Sentiment Analysis: Feature-based Methods

Katharina Kann — CSCI/LING5832

The Big Question

What kind of thing is the meaning of a sentence?

The Big Question

~~What kind of thing is the meaning of a sentence?~~

What can you do with a sentence if you know its meaning?

The Big Question

~~What kind of thing is the meaning of a sentence?~~

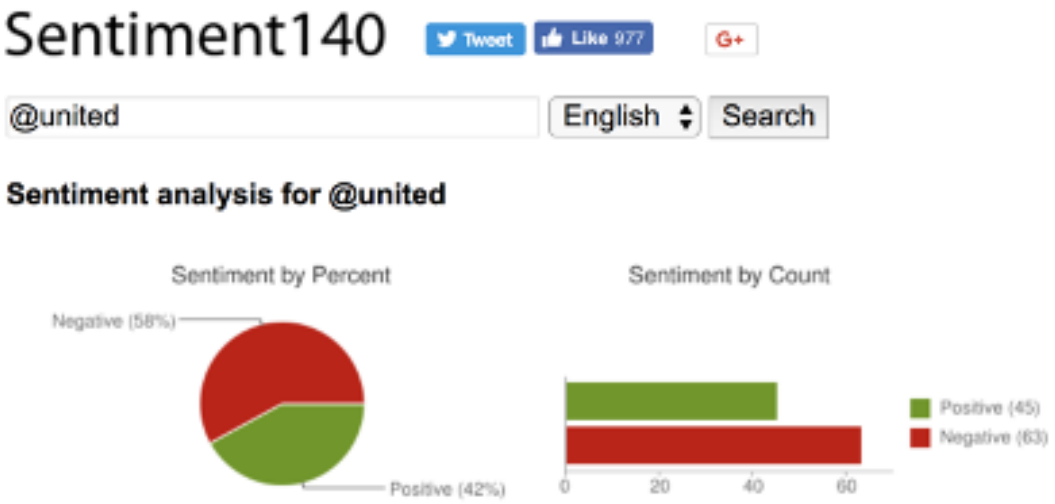
What can you do with a sentence if you know its meaning?

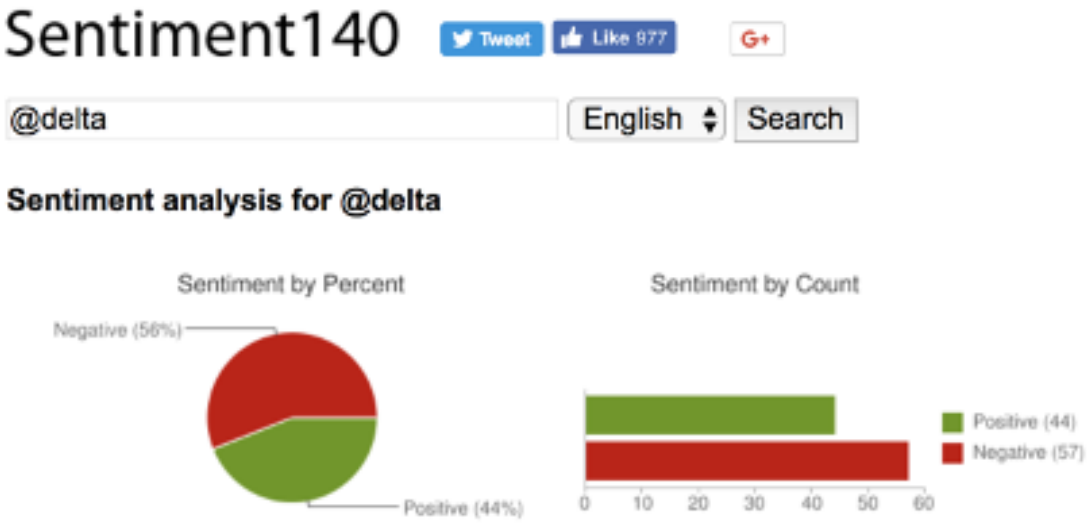
Judge its sentiment!

Positive or Negative?

- unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists
- this is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes.

