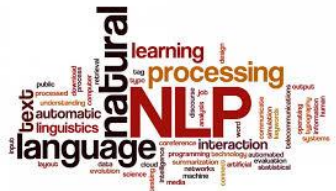


Natural Language Processing

Text Classification

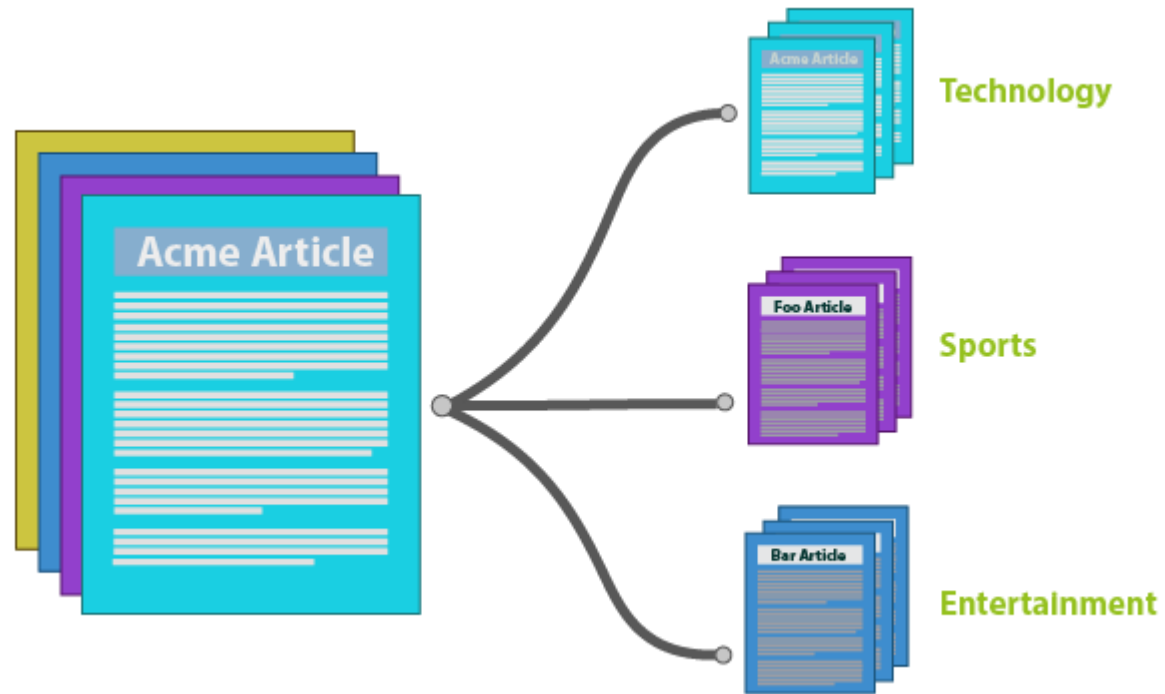


Agenda

- Text Classification - Examples
- Text Classification - Definition
- Role of Classification in Machine Learning

Goal

- is the classification of documents into a fixed number of predefined categories.
- It has been applied successfully multiple times and is integrated in our everyday lives.



Examples

- Newspaper articles and academic papers are often organized by subject or field.
- Text classification provides a solution in automatically organising countless articles and papers.
- Spam filtering.
 - It is becoming more and more common to receive unsolicited emails daily. By using text classification to label these emails as spam, they can be filtered automatically.
- Automated threat detection in social media.
 - classification to detect negative attitudes towards law enforcement.
 - They classify comments to two categories, one with negative attitude comments and one with neutral attitude comments.

Is this spam?

Subject: Important notice!

From: Stanford University <newsforum@stanford.edu>

Date: October 28, 2011 12:34:16 PM PDT

To: undisclosed-recipients;;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

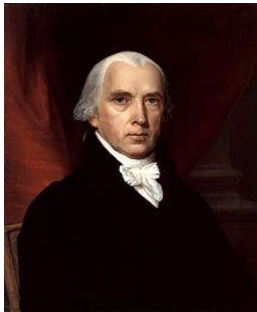
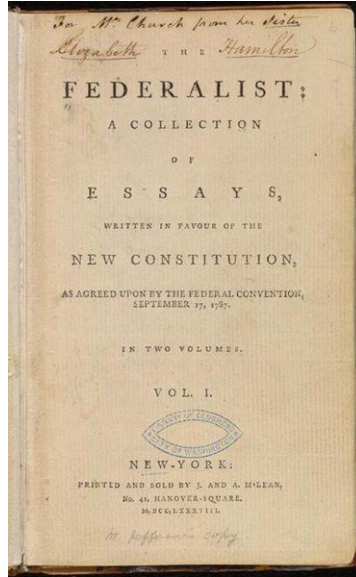
<http://www.123contactform.com/contact-form-StanfordNew1-236335.html>

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information about the new services.

