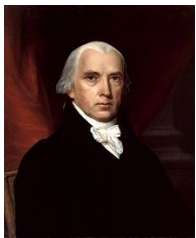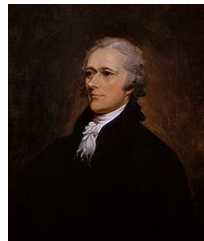# Deep Learning for
Natural Language Processing

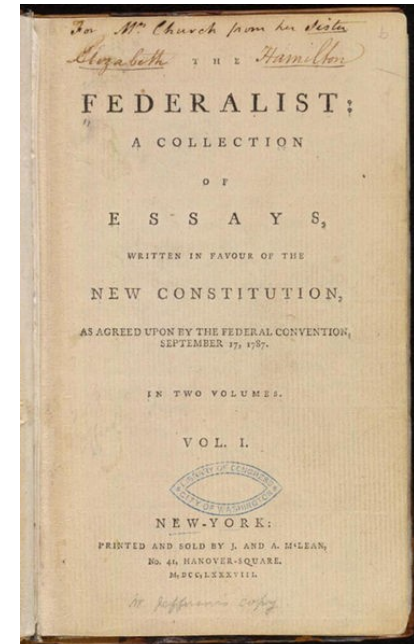## Text categorization

Who wrote which Federalist papers?

- 1787-8: anonymous essays try to convince New York to ratify U.S Constitution: Jay, Madison, Hamilton.

- Authorship of 12 of the letters in dispute

- 1963: solved by Mosteller and Wallace using Bayesian methods

James Madison

Alexander Hamilton

Positive or negative movie review?

- 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.

What is the subject of this article?

MEDLINE Article

MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

4

Text Classification

- Assigning subject categories, topics, or genres

- Spam detection

- Authorship identification

- Age/gender identification

- Language Identification

- Sentiment analysis

- …

Text Classification: definition

Input:
- a document d
- a fixed set of classes $C = \{c_1, c_2, ..., c_J\}$

Output: a predicted class $c \in C$

Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features
  - spam: black-list-address OR ("dollars" AND"have been selected")

- Accuracy can be high
  - If rules carefully refined by expert

- But building and maintaining these rules is expensive

Classification Methods: Supervised Machine Learning

- **Input***:*
  - a document d
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
  - A training set of m hand-labeled documents $(d_1, c_1), \ldots, (d_m, c_m)$

- **Output***:*
  - a learned classifier $\gamma: d \rightarrow c$

Classification Methods: Supervised Machine Learning

- Any kind of classifier
  - Naïve Bayes
  - Logistic regression
  - Support-vector machines
  - k-Nearest Neighbors
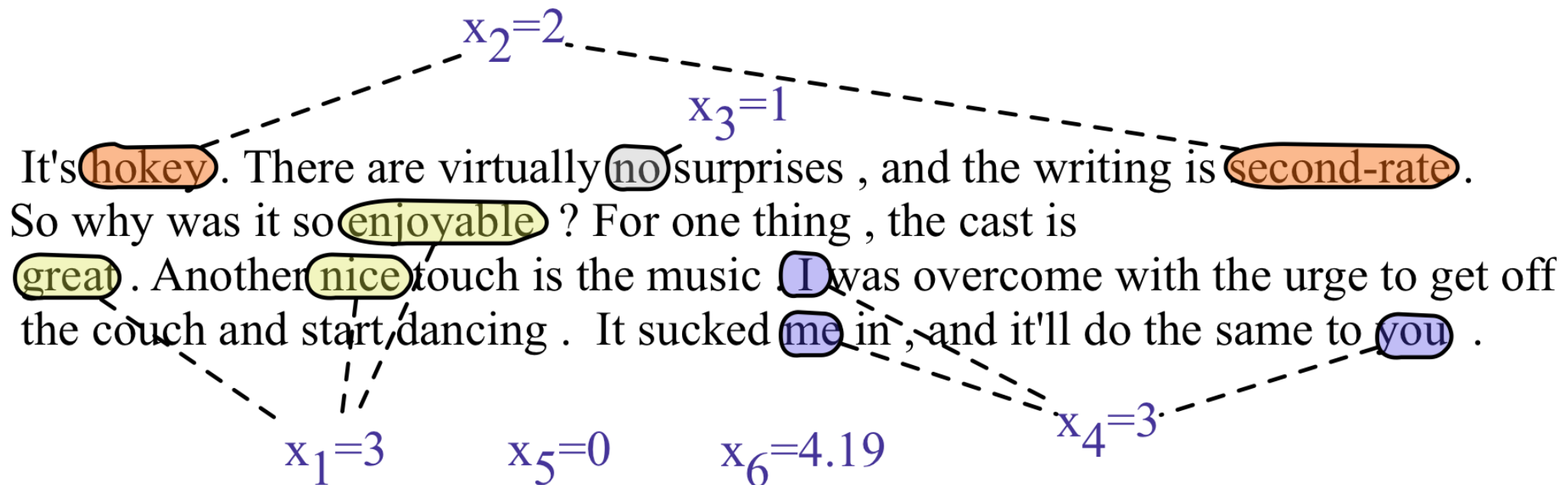
  ….

