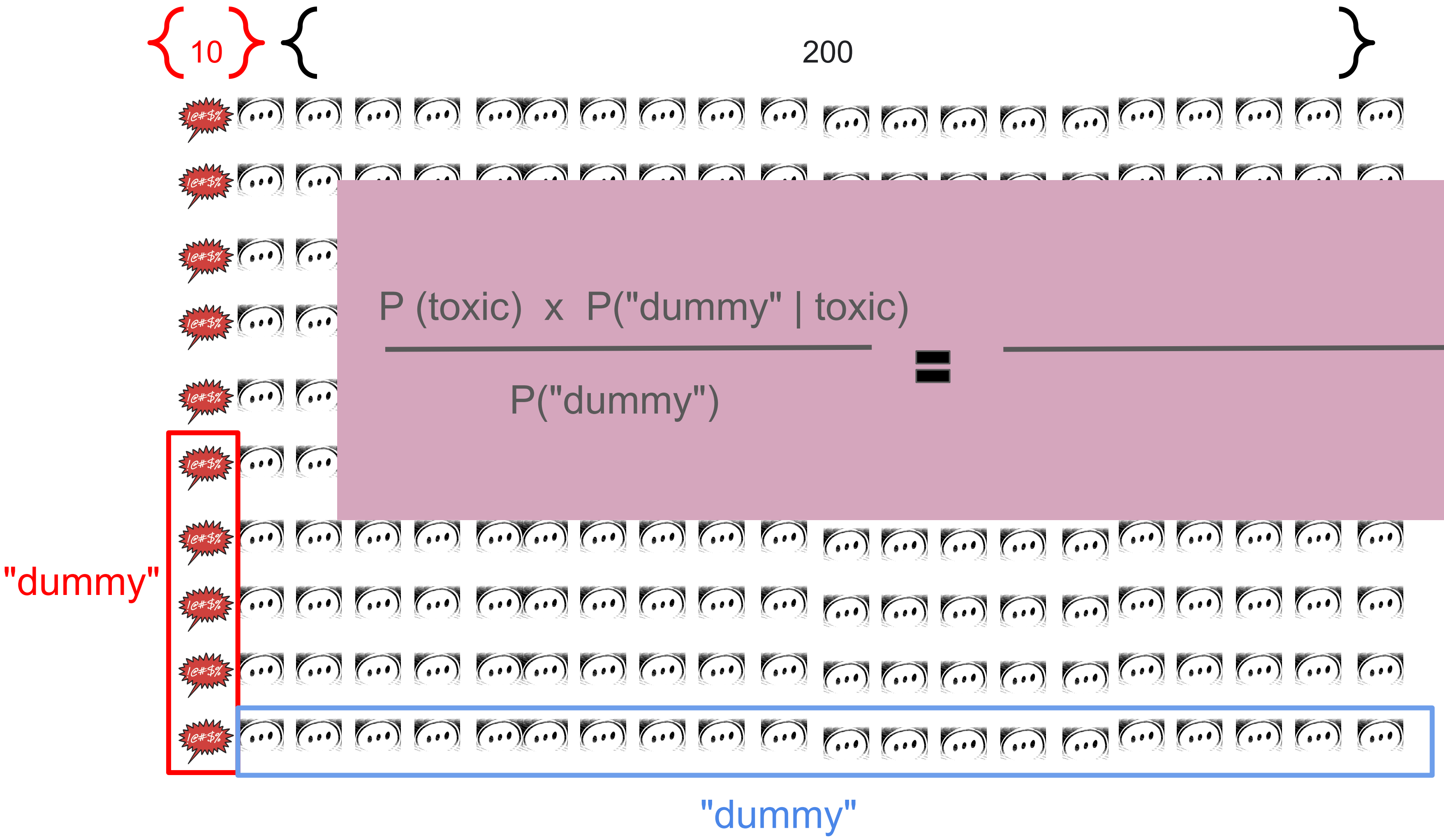
P("toxic" | dummy) = ?

