TEXT CLASSIFICATION

PASQUALE RULLO

TEXT CLASSIFICATION

- Text Classificationis the task of assigning documents to predefined thematic classes (topics) on the basis of the words occurring in them
- Text Classification is also called
  - Text Categorization
  - Document Classification
  - Document Categorization

TEXT CATEGORIZATION

- You want to classify documents into 4 classes: economics, sport, science, life
- There are two approaches that you can take:
  - Manually write a set of rules that classify documents
    - ball $\in$ d and goal $\in$ d $\rightarrow$ sport
  - Automatically create a classifier (machine learning-based approach) using a set of sample documents that are pre-classified into the classes (training data)

TEXT CATEGORIZATION IS A DIFFICULT TASK

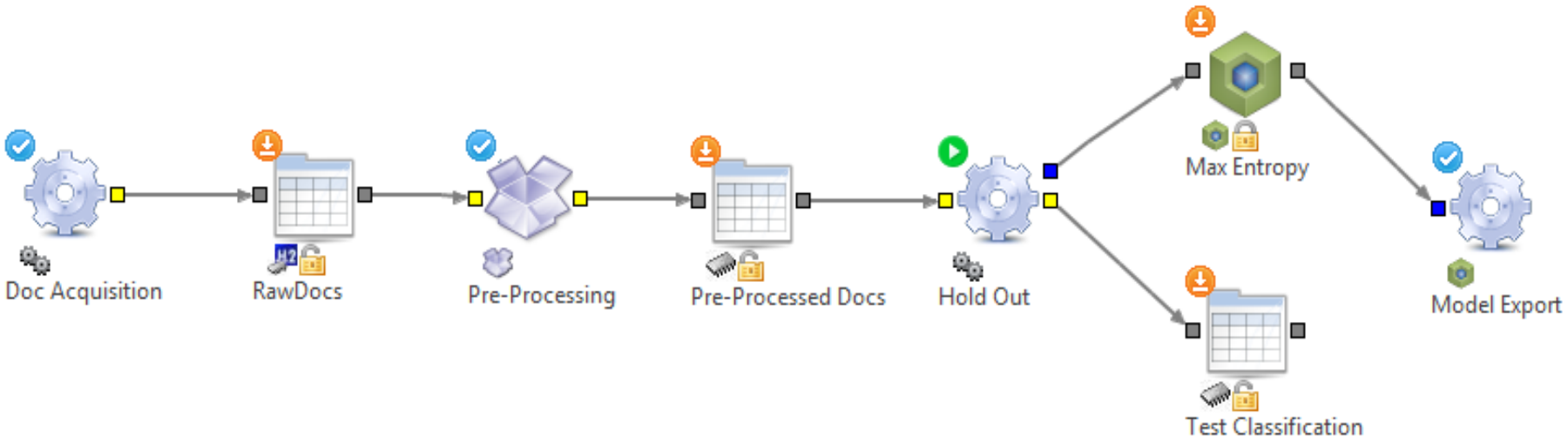
- TC is a difficult task essentially because it has to do with the complexity and richness of the natural language, which allows a concept to be expressed by a variety of constructs and words
- Natural languages are ambiguous
  - Synonymy: two phonemes, the same meaning, e.g., ball and dance (if the context is that of dance)
  - Polisemy: one phoneme, more meanings (ball as dance and ball as the round object used to play soccer)
  - “I have seen a man with the the binoculars” What is the meaning?
- Even manual classification is difficult – high degree of subjectivity

MACHINE LEARNING APPROACH PROBLEM DEFINITION

- Given a training set S={<d, {c_1, .., c_n})>}, where
  - d is a document
  - c_1, .., c_n are the topics of d – also called classes or categories (e.g., sport, gossip, politics, etc.)

  induce from S a model whereby the topics of a new document are determined
- Multi-label classification

