# Classification of Documents

**Pavel Brazdil**
**LIAAD - INESC Porto LA**
**FEP, Univ. of Porto**

**Dec. 2010**

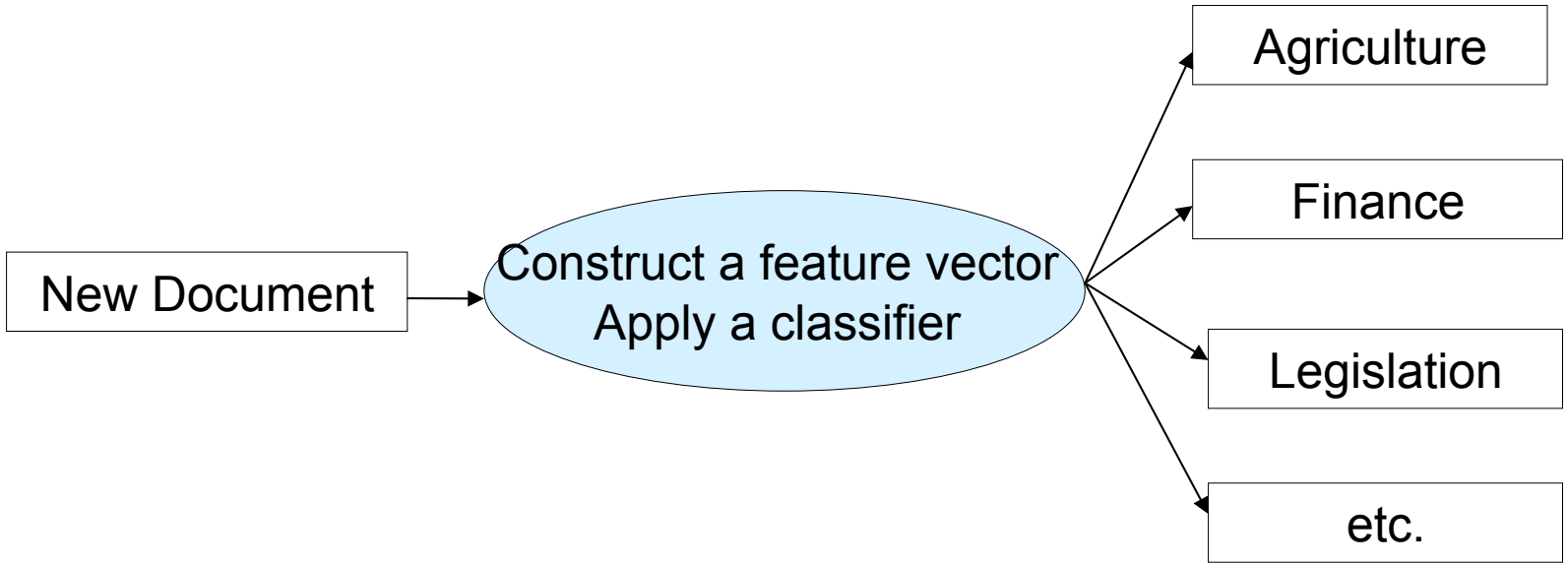
# Overview

Classification of Documents

1. What is document classification
2. Representation of documents
3. ML algorithms used in text classification
4. Some text classification datasets
5. Some comparative results
6. Sentiment analysis

# 1. What is Document Classification

Document classification / categorization enables
to classify a given document into one of pre-defined classes / categories.
First, the document is transformed into a feature vector.
Then classification is done with the help of a (trained) classifier.



# 2. Representation of Documents

As was shown earlier (Introduction to Text Mining, Representation of Docs), documents can be represented in many different ways.

Let us consider one simple possibility here,
where each document is represented by a feature vector,
capturing information about the occurrence of individual words.

The simplest possibility is to create a table, where
the columns identify different features,
each document is represented by a line (vector) of values and
the value in slot i,j captures information about feature j in document i.

These values can be either:

- Binary (e.g. 0/1), representing the presence/absence of a word,
- Word frequencies, indicating how many times a word occurred,
- TF-IDF coding,

# Constructing a Feature Vector: Basic Idea

Consider the following documents (each including just one sentence):

(from The harsh arithmetic behind the banking crisis, The Times, 2 Oct.'08)

D1: *British banks have operated with capital-to-asset ratios of about 5%.*

D2*: Banks have learned how to ensure that their borrowing customers pay back loans.*

D3*: The hash arithmetic behind the banking crisis.*

| Feature | British | banks | have | operated | ... | Banks | learned | how | ... | banking | crisis | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D1 | 1 | 1 | 1 | 1 | | 0 | 0 | 0 | | 0 | 0 | |
| D2 | 0 | 0 | 1 | 0 | | 1 | 1 | 1 | | 0 | 0 | |
| D3 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | | 1 | 1 | |

To construct such a table, we add a column for each new word encountered.
(we have re-used "*have*" of D1, after encountering "*have*" in D2)

The order of words is ignored. This representation is called bag-of-words.
The columns can be reordered, without loss of information.

# Improving the Features

**Separating the contiguous text into words (tokens)**

Spaces (“ “) and punctuation marks (“.”) are used to identify “words”.