$$P(H|E) = \frac{P(H) \cdot P(E|H)}{P(E)}$$

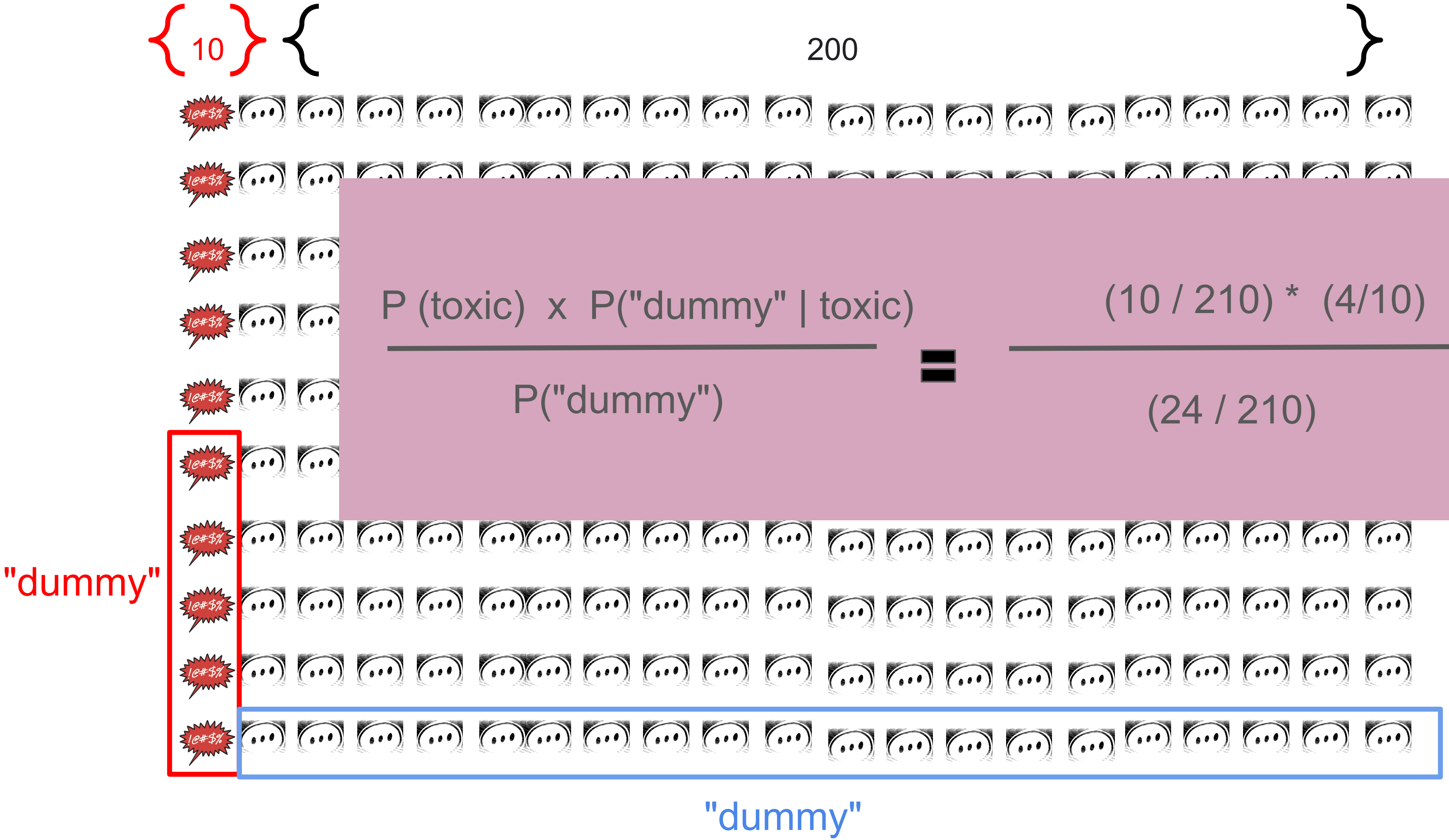
What is the probability that this tweet is toxic, given that it contains "dummy"?

P("toxic" | dummy) = ?

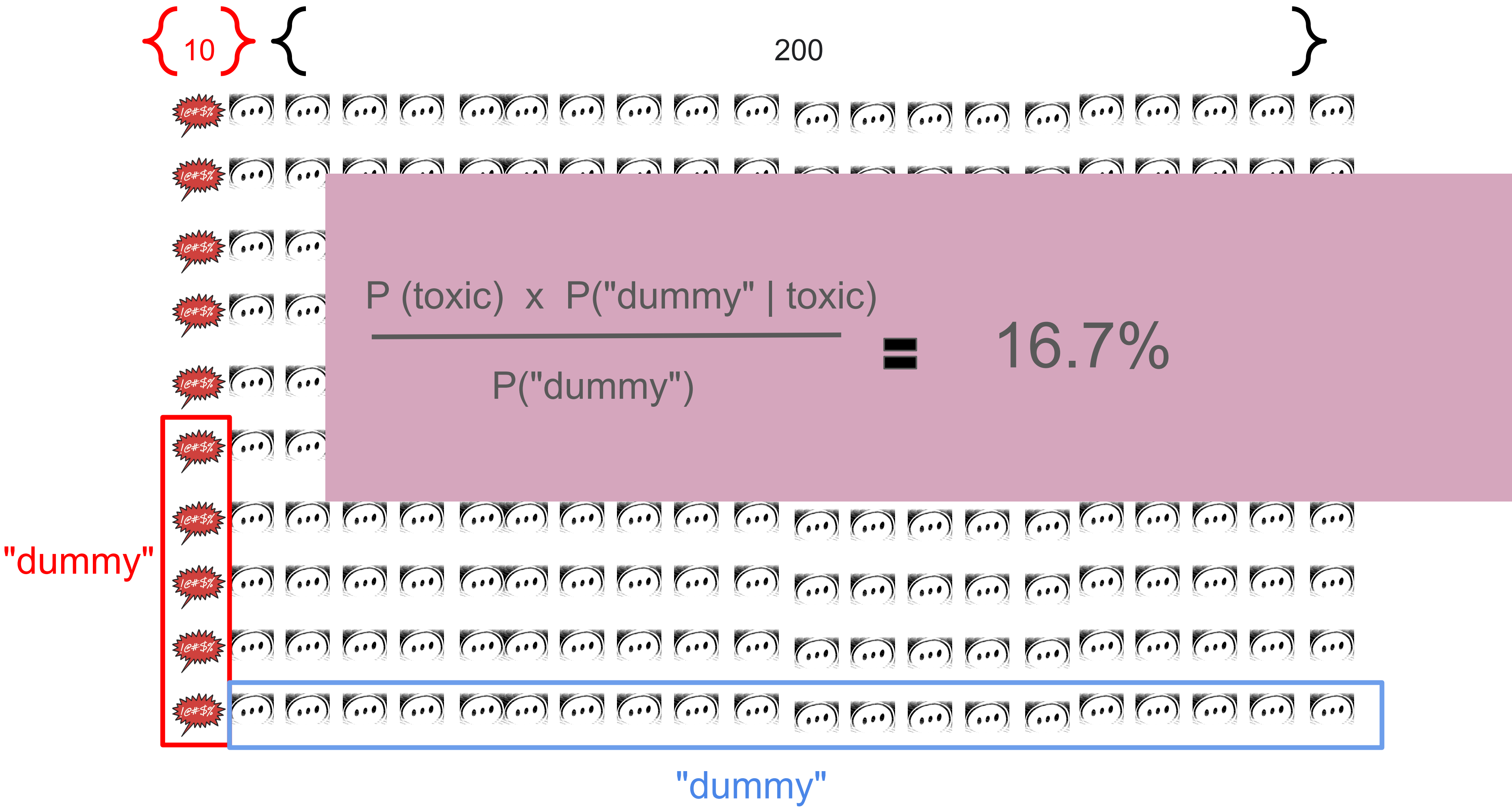
$$\frac{\text{P (toxic) x P("dummy" | toxic)}}{\text{P("dummy")}}$$



