Language Technology (2023)

Text classification

Marco Kuhlmann
Department of Computer and Information Science

LINKÖPING UNIVERSITY

This session

- Announcements
- Overview, questions & answers
- Introduction to lab 1

This session

- Announcements
- Overview, questions & answers
- Introduction to lab 1

Overview, questions and answers

Text classification

- **Text classification** is the task of categorising text documents into predefined classes.
- The term ‘document’ is applied to everything from tweets over press releases to complete books.

Sentiment analysis

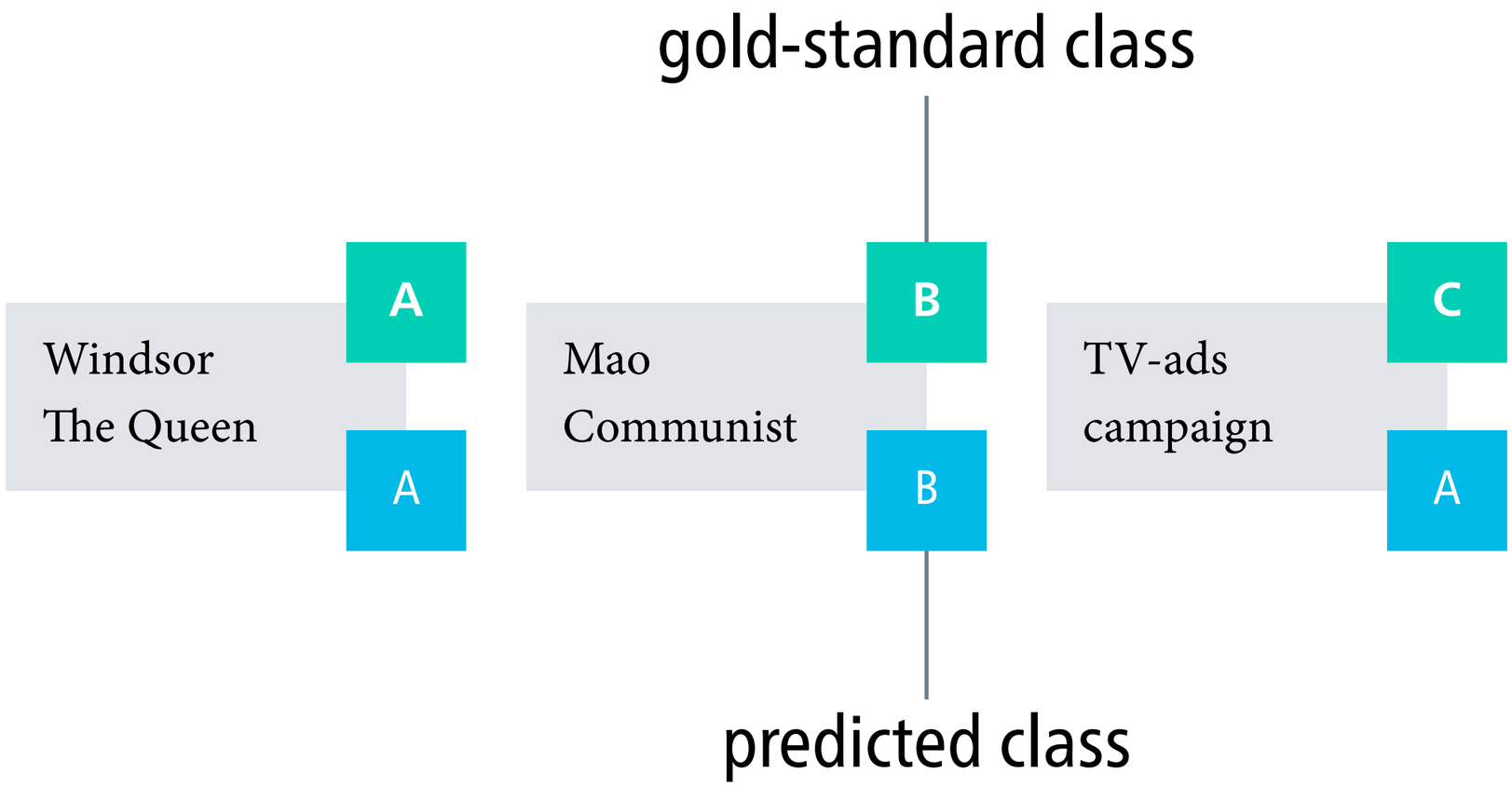
The gorgeously elaborate continuation of “The Lord of the Rings” trilogy is so huge that a column of words cannot adequately describe co-writer/director Peter Jackson’s expanded vision of J.R.R. Tolkien’s Middle-earth