An Example Application

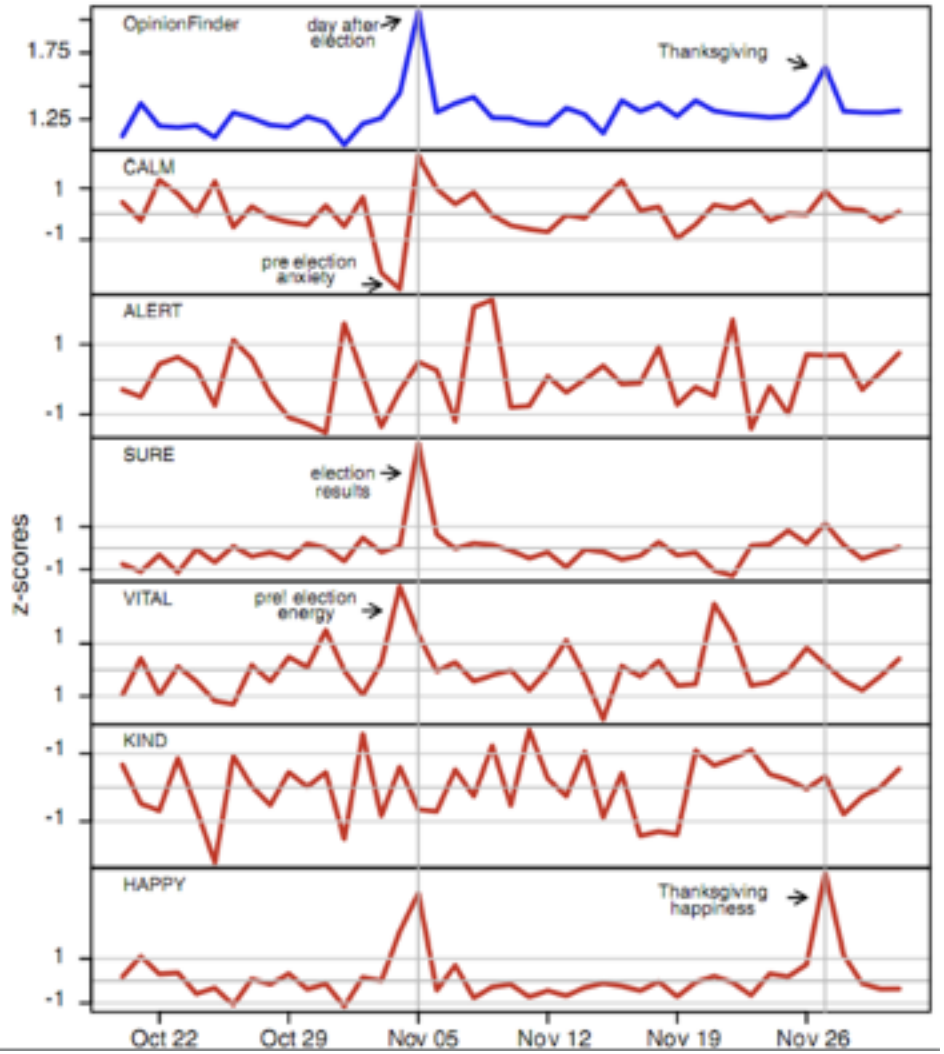




An Example Application

Twitter sentiment:

Johan Bollen, Huina Mao, Xiaojun Zeng. 2011.
Twitter mood predicts the stock market,
Journal of Computational Science 2:1, 1-8.
10.1016/j.jocs.2010.12.007.



An Example Application

Reviews

Write a review

4.3 ★★★★★ Very good

220 reviews on Google

Rooms	Location	Service
2.8	4.9	4.6

What guests are saying MORE

"The **breakfast** is included in the **price** and served downstairs."

"Good **location**, small size - avg **service**"

"Fabulous **place** for stay very comfy **bed room**"

Sentiment Analysis: Broadly Defined

Simplest task:

- Is the attitude of this text positive or negative?

More complex:

- Rank the attitude of this text from 1 to 5

Advanced:

- Detect the target, source, or complex attitude types

Sentiment Analysis: Broadly Defined

Simplest task:

- Is the attitude of this text positive or negative?

More complex:

- Rank the attitude of this text from 1 to 5

Advanced:

- Detect the target, source, or complex attitude types

IMDB Data

✓

when _star wars_ came out some twenty years ago , the image of traveling throughout the stars has become a commonplace image . [...]

when han solo goes light speed , the stars change to bright lines , going towards the viewer in lines that converge at an invisible point .

cool .

october sky offers a much simpler image–that of a single white dot , traveling horizontally across the night sky . [. . .]

✗

“ snake eyes ” is the most aggravating kind of movie : the kind that shows so much potential then becomes unbelievably disappointing .

it’s not just because this is a brian depalma film , and since he’s a great director and one who’s films are always greeted with at least some fanfare .

and it’s not even because this was a film starring nicolas cage and since he gives a brauvara performance , this film is hardly worth his talents .

From Sentences to Words

Tokenization

- For now, we assumed that words were the given units.
- Now, we are working with sentences
  - How do we define words?
  - (For the homework, we accept any tokenization.)

Tokenization

- Tokenization: Turning a sentence/ document into a sequence of symbols.
  - Simple strategy: Split on whitespace
    - doc.split()
  - When does this fail?

Tokenization

“I don’t like any of Ford’s trucks.”

“ I do n’t like any of Ford ’s trucks . ”

Tokenization

Or...

> Gesetze und Bestimmungen, die für Gastgeber relevant sind, fallen typischerweise in mehrere Kategorien, wie zum Beispiel Baugesetze, Wohnraumschutzgesetze, **Zweckentfremdungsverbotsgesetze**, Gewerbebestimmungen und Steuergesetze.
>
> **你能指引我去火车站吗?**

Tokenization

- For English, simple rule-based strategies work fine
  - Toolkits like spaCy, CoreNLP, NLTK, etc. contain tokenizers

Other Common Preprocessing Steps

- Case folding:
  - Does capitalization matter? If not, lowercase everything.

Other Common Preprocessing Steps

- Additional tricks for Twitter, web text, etc.
  - Hashtags:
    - #notesfromnationalemergency
      → # notes from national emergency
  - HTML/XML:
    - here
      → here

A Sentiment Classifier

Classification and Regression

- Classification:
  - Output: One choice from a fixed, finite set of labels
- Regression:
  - Output: One number, possibly from a fixed range

Classification and Regression

- Classification:
  - Output: One choice from a fixed, finite set of labels
- Regression:
  - Output: One number, possibly from a fixed range
- Sentiment can be either:
  - Positive/negative
  - Star ratings (★★★☆☆ vs. ★★★★☆)
  - Real-valued scores out of 100 (86.28 vs. 86.31)
  - But for now, we'll stick to positive/negative classification.

Building a Sentiment Classifier

- What we have:
  - About 10k examples of sentence–label pairs.
  - (SST/Rotten Tomatoes)
  - Some intuition for what the labels mean

Building a Sentiment Classifier

- What we have:
  - About 10k examples of sentence–label pairs.
  - (SST/Rotten Tomatoes)
  - Some intuition for what the labels mean
- What we want:
  - A function that can quickly and accurately identify the label (from the seen set) for a new sentence.