$$\frac{P(\text{toxic}) \times P(\text{"dummy"} \mid \text{toxic})}{P(\text{"dummy"})} = \frac{\qquad}{\qquad}$$



$$\frac{P\,(\text{toxic}) \;\; x \;\; P(\text{"dummy"} \mid \text{toxic})}{P(\text{"dummy"})} = \frac{(10 / 210) * \;\; (4/10)}{(24 / 210)}$$


10
200
!@#$%
"dummy"
"dummy"
P (toxic) x P("dummy" | toxic)
P("dummy")
= 16.7%

$$P(H|E) = \frac{P(H) \cdot P(E|H)}{P(E)}$$

What is the probability that this tweet is toxic, given that it contains "dummy"?

P("toxic" | dummy) = ?

$$\frac{P\text{ (toxic) } \times \text{ P("dummy" | toxic)}}{\text{P("dummy")}} = 16.7\%$$

posterior

prior

likelihood

$$P(H|E) = \frac{P(H) \cdot P(E|H)}{P(E)}$$

What is the probability that this tweet is toxic, given that it contains "dummy"?

P("toxic" | dummy) = ?

$$\frac{\text{P (toxic) x P("dummy" | toxic)}}{\text{P("dummy")}} = 16.7\%$$

Adding more features

If we are actually building a classifier, we want to take into account more than just seeing "dummy".

What are other features we could use?

Adding more features

If we are actually building a classifier, we want to take into account more than just seeing "dummy".

What are other features we could use?

- E1 = "dummy"
- E2 = "!"
- ……

Bayes with more features (multiple pieces of evidence)

E = E1, E2

Evidence is now a complex term

P(H | E) = P(E | H) x P(H) / P(E)

How do we apply Bayes Rule?

Bayes with more features (multiple pieces of evidence)

E = E1, E2

P(H | E) = P(E | H) x P(H) / P(E)

Just substitute!

P(H | E1, E2) = P(E1, E2 | H) × P(H) / P(E1, E2)

Bayes with multiple pieces of evidence

E = E1, E2

P(H | E) = P(E | H) x P(H) / P(E)

Just substitute!

What do we do with these **joint probabilities**?

another **joint probability**

P(H | E1, E2) = P(E1, E2 | H) × P(H) / P(E1, E2)

Remember the Naive in Naive Bayes

independence assumption: we assume that the features in our model don't depend on one another at all. The probabilities of "the" and "dummy" and "!" are totally independente

Remember the Naive in Naive Bayes

independence assumption: we assume that the features in our model don't depend on one another at all. The probabilities of "the" and "dummy" and "!" are totally independente

$$P(a, b) = P(a) \times P(b)$$

This is how we break up a joint probability

Joint probability vs conditional probability

Joint Probability:

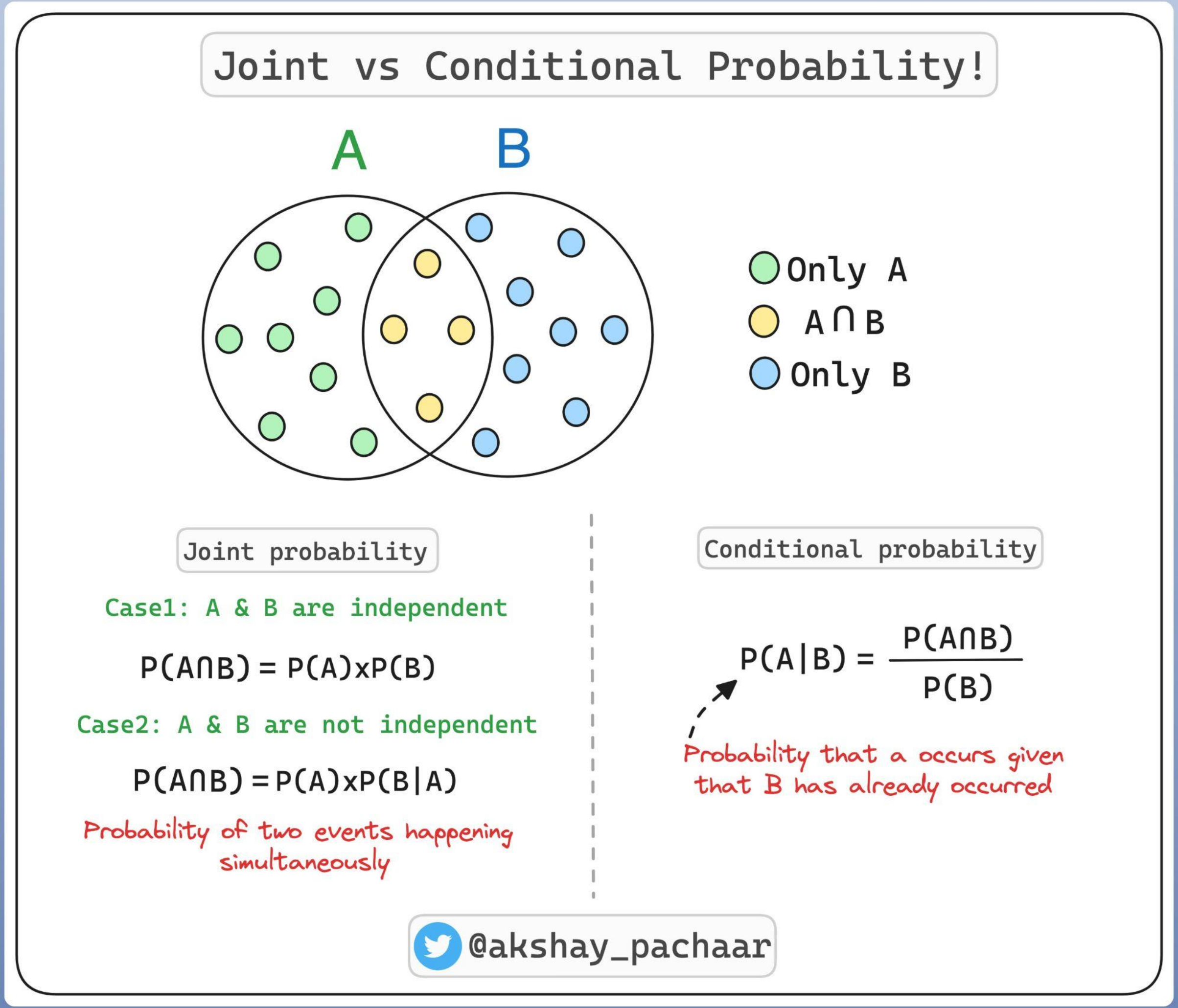
P(A,B)

Probability of A **and** B

Conditional Probability:

P(A | B)

Probability of A **given** B



Remember the Naive in Naive Bayes

independence assumption: we assume that the features in our model don't depend on one another at all. The probabilities of "the" and "dummy" and "!" are totally independente

How do we break these joint probabilities up?

P(a, b) = P(a) x P(b)

P(E1, E2) =

P(E1, E2 | H) =

Remember the Naive in Naive Bayes

independence assumption: we assume that the features in our model don't depend on one another at all. The probabilities of "the" and "dummy" and "!" are totally independente

$$P(\ a,\ b\ ) = P(a) \times P(b)$$

$$P(\ E1,\ E2\ ) = P(\ E1) \times P(E2)$$

conditional independence assumption

$$P(E1,\ E2 \mid H\ ) =$$

Remember the Naive in Naive Bayes

independence assumption: we assume that the features in our model don't depend on one another at all. The probabilities of "the" and "dummy" and "!" are totally independente

$$P(a, b) = P(a) \times P(b)$$

$$P(E1, E2) = P(E1) \times P(E2)$$

conditional independence assumption

$$P(E1, E2 \mid H) = P(E1 \mid H) \times P(E2 \mid H)$$

Bayes with more features (multiple pieces of evidence)

E	= "dummy", "!"
H	= toxic
P(H | E)	= P(E | H) x P(H) / P(E)

Bayes with more features (multiple pieces of evidence)

E	= "dummy", "!"
H	= toxic
P(H | E)	= P(E | H) x P(H) / P(E)

P(toxic | features)

Bayes with more features (multiple pieces of evidence)

E	= "dummy", "!"
H	= toxic
P(H | E)	= P(E | H) x P(H) / P(E)

$$P(\text{toxic} \mid \text{features}) = \frac{P(\text{features} \mid \text{toxic}) \times P(\text{toxic})}{P(\text{features})}$$

Bayes with more features (multiple pieces of evidence)

E = "dummy", "!"

H = toxic

P(H | E) = P(E | H) x P(H) / P(E)

$$P(\text{toxic} \mid \text{features}) = \frac{P(\text{features} \mid \text{toxic}) \times P(\text{toxic})}{P(\text{features})}$$

Just substitute!

P(toxic | "dummy", "!")

Bayes with more features (multiple pieces of evidence)

E = "dummy", "!"