positive

… is a sour little movie at its core; an exploration of the emptiness that underlay the relentless gaiety of the 1920’s, as if to stop would hasten the economic and global political turmoil that was to come.

negative

Evaluation of text classifiers



Evaluation of text classifiers | https://forms.office.com/e/JyLXvjFF6u

Precision and recall with respect to class A

	A	B	C
A	58	6	1
B	5	11	2
C	0	7	43

The role of baselines

- The evaluation measures as such do not really tell us much.

  Whether ‘80% accuracy’ is good or not depends on the task at hand.

- Instead, we should ask for a classifier’s performance relative to a reference result, a **baseline**.

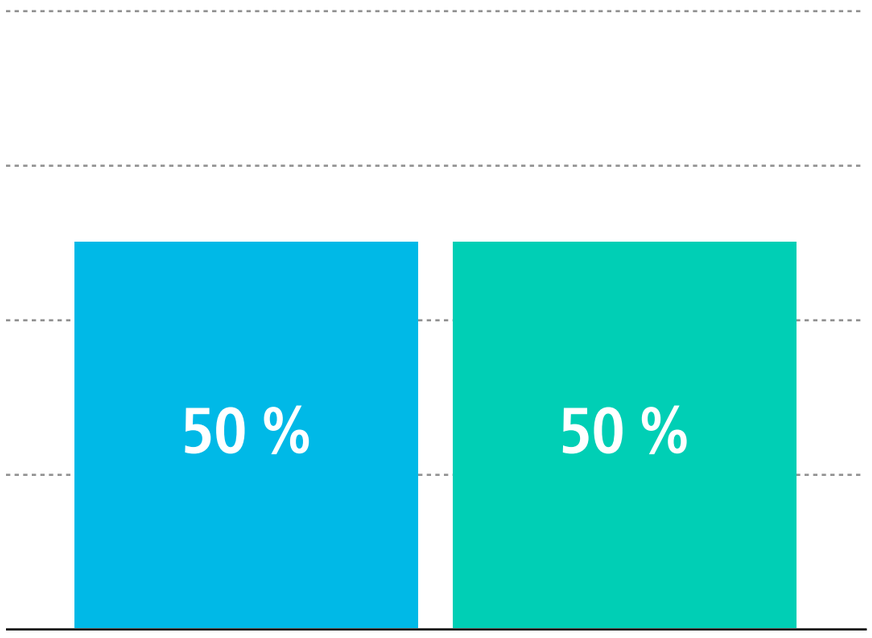
  ‘The accuracy of our system is 5 points higher than that of the baseline.’