Rules?

```
if any of {‘bad’, ‘awful’, ‘terrible’, ‘boring’} in x:
    return NEGATIVE
elif any of {‘great’, ‘fun’, ‘enjoyed’} in x:
        return POSITIVE
elif len(x) > 20 and ‘but’ in x and ‘liked’ in x[0:12]:
        return NEGATIVE
…
…
```

Naïve Bayes?

- Training/fitting is just counting—fast and easy.
  - For most other models we're interested in, this is difficult and slow.
  - Using the model is fast, too!

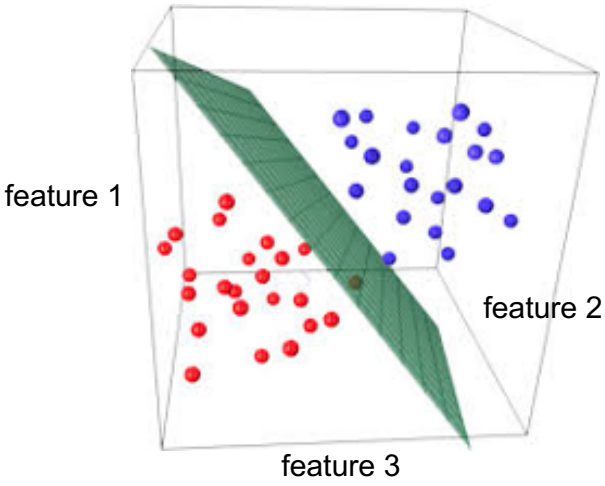
Feedforward Network?

- Usually good performance.
- But… slow to train.
- We will discuss this in the next class!

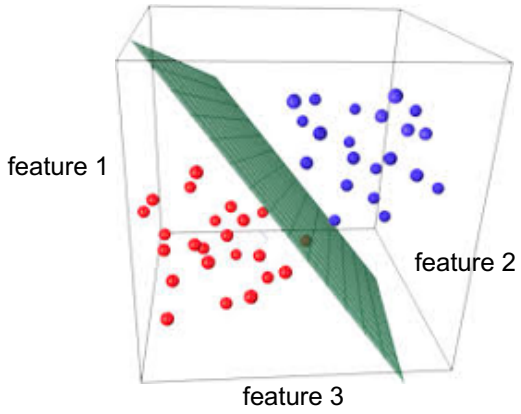
Logistic Regression

Logistic Regression

- Logistic regression models search for a point (1D feature vectors), line (2D), plane (3D), or hyperplane (4D+) that separates the examples of the different classes in feature space.

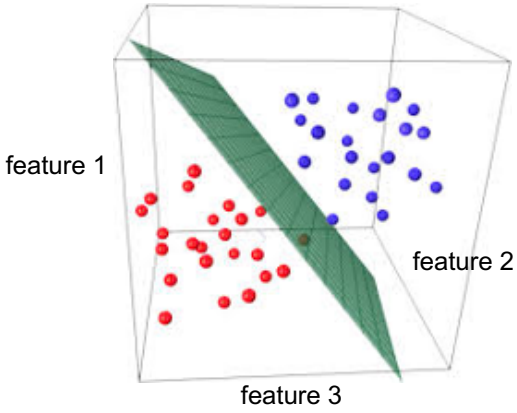


Logistic Regression



- Is basically a perceptron with a sigmoid function
- Is a discriminative classifier, while naive Bayes is a generative classifier
  - **Generative**: has the goal to understand how each class is created
  - **Discriminative**: just learns to distinguish between classes

Logistic Regression



- $y = \text{sigmoid}(b + \Sigma_0^i w_i x_i)$

- x is a feature vector
- w is a weight vector
- b is a bias vector
- y is the probability of class 1
  - (which class is that for binary sentiment classification?)

In-Class Exercise 1

- *the movie was entertaining and funny*
- *brad ’s new movie is super boring*

In-Class Exercise 1

- ***the movie was entertaining and funny***
- ***brad 's new movie is super boring***

- Our features are: 1) contains “funny”; 2) contains “boring”; 3) contains “good”
- Class 1 is “positive”, class 0 is “negative”
- Weights are [10, -10, 20], bias is 0
- Compute the most likely class for above sentences using our logistic regression model!