H = toxic

P(H | E) = P(E | H) x P(H) / P(E)

$$P(\text{toxic} \mid \text{features}) = \frac{P(\text{features} \mid \text{toxic}) \times P(\text{toxic})}{P(\text{features})}$$

Just substitute!

$$P(\text{toxic} \mid \text{"dummy", "!"}) = \frac{P(\text{"dummy", "!"} \mid \text{toxic}) \times P(\text{toxic})}{P(\text{"dummy", "!"})}$$

Bayes with more features (multiple pieces of evidence)

E = "dummy", "!"

H = toxic

P(H | E) = P(E | H) x P(H) / P(E)

$$P(\text{toxic} \mid \text{features}) = \frac{P(\text{features} \mid \text{toxic}) \times P(\text{toxic})}{P(\text{features})}$$

Just substitute!

$$P(\text{toxic} \mid \text{"dummy", "!"}) = \frac{P(\text{"dummy", "!"} \mid \text{toxic}) \times P(\text{toxic})}{P(\text{"dummy", "!"})}$$

$$= \frac{P(\text{"dummy"} \mid \text{toxic}) \times P(\text{"!"} \mid \text{toxic}) \times P(\text{toxic})}{P(\text{"dummy"}) \times P(\text{"!"})}$$

Ignore the Denominator

We usually use Bayes Rule to **compare** probabilities (e.g.probability of toxic vs non-toxic, probability of one candidate correction vs another)

If we are comparing probabilities, we can just calculate the numerator and **ignore the denominator**

Why? Because the denominator is the same for everything we are comparing!

Let's compare **P(toxic | "dummy", "!")** to

P(nontoxic | "dummy", "!")

Ignore the Denominator

We usually use Bayes Rule to **compare** probabilities (e.g.probability of toxic vs non-toxic, probability of one candidate correction vs another)

When we do this, we can just calculate the numerator and ignore the denominator (because it's the same for everything we are comparing)

$$P(\text{toxic} \mid \text{"dummy", "!"}) = \frac{P(\text{"dummy", "!"} \mid \text{toxic}) \times P(\text{toxic})}{P(\text{"dummy", "!"})}$$

$$=$$

Step 1: Apply Bayes Rule

$$P(\text{non-toxic} \mid \text{"dummy", "!"}) = \frac{P(\text{"dummy", "!"} \mid \text{non-toxic}) \times P(\text{non-toxic})}{P(\text{"dummy", "!"})}$$

$$=$$

Ignore the Denominator

We usually use Bayes Rule to **compare** probabilities (e.g.probability of toxic vs non-toxic, probability of one candidate correction vs another)

When we do this, we can just calculate the numerator and ignore the denominator (because it's the same for everything we are comparing)

$$P(\text{toxic} \mid \text{"dummy", "!"}) = \frac{P(\text{"dummy", "!"} \mid \text{toxic}) \times P(\text{toxic})}{P(\text{"dummy", "!"})}$$

$$= \frac{P(\text{"dummy"} \mid \text{toxic}) \times P(\text{"!"} \mid \text{toxic}) \times P(\text{toxic})}{P(\text{"dummy"}) \times P(\text{"!"})}$$

$$P(\text{non-toxic} \mid \text{"dummy", "!"}) = \frac{P(\text{"dummy", "!"} \mid \text{non-toxic}) \times P(\text{non-toxic})}{P(\text{"dummy", "!"})}$$

$$= \frac{P(\text{"dummy"} \mid \text{non-toxic}) \times P(\text{"!"} \mid \text{non-toxic}) \times P(\text{non-toxic})}{P(\text{"dummy"}) \times P(\text{"!"})}$$

Step 2: Expand Joint probabilities

Ignore the Denominator

We usually use Bayes Rule to **compare** probabilities (e.g.probability of toxic vs non-toxic, probability of one candidate correction vs another)

When we do this, we can just calculate the numerator and ignore the denominator (because it's the same for everything we are comparing)

$$P(\text{toxic} \mid \text{"dummy", "!"}) = \frac{P(\text{"dummy", "!"} \mid \text{toxic}) \times P(\text{toxic})}{P(\text{"dummy", "!"})}$$

$$= \frac{P(\text{"dummy"} \mid \text{toxic}) \times P(\text{"!"} \mid \text{toxic}) \times P(\text{toxic})}{P(\text{"dummy"}) \times P(\text{"!"})}$$

Step 3: Simplify. Denominators are the same!

$$P(\text{non-toxic} \mid \text{"dummy", "!"}) = \frac{P(\text{"dummy", "!"} \mid \text{non-toxic}) \times P(\text{non-toxic})}{P(\text{"dummy", "!"})}$$

$$= \frac{P(\text{"dummy"} \mid \text{non-toxic}) \times P(\text{"!"} \mid \text{non-toxic}) \times P(\text{non-toxic})}{P(\text{"dummy"}) \times P(\text{"!"})}$$

Ignore the Denominator

We usually use Bayes Rule to **compare** probabilities (e.g.probability of toxic vs non-toxic, probability of one candidate correction vs another)

When we do this, we can just calculate the numerator and ignore the denominator (because it's the same for everything we are comparing)

$$P(\text{toxic} \mid \text{"dummy", "!"}) \propto P(\text{"dummy", "!"} \mid \text{toxic}) \times P(\text{toxic})$$

$$\propto P(\text{"dummy"} \mid \text{toxic}) \times P(\text{"!"} \mid \text{toxic}) \times P(\text{toxic})$$