**Convering the words to lower case**

In our previous example we had two distinct features “*banks*” and “*Banks*”.

This is undesirable, as both words represent the same concept.

If this were not corrected,

the process of classification would deliver worse results.

This can be corrected by converting all words to lower case.

# Improving the Features

**Stemming / Use of representative words**

We note that *operated* / *operate* / *operates* and banks / banking would also be considered distinct.

This is again undesirable, for similar reasons as explained earlier.

This can be corrected by stripping the suffixes “ed”, “ing” etc., called *stemming* or *lemmatization,* or by selecting a representative form for each word (e.g. the infinitive).

# Problem of High Dimensionality: Feature Reduction Techniques

Structured representations of natural languange documents leads usually to very large number of features.

For instance, one small collection of Reuters of 15,000 documents contains 25,000 non-trivial features (word stems).

Some algorithms do not deal very well with large numbers of features and hence it is necessary to employ feature reduction techniques, such as:

- Elimination of common words (stop words) (words that appear in many documents, e.g., *have*, *the* etc.),
- Selection of the most informative features (using e.g. information gain)

# Problem of Feature Sparcity

Another problem is feature sparcity:

Each document conctains only a small number of all potential features.

This requires that special representations be used.

- avoid using sparsely populated matrices,
- replace each word by a number (index in a unique word list)
- store information about where each word is used.

# Classification Strategies

If we need to discriminate N classes, we can define:
- a single multiclass problem,
- N problems 1-vs-All
- N*(N-1) / 2 problems 1-vs-1
  Ex. 20*(20-1) / 2 = 190

If we opt for binary problems, it is advisable to balance classes:
Use all examples of the minority class and
  about the same number of examples from the other class.

# 3. ML Algorithms used Text Classification

The ML algorithms that have proved to be useful in document classification:

- Naive Bayes
  The first ML algorithm used successfully for this task
  Mentioned already in the book of T.Mitchell: Machine Learning, 1997

- k-NN
  Particularly successful if used with *tf-idf* coding (and similar variants)

- SVM
  used typically with linear kernel
  uses 1-vs-All strategy

# 4. Some Text Classification Datasets

Some datasets used in previous comparative studies:

- 20Newsgroups
- Ohsumed
- Reuters

# 20Newsgroups

Short emails with context / class specific words

Available from

http://people.csail.mit.edu/jrennie/20Newsgroups/

Unpacked data 91MB

20.000 documents in 20 classes

(some overlapping classes)

Removed email headers, lower case, no stemming

Around 70MB for the document-term matrix

# 20Newsgroups

Subgroup “comp”
- comp.graphics
- comp.os.ms-windows.misc
- comp.sys.ibm.pc.hardware
- comp.sys.mac.hardware
- comp.windows.x

Subgroup “misc”
- misc.forsale

Subgroup “rec”
- rec.autos
- rec.motorcycles
- rec.sport.baseball
- rec.sport.hockey

Subgroup “sci”
- sci.crypt
- sci.electronics
- sci.med
- sci.space

Subgroup “talk.politics”
- talk.politics.guns
- talk.politics.mideast
- talk.politics.misc

Subgroup “religion”
- talk.religion.misc
- alt.atheism
- soc.religion.christian

# 20Newsgroups - Example document

…
**Newsgroups: comp.os.ms-windows.misc**
Subject: WIN STORM PC
Message-ID: <1993Apr27.153409.49548@kuhub.cc.ukans.edu>
From: srini@shannon.tisl.ukans.edu (Srini Seetharam)
Date: 27 Apr 93 15:33:57 CST
Reply-To: srini@shannon.tisl.ukans.edu (Srini Seetharam)
...
Anyone have any info. on the video/sound card from SIGMA designs.
It is called WIN STORM PC.
They also have another card called the legend 24lx
any info would be appreciated, incuding performance, pricing and availability.
thanks
srini

# Ohsumed

Medical abstracts

50.000 documents in 23 classes

Lower case, no stemming

Categories (classes):

| | |
|---|---|
| Bacterial Infections and Mycoses | C01 |
| Virus Diseases | C02 |
| Parasitic Diseases | C03 |
| Neoplasms | C04 |
| Musculoskeletal Diseases | C05 |
| Digestive System Diseases | C06 |
| Stomatognathic Diseases | C07 |
| Respiratory Tract Diseases | C08 |
| Otorhinolaryngologic Diseases | C09 |
| Nervous System Diseases | C10 |

(continued)

# Ohsumed

| | |
|---|---|
| Eye Diseases | C11 |
| Urologic and Male Genital Diseases | C12 |
| Female Genital Diseases and Pregnancy Complications | C13 |
| Cardiovascular Diseases | C14 |
| Cardiovascular Diseases | C14 |
| Hemic and Lymphatic Diseases | C15 |
| Neonatal Diseases and Abnormalities | C16 |
| Skin and Connective Tissue Diseases | C17 |
| Nutritional and Metabolic Diseases | C18 |
| Endocrine Diseases | C19 |
| Immunologic Diseases | C20 |
| Disorders of Environmental Origin | C21 |
| Animal Diseases | C22 |

# Ohsumed - example document

Catabolic illness. Strategies for enhancing recovery.