- Deep neural networks

Sentiment classification

- Binary sentiment classification on movie review.

- We would like to know whether to assign the sentiment class + or − to a review document.

Var	Definition	Value in Fig. 5.2
x_1	count(positive lexicon) $\in$ doc)	3
x_2	count(negative lexicon) $\in$ doc)	2
x_3	$\begin{cases} 1 & \text{if "no"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$	1
x_4	count(1st and 2nd pronouns $\in$ doc)	3
x_5	$\begin{cases} 1 & \text{if "!"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$	0
x_6	log(word count of doc)	$\ln(66) = 4.19$

23

Sentiment classification

- A sample mini test document showing the extracted features in the vector x:

$x_2=2$

$x_3=1$

It's hokey . There are virtually no surprises , and the writing is second-rate .
So why was it so enjoyable ? For one thing , the cast is
great . Another nice touch is the music . I was overcome with the urge to get off
the couch and start dancing . It sucked me in , and it'll do the same to you .

$x_1=3$ $x_5=0$ $x_6=4.19$ $x_4=3$

24

Sentiment classification

- Given 6 features and the input review x, P(+|x) and P(−|x) can be computed using:

$$
\begin{aligned}
p(+|x) = P(Y = 1|x) &= \sigma(w \cdot x + b) \\
&= \sigma([2.5, -5.0, -1.2, 0.5, 2.0, 0.7] \cdot [3, 2, 1, 3, 0, 4.19] + 0.1) \\
&= \sigma(.833) \\
&= 0.70 \\
p(-|x) = P(Y = 0|x) &= 1 - \sigma(w \cdot x + b) \\
&= 0.30
\end{aligned} \tag{5.6}
$$

- Feature interactions / templates should be designed by hand.

Word and document embeddings as features for LR, FFNN, RNN, ...

- Use the pre-trained word and document embeddings

- Instead of just bag of words

- Learn "a model" that associate these features with the target label in a generalizable way.

Argument mining as text categorization

- https://www.dialog-21.ru/evaluation/2022/ruarg/
- The annotation process included two stages: labelling by stance and labelling by premises. At both stages sentences were labelled in relation to the following claims:
  - "Vaccination is beneficial for society"
  - "The introduction and observance of quarantine is beneficial for society"
  - "Wearing masks is beneficial for society"

Stance	Premise	Numerical label
for	for	2
other (neutral/contradictory/unclear)	no argument	1
against	against	0
irrelevant	irrelevant	-1

Table 1: System of categories used to label the dataset.

Stance

- **for:** positive stance, which means that the speaker expresses his support for the topic;
- **against:** negative stance — the topic of discussion is not endorsed by the speaker;
- **other:** neutral stance (this label is used for factual sentences without any visible attitudes fi author); contradictory stance (for such a label, evident positive and negative attitudes should in a message); unclear stance (the presence of a stance is seen, but the context of sentence d give possibility to determine it);
- **irrelevant:** text does not contain stance on the topic.