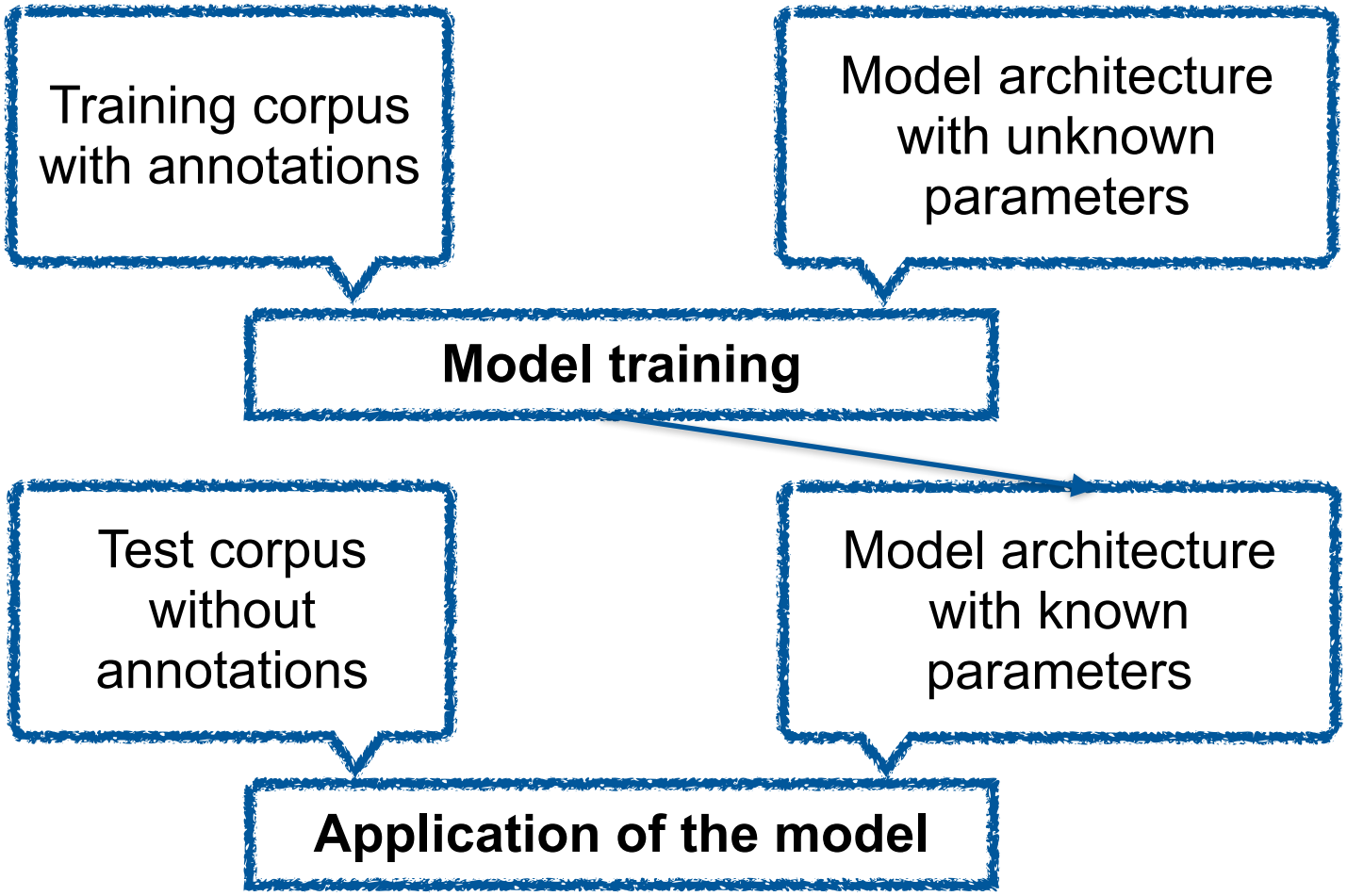
Optimization and Training

- How to obtain our parameters?
- Many models come with cost functions.
  - Specifies how badly the model did on a particular example.
  - The glass is half empty:
    - Cost is generally a positive number, even when we get the example right.
    - To train: Search for parameters that minimize the value cost function, summed over the training data.