Step 3: Simplify. Denominators are the same!

$$P(\text{non-toxic} \mid \text{"dummy", "!"}) \propto P(\text{"dummy", "!"} \mid \text{non-toxic}) \times P(\text{non-toxic})$$

$$\propto P(\text{"dummy"} \mid \text{non-toxic}) \times P(\text{"!"} \mid \text{non-toxic}) \times P(\text{non-toxic})$$

Ignore the Denominator

We usually use Bayes Rule to **compare** probabilities (e.g.probability of toxic vs non-toxic, probability of one candidate correction vs another)

When we do this, we can just calculate the numerator and ignore the denominator (because it's the same for everything we are comparing)

$$P(H \mid E) \propto P(E \mid H) \times P(H)$$

$\propto$ means "proportional to"

Ignore the Denominator

We usually use Bayes Rule to **compare** probabilities (e.g.probability of toxic vs non-toxic, probability of one candidate correction vs another)

When we do this, we can just calculate the numerator and ignore the denominator (because it's the same for everything we are comparing)

$$P(H \mid E) \propto P(E \mid H) \times P(H)$$

∝ means "proportional to"

$$P(\text{toxic} \mid \text{"dummy"}, \text{"!"}) \propto P(\text{"dummy"}, \text{"!"} \mid \text{toxic}) \times P(\text{toxic})$$

$$\propto P(\text{"dummy"} \mid \text{toxic}) \times P(\text{"!"} \mid \text{toxic}) \times P(\text{toxic})$$

$$P(\text{non-toxic} \mid \text{"dummy"}, \text{"!"}) \propto P(\text{"dummy"}, \text{"!"} \mid \text{non-toxic}) \times P(\text{non-toxic})$$

$$\propto P(\text{"dummy"} \mid \text{toxic}) \times P(\text{"!"} \mid \text{toxic}) \times P(\text{toxic})$$

Ignore the Denominator

We usually use Bayes Rule to **compare** probabilities (e.g.probability of toxic vs non-toxic, probability of one candidate correction vs another)

When we do this, we can just calculate the numerator and ignore the denominator (because it's the same for everything we are comparing)

$$\boxed{P(H \mid E) \propto P(E \mid H) \times P(H)}$$

P(toxic | "dummy", "!") ∝ P("dummy", "!" | toxic) × P(toxic)

mmy" | toxic) x P("!" | toxic) x P(toxic)

the formula you need for the homework!!!

P(non-toxic | "du mmy", "!" | non-toxic) × P(non-toxic)

dummy" | toxic) x P("!" | toxic) x P(toxic)

A visual Introduction to machine learning

part 1: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

part 2: http://www.r2d3.us/visual-intro-to-machine-learning-part-2/

visual aide for understanding

- features/dimensions
- true positives / false positives, true negatives, false negatives
- overfitting (part 2)

How do we evaluate performance on the test set?

Imagine we're building a hate speech classifier. The model should output 1

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

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

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

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

Our test set contains 100 examples.

The model was right on 91 of them.

Acc = 91/ 100
= 91%

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.*

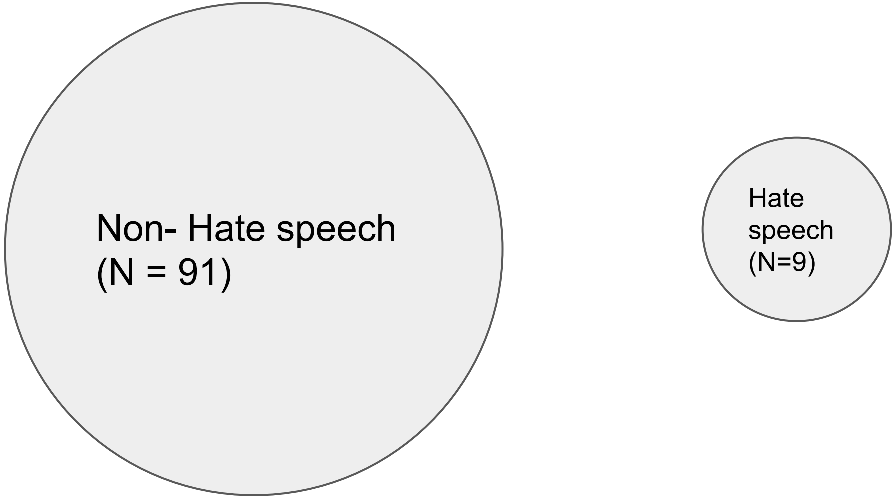
Maybe 0.05% of real life comments actually contain hate speech.

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

How accurate would the model be?

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?