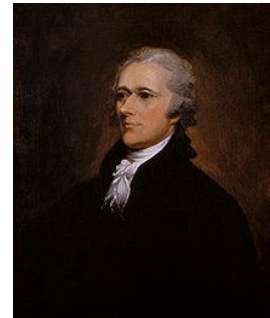
© Stanford University. All Rights Reserved.

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

Male or female author?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam...
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets...

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.

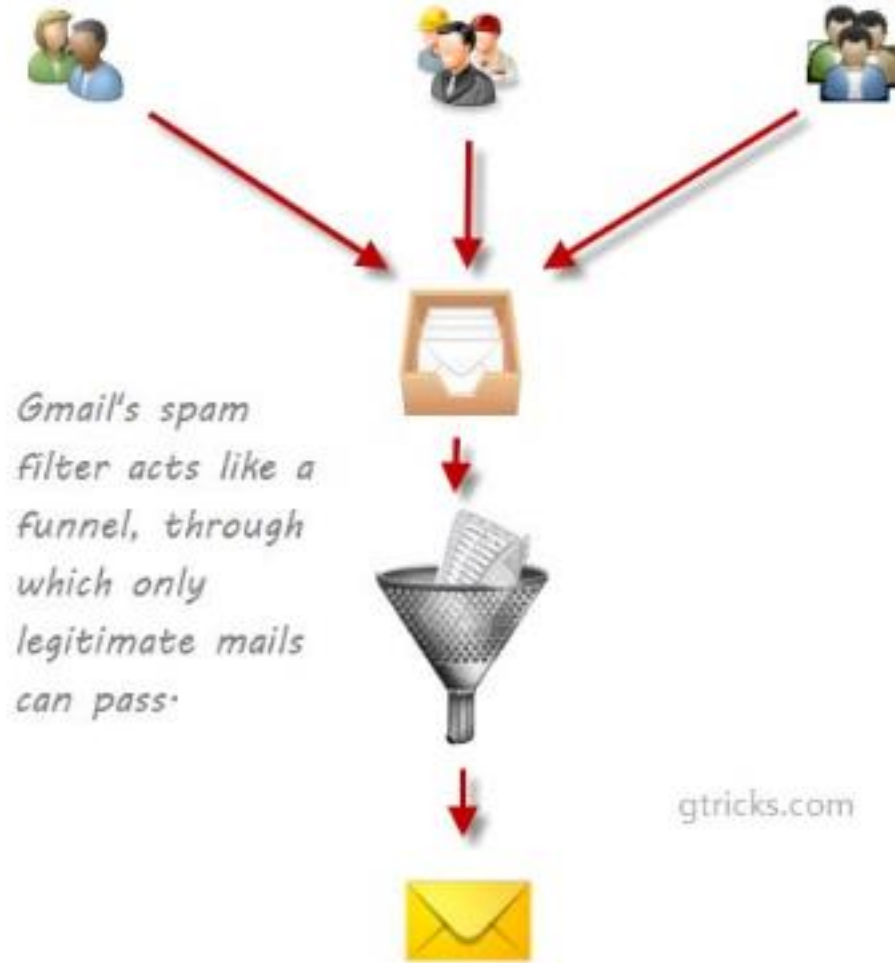
What is the subject of this article?

MEDLINE Article



MeSH Subject Category Hierarchy

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...



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 $C = \{c_1, c_2, \dots, c_J\}$
- *Output:* a predicted class $c \in C$