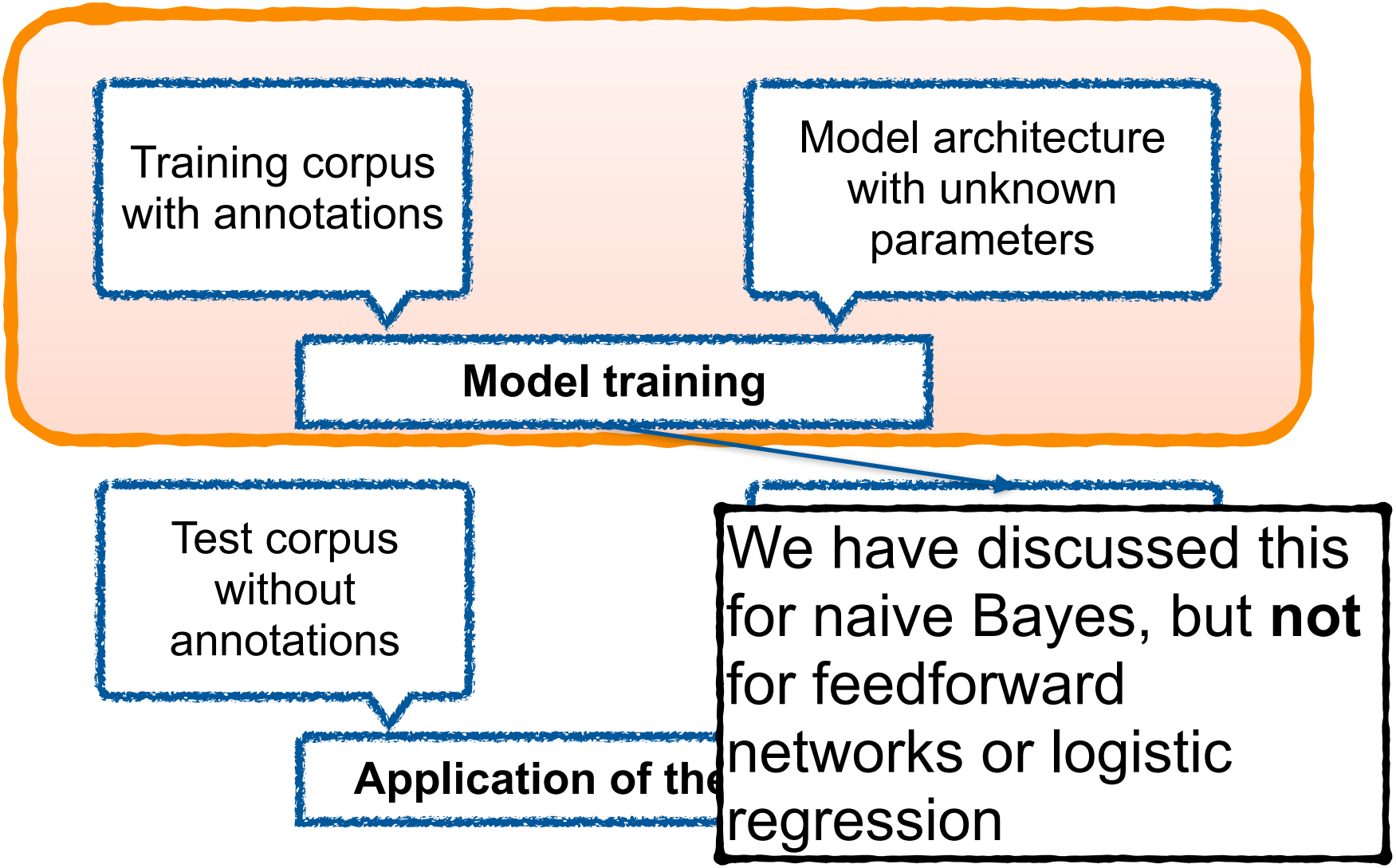
Optimization and Training

- How to obtain our parameters?
- Many models come with cost functions.
  - Specifies how badly the model did on a particular example.
  - The glass is half empty:
    - Cost is generally a positive number, even when we get the example right.
    - To train: Search for parameters that minimize the value cost function, summed over the training data.
- We will discuss this soon!

Optimization and Training

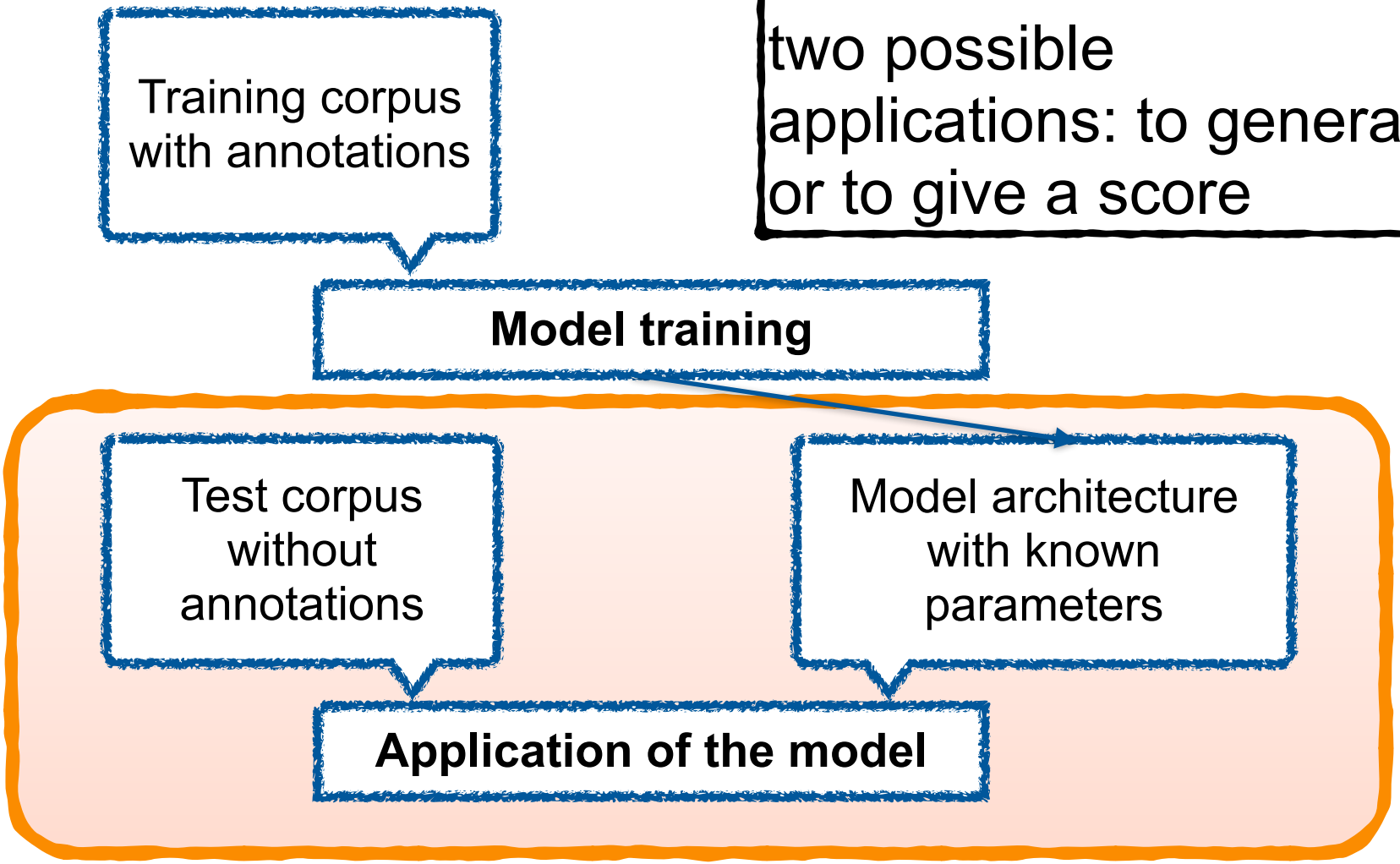


Optimization and Training



Optimization and Training

Generative models have two possible applications: to generate or to give a score



Feature Engineering

For Most Classifiers…

- Input sentences should be represented as vectors.
- These vectors should give you information that helps you separate examples belonging to the different classes.
- To build those vectors, there's room for some creativity…

Features: Beyond the Basics

- Features don’t have to be single words!