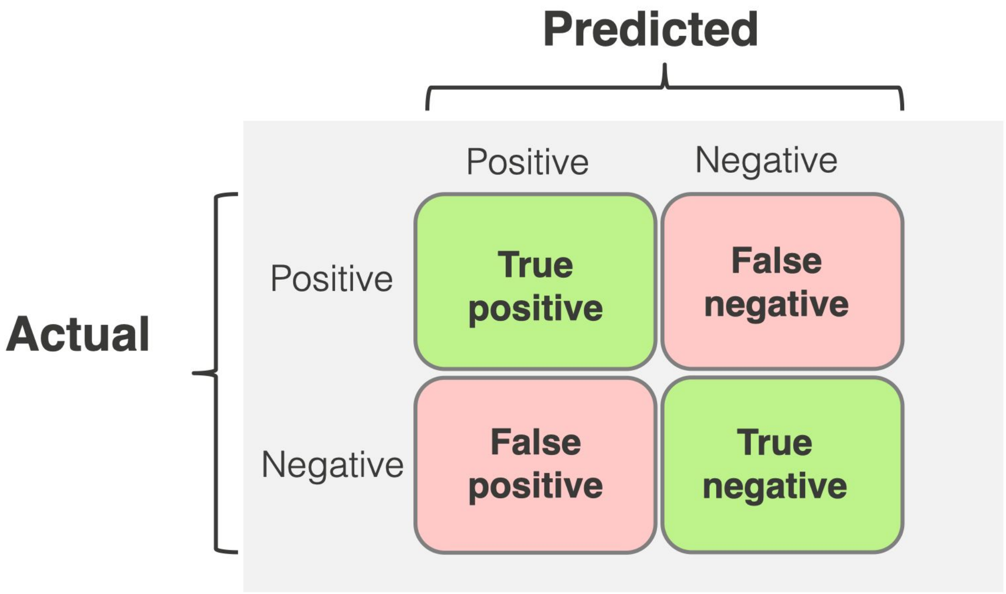
Types of Error

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

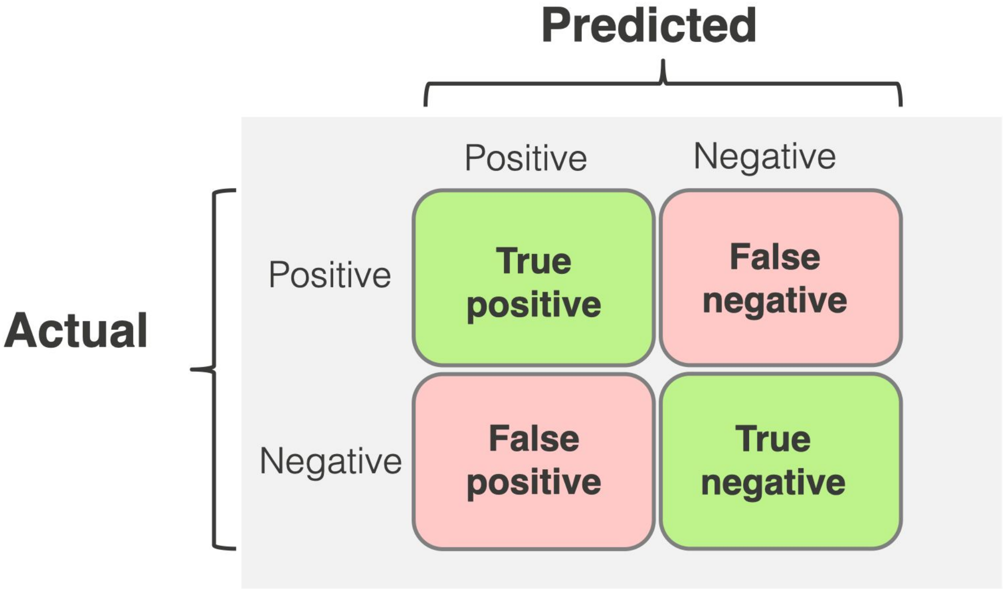
We can visualize types of error like this. (called a **confusion matrix)**

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

True Positives • Reality: Hate Speech • Classifier Prediction: Hate Speech • Number of TP results: **1**	**False Positives** • Reality: Normal • Classifier Prediction: Hate Speech • Number of FP results: **1**
False Negatives • Reality: Hate Speech • Classifier Prediction: Benign • Number of FN results: **8**	**True Negatives** • Reality: Normal • Classifier Prediction: Benign • Number of TN results: **90**

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

Accuracy Of a Hate Speech Detector

True Positives • Reality: Hate Speech • Classifier Prediction: Hate Speech • Number of TP results: **1**	**False Positives** • Reality: Normal • Classifier Prediction: Hate Speech • Number of FP results: **1**
False Negatives • Reality: Hate Speech • Classifier Prediction: Benign • Number of FN results: **8**	**True Negatives** • Reality: Normal • Classifier Prediction: Benign • Number of TN results: **90**

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

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

Accuracy Of a Hate Speech Detector

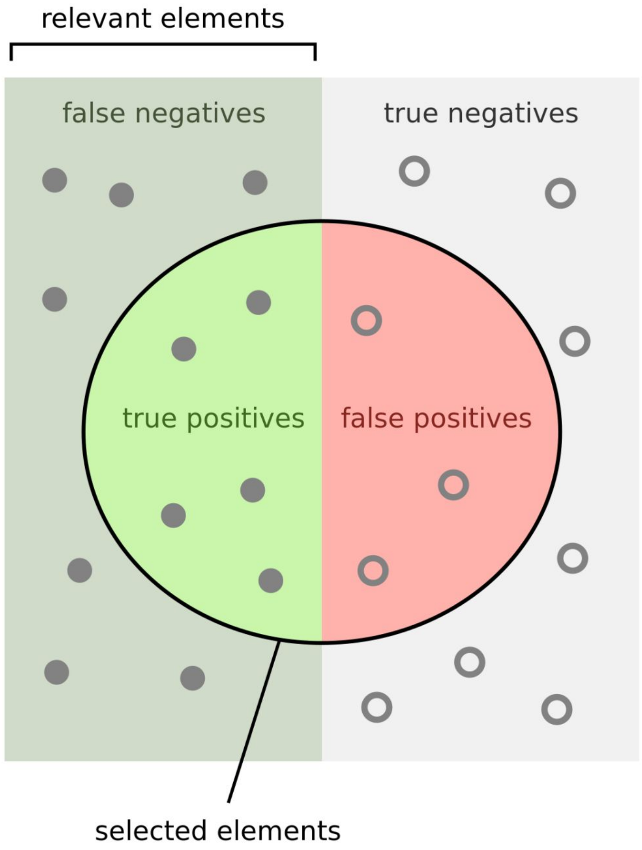
$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Number of total predictions}}$$