Classification Methods: Supervised Machine Learning

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

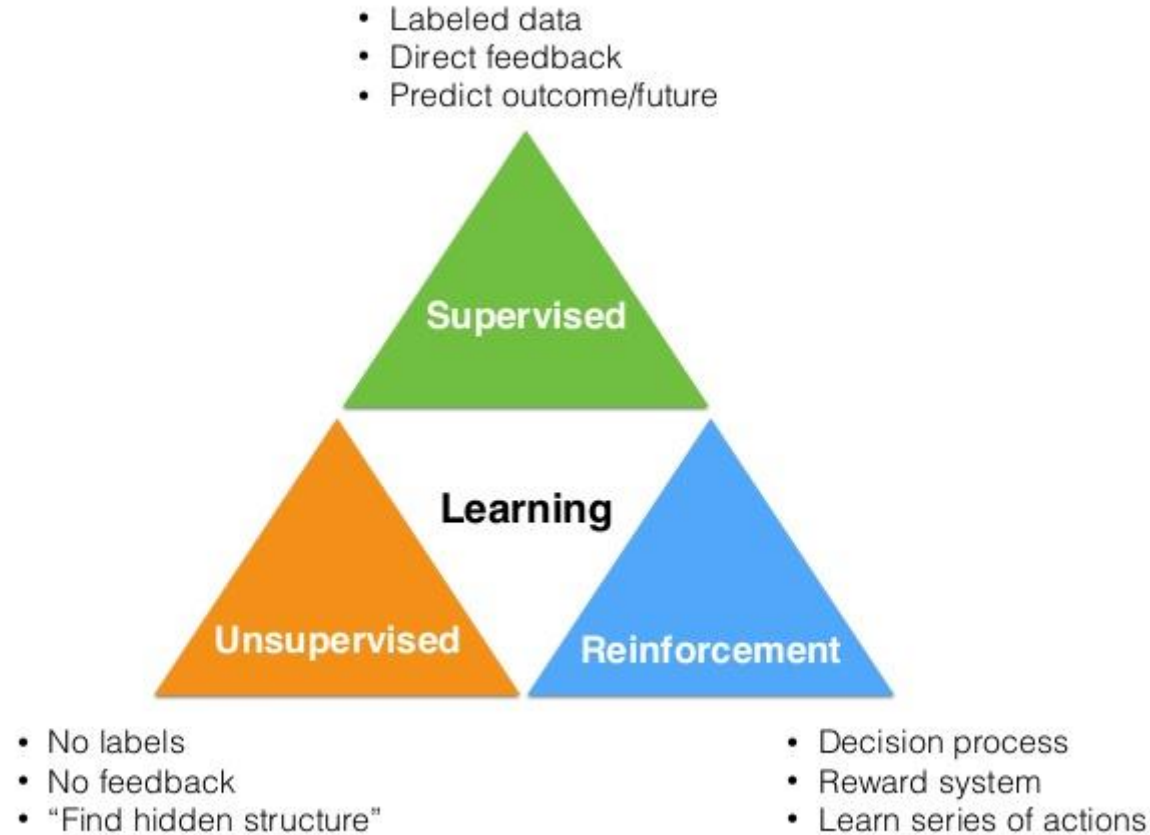
Machine Learning

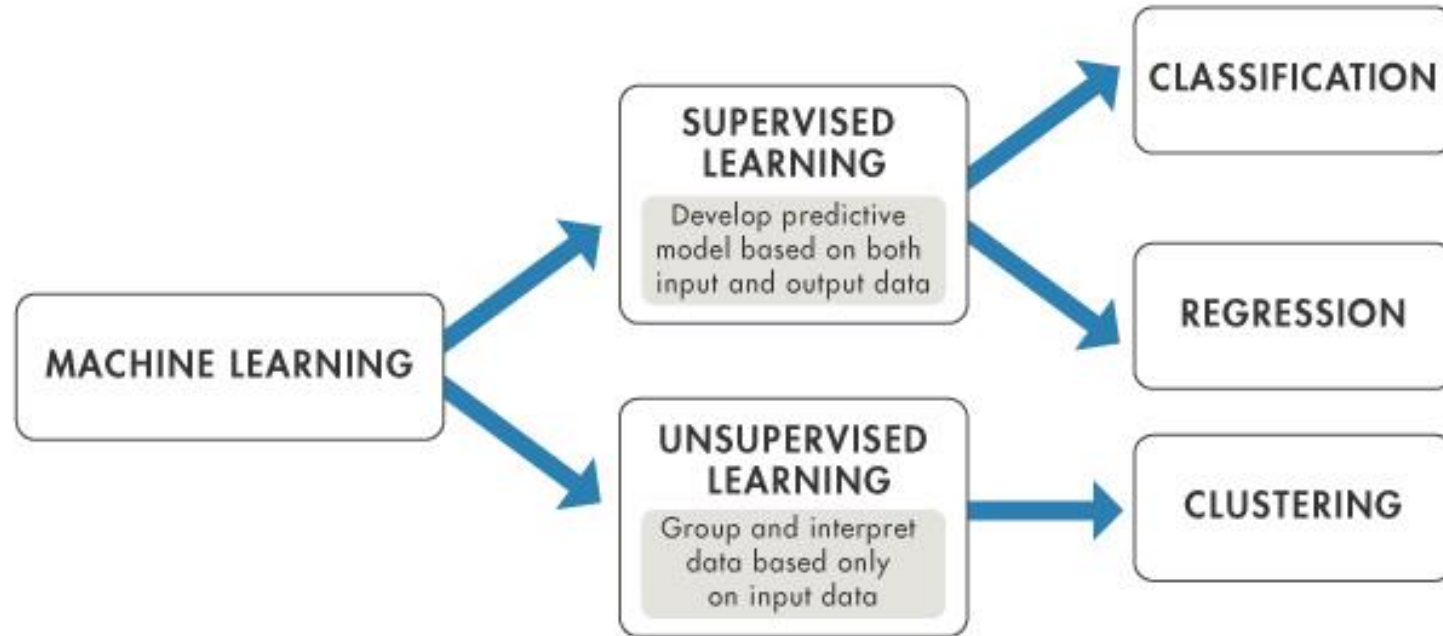
is the field of study that gives computers the ability to learn without being explicitly programmed.