“ snake eyes ” is the most aggravating kind of movie : the kind that shows so much potential then becomes unbelievably disappointing .

a a	0
a Aaron	0
…	…
aggravating kind	1
…	…
apple butter	0
…	…
becomes unbelievably	1
…	…
cab driver	0
…	…
disappointing .	1
…	…

Features: Beyond the Basics

- Features don’t have to be 0 or 1!

“ snake eyes ” is the most aggravating kind of movie : the kind that shows so much potential then becomes unbelievably disappointing .

a	0
Aaron	0
…	…
aggravating	1
…	…
apple	0
…	…
unbelievably	1
…	…
cab	0
…	…
the	**2**
…	…

Features: Beyond the Basics

- Features can also use external tools, for instance a lemmatizer!

“ snake eyes ” is the most aggravating kind of movie : the kind that shows so much potential then becomes unbelievably disappointing .

a	0
Aaron	0
…	…
aggravate	1
…	…
apple	0
…	…
become	1
…	…
cab	0
…	…
disappoint	1
…	…

Features: Beyond the Basics

- Features don't have to be integers!
  - One common approach: Represent a sentence as the sum of its (distributional) word embeddings!

" snake eyes " is the most aggravating kind of movie : the kind that shows so much potential then becomes unbelievably disappointing .

d1	-0.16
d2	0.13
d3	0.42
d4	-0.01
…	…
d299	0.14
d300	0.21

Features for Spam Detection

SpamAssassin Features:

Features for Spam Detection

SpamAssassin Features:

- Mentions Generic Viagra

Features for Spam Detection

SpamAssassin Features:

- Mentions Generic Viagra
- Online Pharmacy

Features for Spam Detection

SpamAssassin Features:

- Mentions Generic Viagra
- Online Pharmacy
- Mentions millions of (dollar) ((dollar) NN,NNN,NNN.NN)

Features for Spam Detection

SpamAssassin Features:

- Mentions Generic Viagra
- Online Pharmacy
- Mentions millions of (dollar) ((dollar) NN,NNN,NNN.NN)
- Phrase: impress ... girl

Features for Spam Detection

SpamAssassin Features:

- Mentions Generic Viagra
- Online Pharmacy
- Mentions millions of (dollar) ((dollar) NN,NNN,NNN.NN)
- Phrase: impress ... girl
- From: starts with many numbers

Features for Spam Detection

SpamAssassin Features:

- Mentions Generic Viagra
- Online Pharmacy
- Mentions millions of (dollar) ((dollar) NN,NNN,NNN.NN)
- Phrase: impress ... girl
- From: starts with many numbers
- Subject is all capitals

Features for Spam Detection

SpamAssassin Features:

- Mentions Generic Viagra
- Online Pharmacy
- Mentions millions of (dollar) ((dollar) NN,NNN,NNN.NN)
- Phrase: impress ... girl
- From: starts with many numbers
- Subject is all capitals
- HTML has a low ratio of text to image area

Features for Spam Detection

SpamAssassin Features:

- Mentions Generic Viagra
- Online Pharmacy
- Mentions millions of (dollar) ((dollar) NN,NNN,NNN.NN)
- Phrase: impress ... girl
- From: starts with many numbers
- Subject is all capitals
- HTML has a low ratio of text to image area
- One hundred percent guaranteed

Features for Spam Detection

SpamAssassin Features:

- Mentions Generic Viagra
- Online Pharmacy
- Mentions millions of (dollar) ((dollar) NN,NNN,NNN.NN)
- Phrase: impress ... girl
- From: starts with many numbers
- Subject is all capitals
- HTML has a low ratio of text to image area
- One hundred percent guaranteed
- Claims you can be removed from the list

Features for Spam Detection

SpamAssassin Features:

- Mentions Generic Viagra
- Online Pharmacy
- Mentions millions of (dollar) ((dollar) NN,NNN,NNN.NN)
- Phrase: impress ... girl
- From: starts with many numbers
- Subject is all capitals
- HTML has a low ratio of text to image area
- One hundred percent guaranteed
- Claims you can be removed from the list
- 'Prestigious Non-Accredited Universities'

Features: Two Views

- The statistical NLP approach:
  - Features should be carefully chosen, and should capture the engineer's intuitions about the kinds of information the model will need. Feature functions can be complex, and can use a variety of external resources.
  - Models should be relatively simple, and should focus on efficiently mapping feature vectors to predictions.

Features: Two Views

- The statistical NLP approach:
  - Features should be carefully chosen, and should capture the engineer's intuitions about the kinds of information the model will need. Feature functions can be complex, and can use a variety of external resources.
  - Models should be relatively simple, and should focus on efficiently mapping feature vectors to predictions.
- The deep learning approach:
  - Features should be a direct representation of the input, with little or no problem-specific engineering.
  - Models should be complex enough for end-to-end learning: They should learn to build their own feature functions that transform the raw inputs into something more useful.
  - Easier, but requires lots of training data.

Features: Two Views

- The borders between these can be fuzzy:
  - GloVe vectors in logistic regression?
  - Neural networks over part-of-speech tags?
  - …
- If you want to build a system that works, you'll often want to use a bit of each.