After injury, infection, extensive chemotherapy, and other critical illnesses, both protein and fat are lost from the body.

Although minor alterations in body composition are probably of little clinical importance, losses of body protein of 10 percent or more contribute to morbidity and debility.

This catabolic response can be modified and recovery can be accelerated by a variety of approaches.

First, the inflammatory response can be reduced; second, specific nutrients can be provided to support the patient's tissue requirements during catabolic illness; and third, growth factors can be used to enhance protein synthesis and tissue repair.

These approaches, whether used alone or in combination, will reduce the loss of body protein, which should accelerate recovery, shorten the length of hospitalization, and reduce convalescence.

# 5. Some comparative results

Work of Fabrice Colas (M.Sc student at LIAAD in 2006) and P. Brazdil:

SVM

- outperforms k-NN and NBayes for small feature set sizes,
- is slow in comparison

k-NN and NBayes

- are simple, well understood,
- have quite good performance,
- are fast

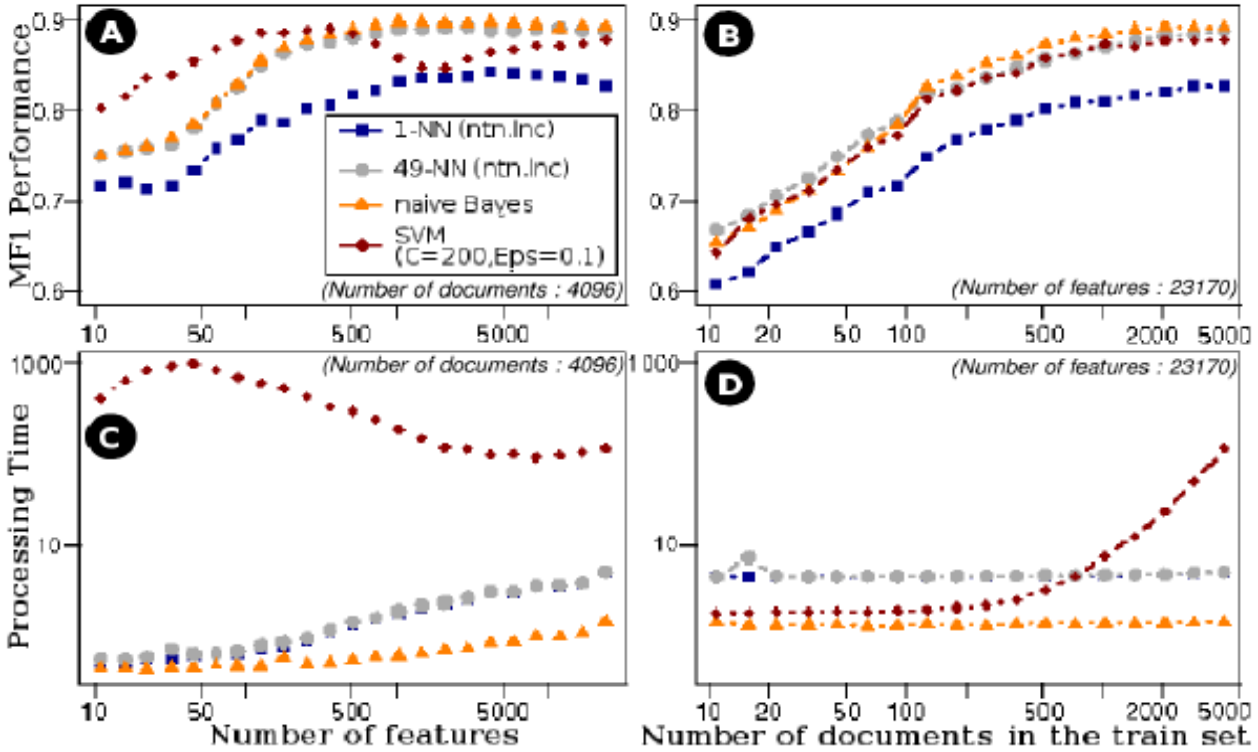
See the graph on next page

# Some comparative results

## Comparisons for increasing document and feature sizes

### Results of one typical classification task

Bacterial infections and mycoses against Disorders of environmental origin



# 6. Sentiment Analysis

Motivation

Decision makers require user's feedback
regards certain products or services
to determine whether to increase decrease production / correct something

Previous approaches relied on questionnaires, which are costly.

Sentiment analysis permits to obtain a similar information in a cheaper way
by analysing forums, discussion groups, blogs etc.

We distinguish between :

- positive or negative opinion orientation to the whole document,
  (can be seen as a classification task; involves positive / negative class)
- positive or negative opinion orientation with respect to some item
  (e.g. camera or some of its aspects, e.g. its size)
  The item needs to be identified beforehand (as in extraction)

# Features in Sentiment Analysis

Features:

- sentiment or opinion words or phrases
  - adjectives: great, excellent, amazing, bad, horrible etc.
    verbs: like, hate etc. ;
    phrases (camera is too heavy etc.)

Sentinent Lexicon