TEXT CLASSIFICATION PROCESS

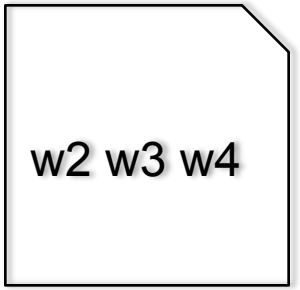


DOCUMENT REPRESENTATION

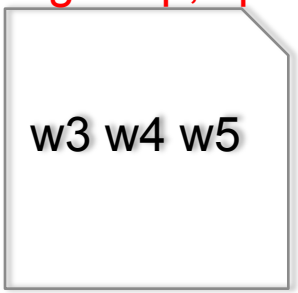
- Bag-of-word representation: a document is regarded as a bag of words regardless of the word order and grammar
- Binary representation: 0/1 as absence/presence
- Frequency: number of times a word/n-gram appears in a document

DOCUMENT REPRESENTATION

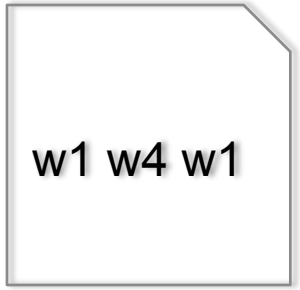
d1 - sport



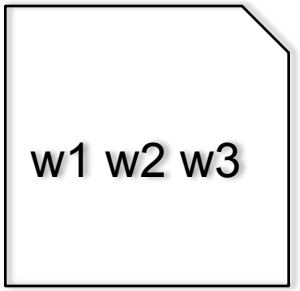
d2 – gossip, sport



d3 - sport



d4 - politics



- A document is a bag of words
- Assume that
  - d1 and d3 are about sport
  - d2 is about both gossip and sport
  - d4 is about politics

DOCUMENT REPRESENTATION

- Documents are the examples
- Words are the features (attributes)

	w1	w2	w3	w4	w5	Class
d1	0	1	1	1	0	sport
d2	0	0	1	1	1	gossip, sport
d3	1	0	0	1	0	sport
d4	1	1	1	0	0	politics

- Each attribute values represents presence/absence of a word

DOCUMENT REPRESENTATION

- Documents are the examples
- Words are the features (attributes)

	w1	**w2**	**w3**	**w4**	**w5**	**Class**
d1	0	1	1	1	0	sport
d2	0	0	1	1	1	gossip, sport
d3	2	0	0	1	0	sport
d4	1	1	1	0	0	politics

- Each attribute values represents the frequency of a word

TEXT CLASSIFICATION vs DATA CLASSIFICATION

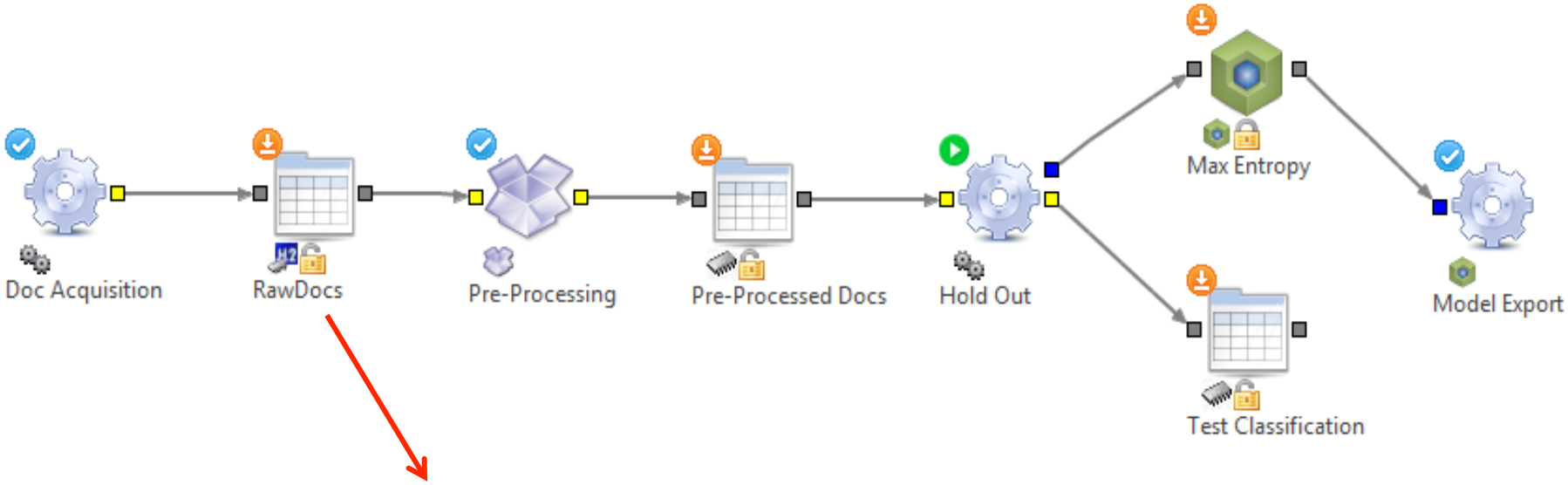
- Unlike data classification, TC is
  - Multi-label: one document may belong to different categories
  - High dimensional: thousands of attributes