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

True Positives	False Positives
• Ground Truth: Hate Speech • Classifier Prediction: Hate Speech • Number of TP results: **1**	• Reality: Benign • Classifier Prediction: Hate Speech • Number of FP results: **1**
False Negatives	**True Negatives**
• Reality: Hate Speech • Classifier Prediction: Benign • Number of FN results: **8**	• Reality: Benign • ML model predicted: Benign • Number of TN results: **90**

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

Precision/Recall



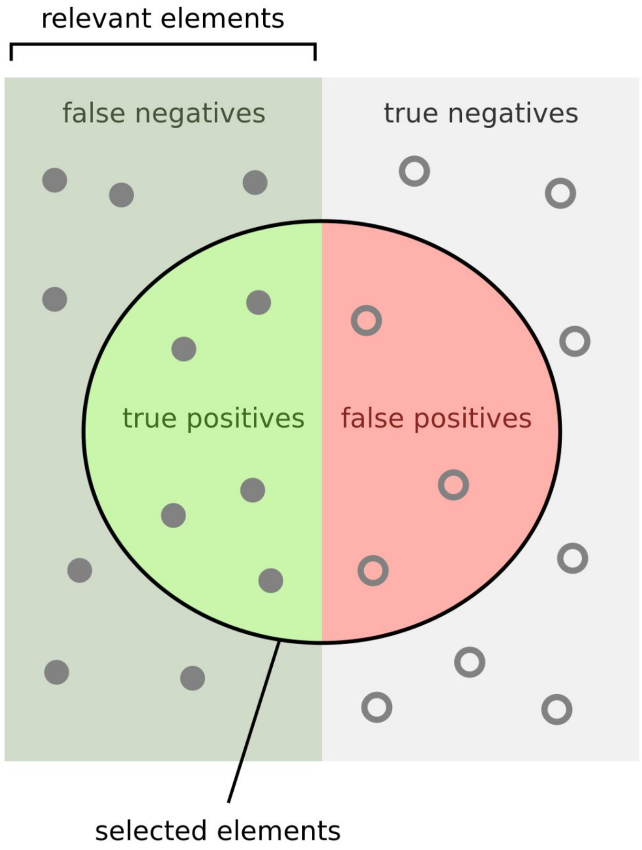
Precision:

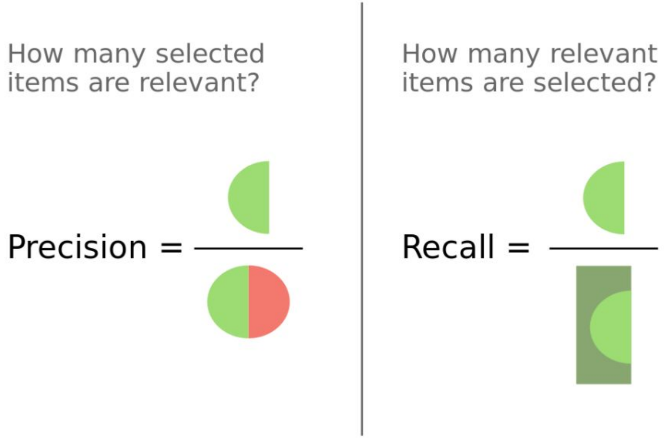
- Of all the tweets that you predicted to be POS, what proportion was correct?

Recall

- Of all the POS tweets in the test set, how many did you recall correctly?

Precision/Recall





Precision & Recall Of the Hate Speech Detector

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

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

True Positives • Ground Truth: Hate Speech • Classifier Prediction: Hate Speech • Number of TP results: **1**	**False Positives** • Reality: Benign • Classifier Prediction: Hate Speech • Number of FP results: **1**
False Negatives • Reality: Hate Speech • Classifier Prediction: Benign • Number of FN results: **8**	**True Negatives** • Reality: Benign • ML model predicted: Benign • Number of TN results: **90**

Precision & Recall Of the Hate Speech Detector

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

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

True Positives • Ground Truth: Hate Speech • Classifier Prediction: Hate Speech • Number of TP results: **1**	**False Positives** • Reality: Benign • Classifier Prediction: Hate Speech • Number of FP results: **1**
False Negatives • Reality: Hate Speech • Classifier Prediction: Benign • Number of FN results: **8**	**True Negatives** • Reality: Benign • ML model predicted: Benign • Number of TN results: **90**

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \frac{1}{1+8} = \frac{1}{9} = .11$$

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