Machine Learning - Types

- **Supervised machine learning:**
 - The program is “trained” on a pre-defined set of “training examples”, which then facilitate its ability to reach an accurate conclusion when given new data.
- **Unsupervised machine learning:**
 - The program is given a bunch of data and must find patterns and relationships therein.
- **Reinforcement learning:**
 - Reinforcement learning is a type of machine learning algorithm that allows the agent to decide the best next action based on its current state, by learning behaviours that will maximize the reward.

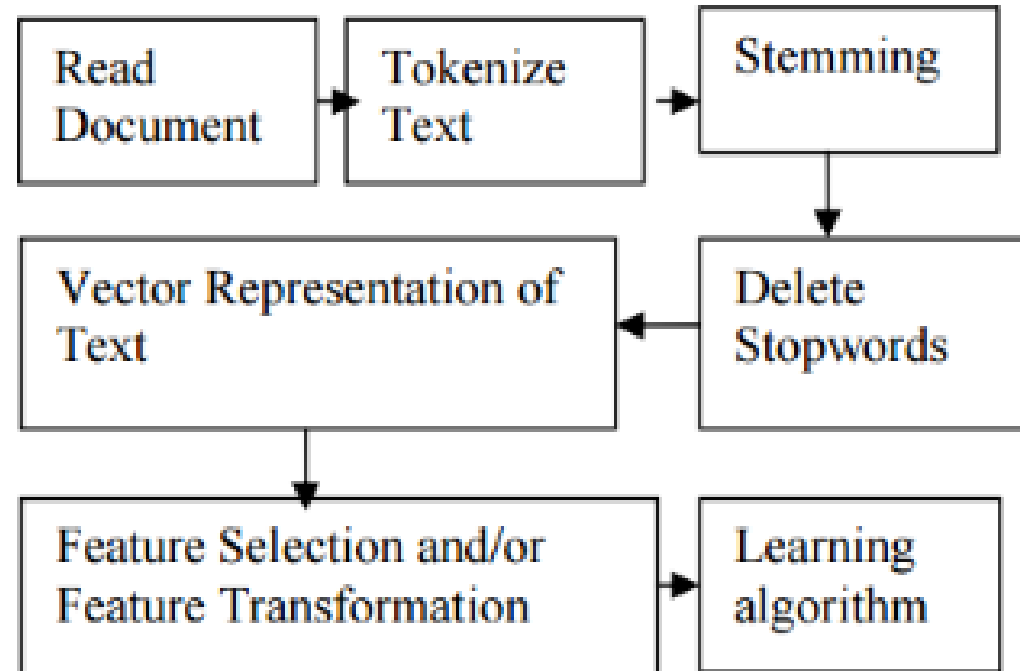
Machine Learning - Types





Text classification - 2 Sections

- pre-processing and
- classification.



Text Classification

- Text Classification is an example of supervised machine learning task
- since a labelled dataset containing text documents and their labels is used for train a classifier.

Machine Learning – Primary Goal

- Highly accurate predictions on test data

Supervised Learning

- In a typical supervised learning scenario,
- a training set is given and
- the goal is to form a description that can be used to predict previously unseen examples

Naive Bayes

- Naive Bayes (NB) is a simple method based on the Bayes rule.
- The probability each feature contributes independently to the final probability to be a class, each one has its distribution.
- The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors.

$$P(c | x) = \frac{P(x | c) P(c)}{P(x)}$$

Diagram illustrating the components of the equation:

- $P(c | x)$ is labeled as **Posterior Probability** (indicated by a downward arrow).
- $P(x | c)$ is labeled as **Likelihood** (indicated by an upward arrow).
- $P(c)$ is labeled as **Class Prior Probability** (indicated by an upward arrow).
- $P(x)$ is labeled as **Predictor Prior Probability** (indicated by a downward arrow).

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \cdots \times P(x_n | c) \times P(c)$$