TEXT CLASSIFICATION

	w_1	w_2	w_3	w_4	w_5	…	$w_{100.000}$	**Class**
d1	0	1	1	1	0	…	1	Sport, politics
d2	0	0	1	1	1	…	0	gossip
d3	1	0	0	1	0	…	0	Sport, gossip
d4	1	1	1	0	0	…	1	politics

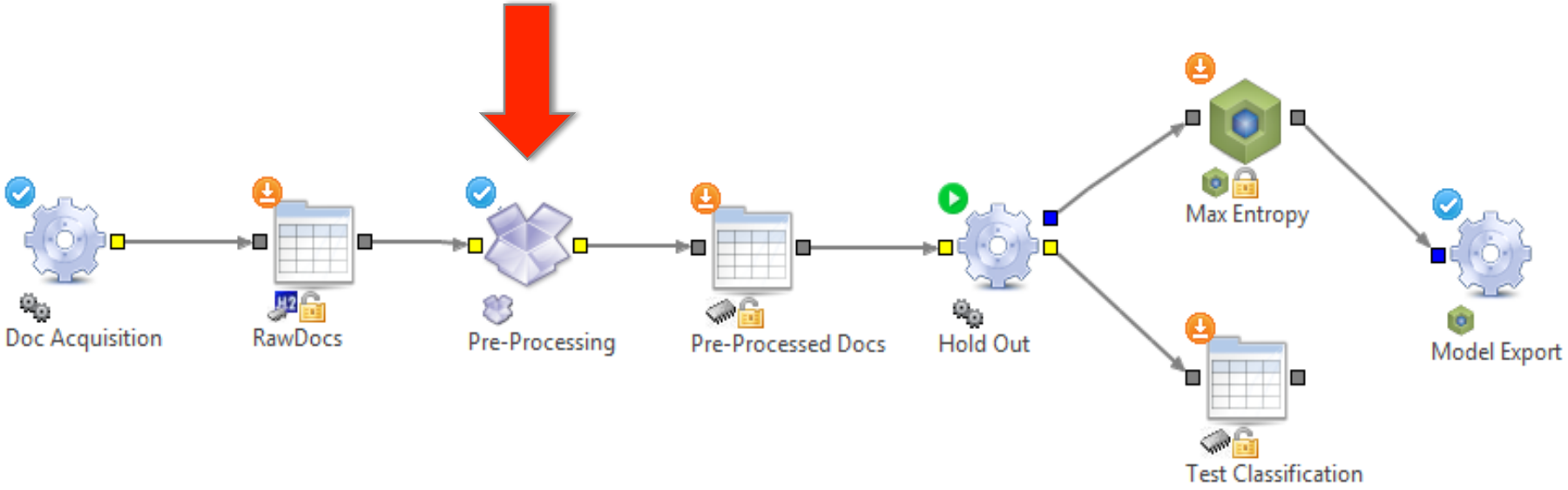
- A training set may be multi-label
- W_1 … $w_{100.000}$ are all words occurring in the docs of the training set

