

Introduction to text classification

Marco Kuhlmann

Department of Computer and Information Science

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.

Topic classification

UK	China	Elections	Sports
congestion London	Olympics Beijing	recount votes	diamond baseball
Parliament Big Ben	tourism Great Wall	seat run-off	forward soccer
Windsor The Queen	Mao Communist	TV-ads campaign	team captain

Adapted from Manning et al. (2008)

Forensic linguistics



‘I realized the faxed copy I just received was an outline of the manifesto, using much of the same wording, definitely the same topics and themes. ... I invented [the language analysis] for this case and really, forensic linguistics took off after that.’

James Fitzgerald, profiler

Sources: [Wikipedia](#), [Newsweek](#)

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

Use cases related to sentiment analysis

Eisenstein § 4.1.1

- **Subjectivity detection**

Identify those parts of a text that express subjective opinions, speculations, and hypotheticals.

[Riloff and Wiebe \(2003\)](#)

- **Stance classification**

Identify whether an author is in support or opposition of an argument or a proposal, such as a law.

[Anand et al. \(2011\)](#)

Text classification as supervised machine learning

congestion London	A	Olympics Beijing	B	recount votes	C
Parliament Big Ben	A	tourism Great Wall	B	seat run-off	C
Windsor The Queen	A	Mao Communist	B	TV-ads campaign	C

Text classification as supervised machine learning

training set
learning

congestion
London

A

Olympics
Beijing

B

recount
votes

C

Parliament
Big Ben

A

tourism
Great Wall

B

seat
run-off

C

test set
evaluation

Windsor
The Queen

A

Mao
Communist

B

TV-ads
campaign

C

Training and testing

- **Training**

When we train a classifier, we present it with a document x and its gold-standard class y and apply some learning algorithm.

- **Testing**

When we evaluate a classifier, we present it with x and compare the predicted class for this input with the gold-standard class y .

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

Bag-of-words features

Feature	Value
gorgeously	1
elaborate	1
huge	1
sour	0
little	0
...	

positive

Feature	Value
gorgeously	0
elaborate	0
huge	0
sour	1
little	1
...	

negative