Premise

- **for:** the stance is supported with argument in favor of the topic;
- **against:** the argument explains the author's negative outlook on the topic;
- **no argument:** no explanation is given for supporting/critisism of the topic;
- **irrelevant:** text does not contain stance and, consequently, premise on the topic.

A sentence was considered as a premise if the annotator could use it to convince an opponent about the given claim, such as "Masks help prevent the spread of disease."

Text	Masks		Quarantine		Vaccines	
	Stance	Premise	Stance	Premise	Stance	Premise
И какой смысл в вакцине если антитела только 3 месяца? *(And what's the point of a vaccine if the antibodies work only for 3 months?)*	—	—	—	—	against	against
Должна быть вакцина которую, будут прививать с детства!!! *(There must be a vaccine that will be vaccinated from childhood!!!)*	—	—	—	—	for	no argument
Вот только там на момент, когда была 1000 выявленных, уже неделю карантин действовал. *(At the time when there were 1000 identified, quarantine had been in effect for a week.)*	—	—	other	against	—	—
Развитие ситуации: если соблюдать карантин месяц, то вирус будет остановлен. *(The development of the situation: if the quarantine is observed for a month, the virus will be stopped.)*	—	—	for	for	—	—
Вопрос к властям :почему из гос резерва не получили люди масок когда их не хватало или и резерва уже нет *(Question to the authorities : why didn't people get masks from the state reserve when there were not enough of them or there is no reserve anymore)*	for	no argument	—	—	—	—
Любители масок не ужели вы думаете что эта косметическая тряпочка поможет от вируса?! *(Mask lovers don't you really think that this cosmetic rag will help against the virus?!)*	against	no argument	—	—	—	—

28

Evaluation: Precision and Recall

- The 2-by-2 contingency table:

	correct	not correct
selected	tp	fp
not selected	fn	tn

- **Precision**: % of selected items that are correct

- **Recall**: % of correct items that are selected

Evaluation: F1 score

- A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

$$F = \frac{1}{\alpha \dfrac{1}{P} + (1-\alpha)\dfrac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- The harmonic mean is a very conservative average;

- People usually use balanced F1 measure
  i.e., with $\beta = 1$ (that is, $\square = \frac{1}{2}$):   *F = 2PR/(P+R)*

Evaluation: Confusion matrix c

- For each pair of classes $<c_1,c_2>$ how many documents from c_1 were incorrectly assigned to c_2?
  - $c_{3,2}$: 90 wheat documents incorrectly assigned to poultry

Docs in test set	Assigned UK	Assigned poultry	Assigned wheat	Assigned coffee	Assigned interest	Assigned trade
True UK	95	1	13	0	1	0
True poultry	0	1	0	0	0	0
True wheat	10	90	0	1	0	0
True coffee	0	0	0	34	3	7
True interest	-	1	2	13	26	5
True trade	0	0	2	14	5	10

Evaluation: per class measures

Recall:
Fraction of docs in class *i* classified correctly:

$$\frac{c_{ii}}{\sum_j c_{ij}}$$

Precision:
Fraction of docs assigned class *i* that are actually about class *i*:

$$\frac{c_{ii}}{\sum_j c_{ji}}$$

Accuracy: (1 - error rate)
Fraction of docs classified correctly:

$$\frac{\sum_i c_{ii}}{\sum_j \sum_i c_{ij}}$$

Development Test Sets and Cross-validation

Training set	Development Test Set	Test Set

- Metric: P/R/F1 or Accuracy
- Unseen test set
  - avoid overfitting ('tuning to the test set')
  - more conservative estimate of performance
- Cross-validation over multiple splits
  - Handle sampling errors from different datasets
  - Pool results over each split
  - Compute pooled dev set performance

Training Set	Dev Test	

Training Set	Dev Test

Dev Test	Training Set

Test Set