TEXT CLASSIFICATION PROCESS



	w_1	w_2	w_3	w_4	w_5	…	$w_{100.000}$	Class
d1	0	1	1	1	0	…	1	Sport, politics
d2	0	0	1	1	1	…	0	gossip
d3	1	0	0	1	0	…	0	Sport, gossip
d4	1	1	1	0	0	…	1	politics

TEXT CLASSIFICATION PROCESS



DOCUMENT PRE-PROCESSING

- Question: which words should be selected as representative features?
- Pre-processing main steps
  - N-gram extraction
  - Stop-words removal
  - Lemmatization
  - Feature selection

DOCUMENT PRE-PROCESSING

- n-gram: sequence of n consecutive words, e.g., in a medical domain, a 3-gram is

  "immunologic deficiency syndromes"

- This 3-gram is more meaningful than each single word – its meaning is very different from that of "deficiency" alone
- A document is in general a bag on n-grams

DOCUMENT PRE-PROCESSING

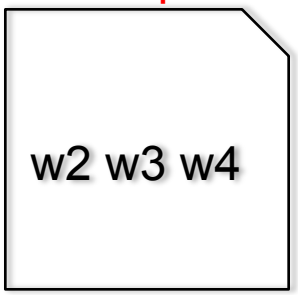
- Lemma is the canonical form, dictionary form, or citation form of a set of words
- Lemmatization: reduction to basic forms, e.g., jumps => jump, working => work
- Stop-word removal
  - Ignore common words, e.g., the, a, to, that, and, at, …. (high entropy words)

DOCUMENT PRE-PROCESSING

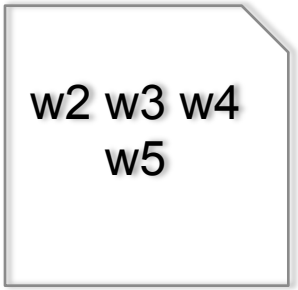
- **Features Selection**
  - Select features with high discriminative power, i.e., features inducing a high information gain (reduction of entropy)
  - Example: if the word “house” is evenly distributed across the various classes (high entropy) it is not useful for the purpose of discriminating the documents of a class w.r.t. those of the other classes
  - FS is beneficial in that it
    - Reduces noise, thus improving the learning effectiveness
    - Reduces the high-dimensionality problem, thus increasing efficiency
  - FS functions: Information Gain, CHI square, IG, Odds Ratio, etc.

DOCUMENT PRE-PROCESSING

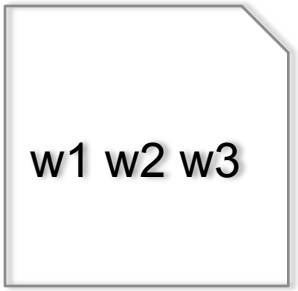
d1 - sport



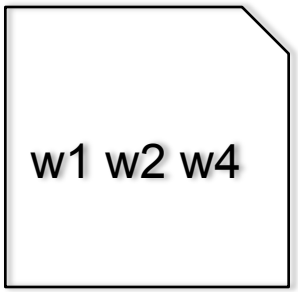
d2 – gossip, sport



d3 - sport



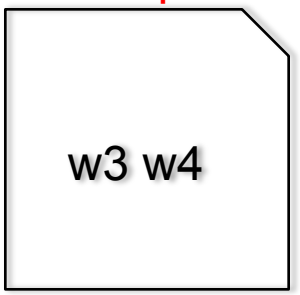
d4 - politics



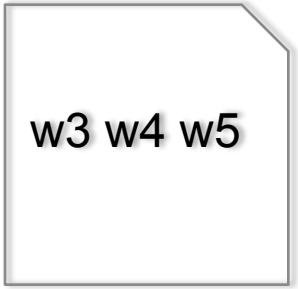
- w1 is an article, e.g., the
- w2 is uniformly distributed over all classes, so it has no discriminating power (high entropy)
- w3 occurs only within documents under sport
- w5 occurs only within documents about gossip
- …