Evaluating a Two-Way Classifier

Accuracy

- Count +1 for each correct answer and divide this by the number of examples
  - For instance: You are classifying 100 sentences into “positive” or “negative”. You get 40 of them right.

Accuracy

- Count +1 for each correct answer and divide this by the number of examples
  - For instance: You are classifying 100 sentences into “positive” or “negative”. You get 40 of them right.
  - Accuracy: 40/100 = 0.4

Why Not Accuracy?

- Accuracy doesn’t give us much information if the classes are unbalanced
  - For instance: You are classifying 100 patients into “ill” or “not ill”. 95 are healthy and you have a system that always predicts healthy.

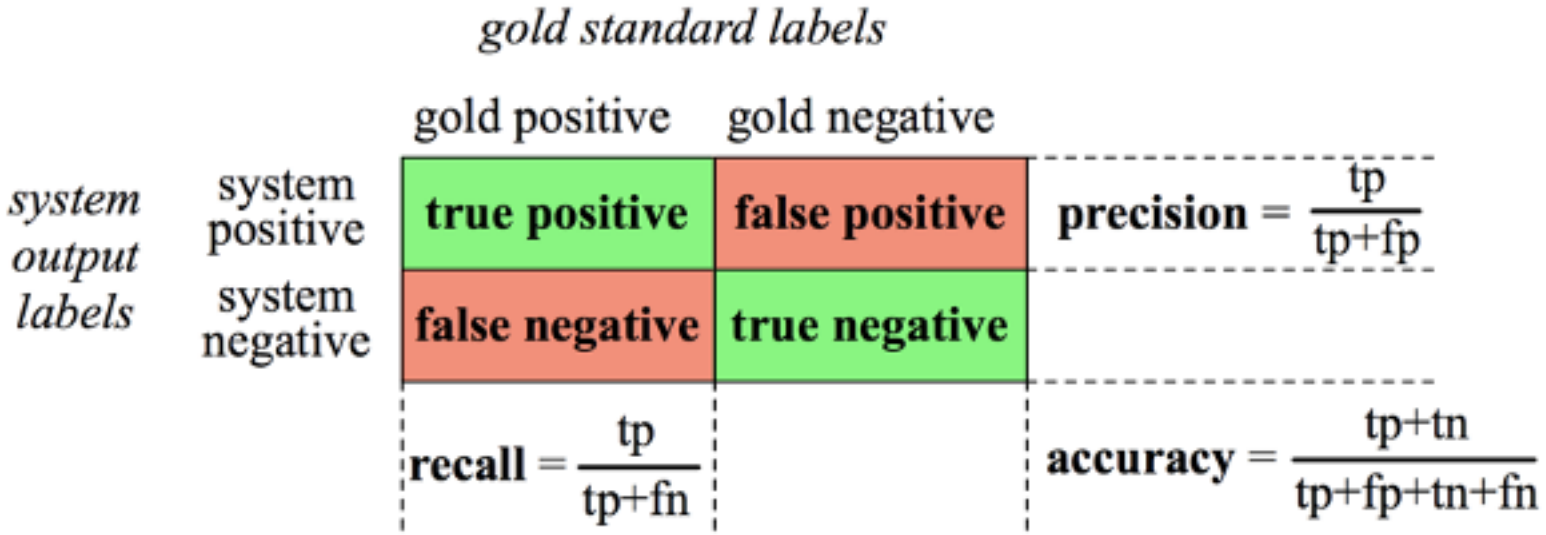
Why Not Accuracy?

- Accuracy doesn’t give us much information if the classes are unbalanced
  - For instance: You are classifying 100 patients into “ill” or “not ill”. 95 are healthy and you have a system that always predicts healthy.
  - Accuracy: 95/100 = 0.95

Why Not Accuracy?

- Accuracy doesn’t give us much information if the classes are unbalanced
  - For instance: You are classifying 100 patients into “ill” or “not ill”. 95 are healthy and you have a system that always predicts healthy.
  - Accuracy: 95/100 = 0.95
  - Is this a good system?

F1 Score



$$F_1 = \frac{2PR}{P+R}$$

In-Class Exercise 2

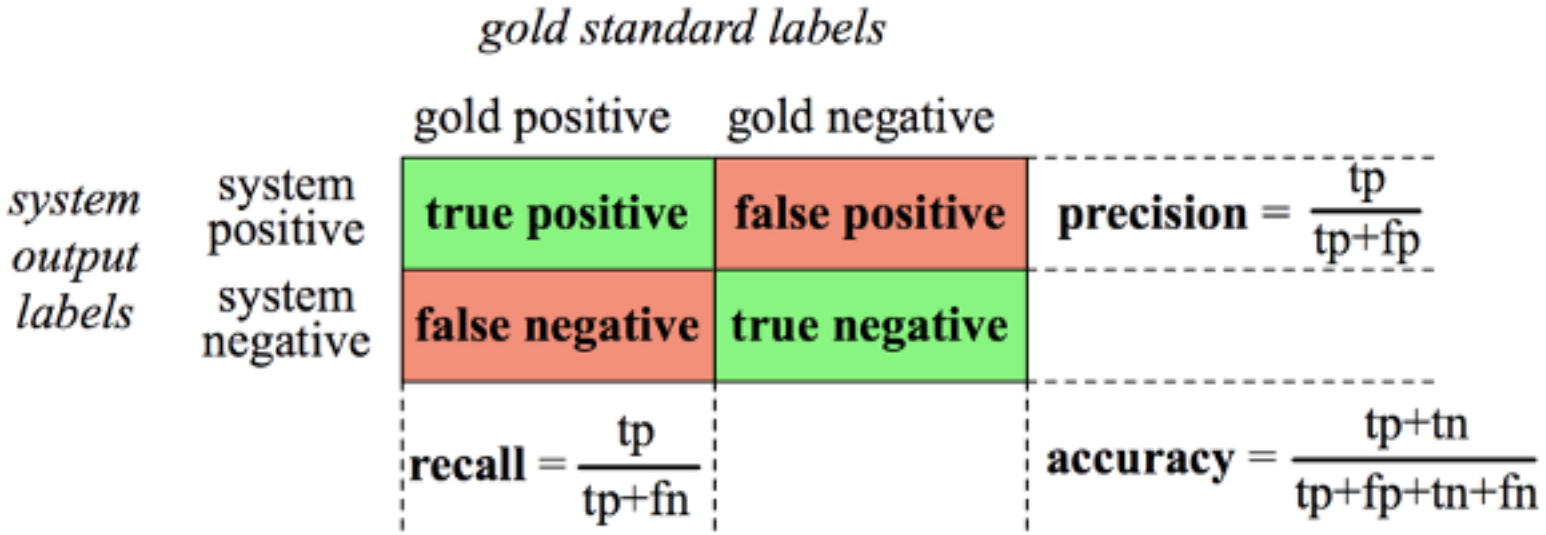
You have data from 100 patients and want to classify who has cancer. 53 of them in fact have cancer, but your system predicts positive for only 45 of them, and negative for 8. For the 47 healthy patients, your system predicts negative for 40, and positive for 7.