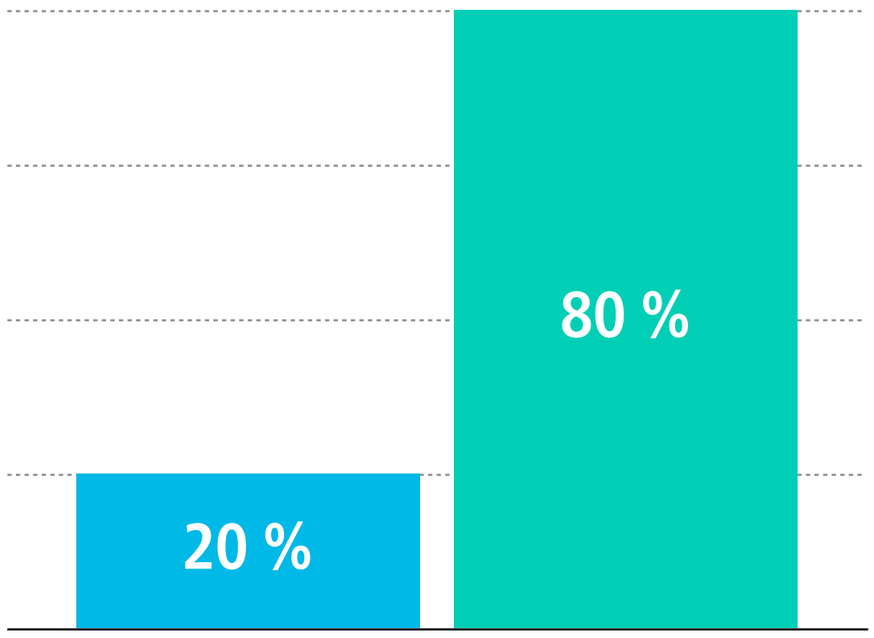
- A simple baseline for classification is to always predict the class which occurred most often in the training data.

  Most Frequent Class

Everything is relative, even accuracy

Is 80% accuracy good or bad?





Naive Bayes

- The **Naive Bayes classifier** is a simple but surprisingly effective probabilistic text classifier that builds on Bayes' rule.

- It is called 'naive' because it makes strong (unrealistic) independence assumptions about probabilities.

- It uses a representation of texts as **bags of words**, that is, it does not pay attention to word order.

Naive Bayes classification rule, informally

elaborate
gorgeously

positive

elaborate
gorgeously

score(**pos**) =

$P($**pos**$)\ P($elaborate$|$**pos**$)\ P($gorgeously$|$**pos**$)$

70%

negative

elaborate
gorgeously

score(**neg**) =

$P($**neg**$)\ P($elaborate$|$**neg**$)\ P($gorgeously$|$**neg**$)$

30%

Naive Bayes classification rule, formally

choose the class c that maximises
the term to the right of the ‘arg max’

$$\hat{c} = \arg\max_{c \in C} P(c) \cdot \prod_{w \in V} P(w \mid c)^{\#(w)}$$

predicted class

count of the word w

Implementing the Naive Bayes classification rule

- **Problem 1:** takes long time to loop over a large vocabulary

  Solution: loop over the words in the document instead

- **Problem 2:** words not in the vocabulary

  Solution: skip unknown words (this is what the model says!)

- **Problem 3:** underflow as one multiplies probabilities

  Solution: use log probabilities instead

Log probabilities | https://forms.office.com/e/QL3uZ6nuFt

MLE for the Naive Bayes classifier

$\#(c)$ number of documents with gold-standard class c

$\#(w, c)$ number of occurrences of w in documents with class c

$$P(c) = \frac{\#(c)}{\sum_{x \in C} \#(x)} \qquad P(w \mid c) = \frac{\#(w, c)}{\sum_{x \in V} \#(x, c)}$$

MLE with additive smoothing

$\#(c)$ number of documents with gold-standard class c

$\#(w, c)$ number of occurrences of w in documents with class c

$$P(c) = \frac{\#(c)}{\sum_{x \in C} \#(x)} \qquad P(w \mid c) = \frac{\#(w, c) + k}{\left[\sum_{x \in V} \#(x, c)\right] + k \cdot |V|}$$

Sample exam problem

This session

- Announcements
- Overview, questions & answers
- Introduction to lab 1

This session

- Announcements
- Overview, questions & answers
- Introduction to lab 1