DOCUMENT PRE-PROCESSING

d1 - sport

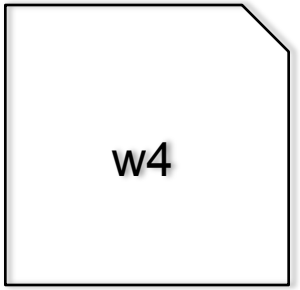


d2 – gossip, sport



d3 - sport

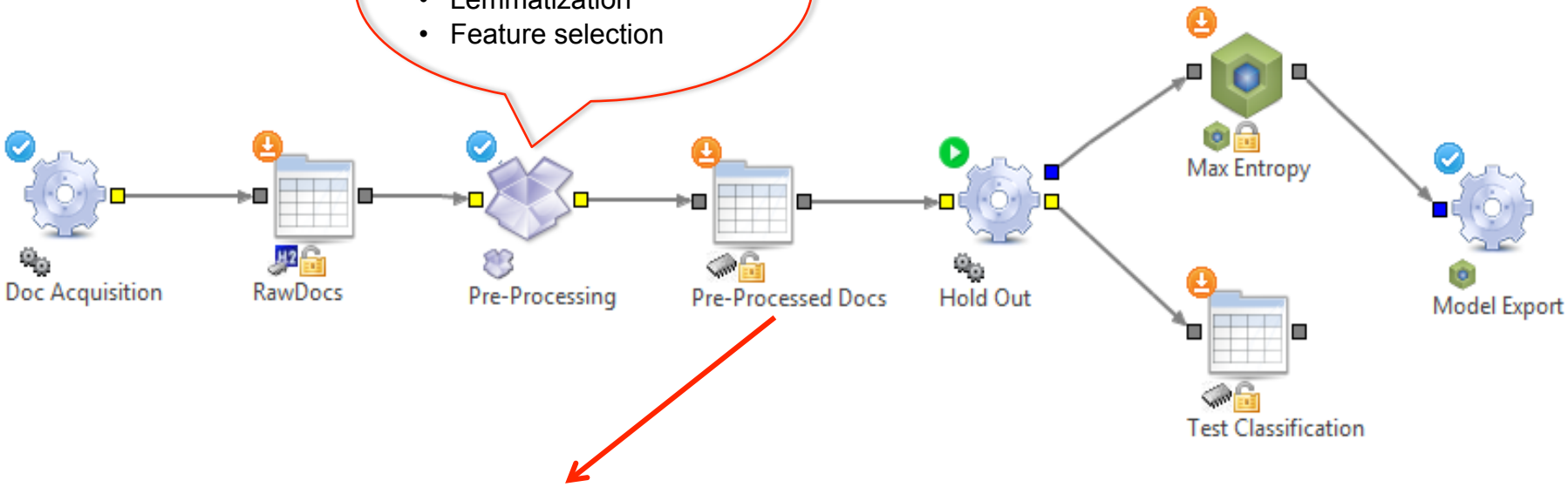
d4 - politics



After feature selection

TEXT CLASSIFICATION PROCESS

- N-gram extraction
- Stop-words removal
- Lemmatization
- Feature selection



	ng_1	ng_2	ng_3	ng_4	…	$ng_{1.000}$	Class
d1	0	1	1	1	…	1	Sport, politics
d2	0	0	1	1	…	0	gossip
d3	1	0	0	1	…	0	Sport, gossip
d4	1	1	1	0	…	1	politics

MODEL INDUCTION

- Given the above representation of a training set, either traditional classifiers like Ripper, C4.5, Naïve Bayes, SVMs, etc., or text-specific classifiers, like MaxEntropy, CNB, etc., can be used for the purpose of TC
- RIPPER: *soccer team=1 and goal = 1 → sport* – if "soccer team" and "goal" occur in a document d then classify d under sport
- Naïve Bayes: *p(sport|d)*, *p(politics|d)*, …,

  where $d = \langle ng_1, \ldots, ng_n \rangle$

TEXT CATEGORIZATION APPLICATIONS

- E-mail spam filtering
- Categorize newspaper articles and newswires into Topics
- Organize Web pages into hierarchical categories
- Sort journals and abstracts by subject categories (e.g., MEDLINE, etc.)
- …