In-Class Exercise 2

You have data from 100 patients and want to classify who has cancer. 53 of them in fact have cancer, but your system predicts positive for only 45 of them, and negative for 8. For the 47 healthy patients, your system predicts negative for 40, and positive for 7.

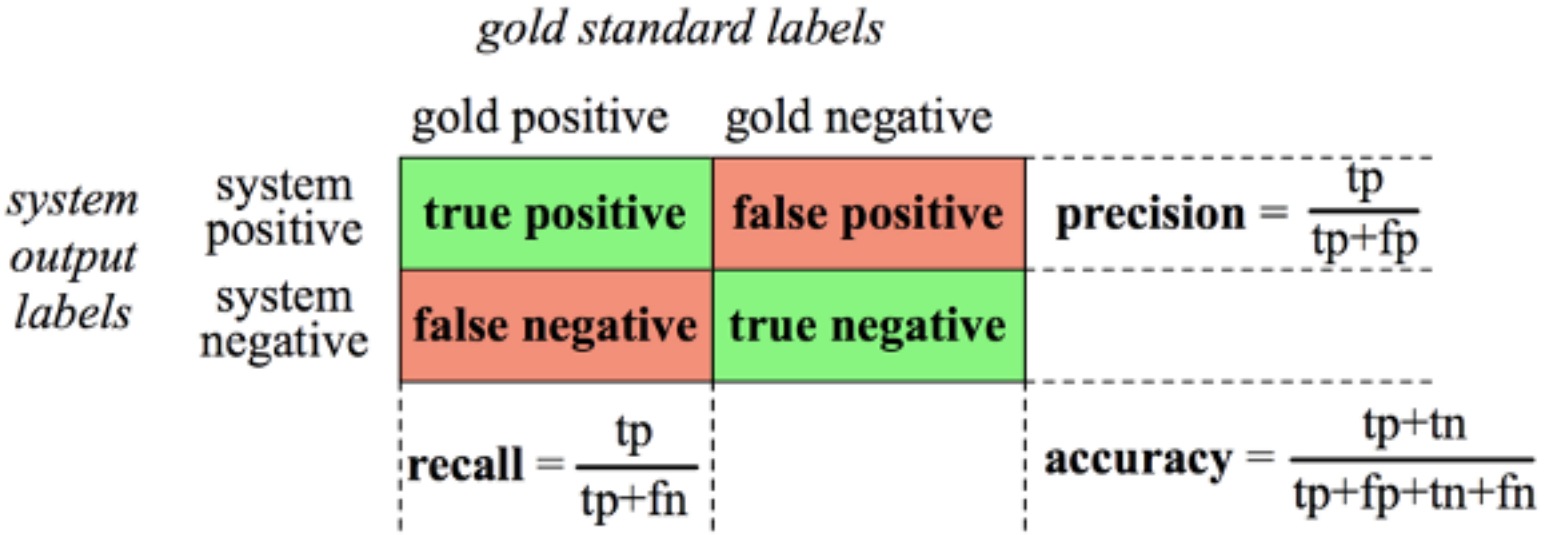
- Compute: precision, recall, and F1 score for your system!

F1 Score



$$F_1 = \frac{2PR}{P+R}$$

F1 Score



What's good for what?

$$F_1 = \frac{2PR}{P+R}$$

Multinomial Logistic Regression

Multi-Class Classification

- Needed for tasks that don't consist of a binary decision
- Many possible examples:
  - Sentiment: positive, negative, neutral
  - Natural language inference: entailment, contradiction, neutral
  - Predicting the next word
  - …

Multinomial Logistic Regression

- Also called **softmax regression** or **maxent classifier**.
- Uses a softmax function for the output!
- This requires weights for each class.

Recap: Softmax

- Use for one-of-many predictions
- Gives you a probability distribution
  - Values are all between 0 and 1
  - Values add up to 1

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \quad 1 \leq i \leq D$$

Wrapping up

- Discussed today:
  - Tokenization and preprocessing
  - Logistic regression
  - Feature engineering
  - Precision/recall/F1 score
  - Multinomial logistic regression

- On Wednesday: Sentiment Analysis with Neural Models