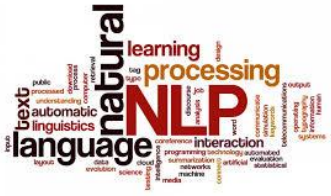
Natural Language Processing

Text Classification

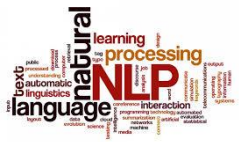


Agenda

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

Machine Learning

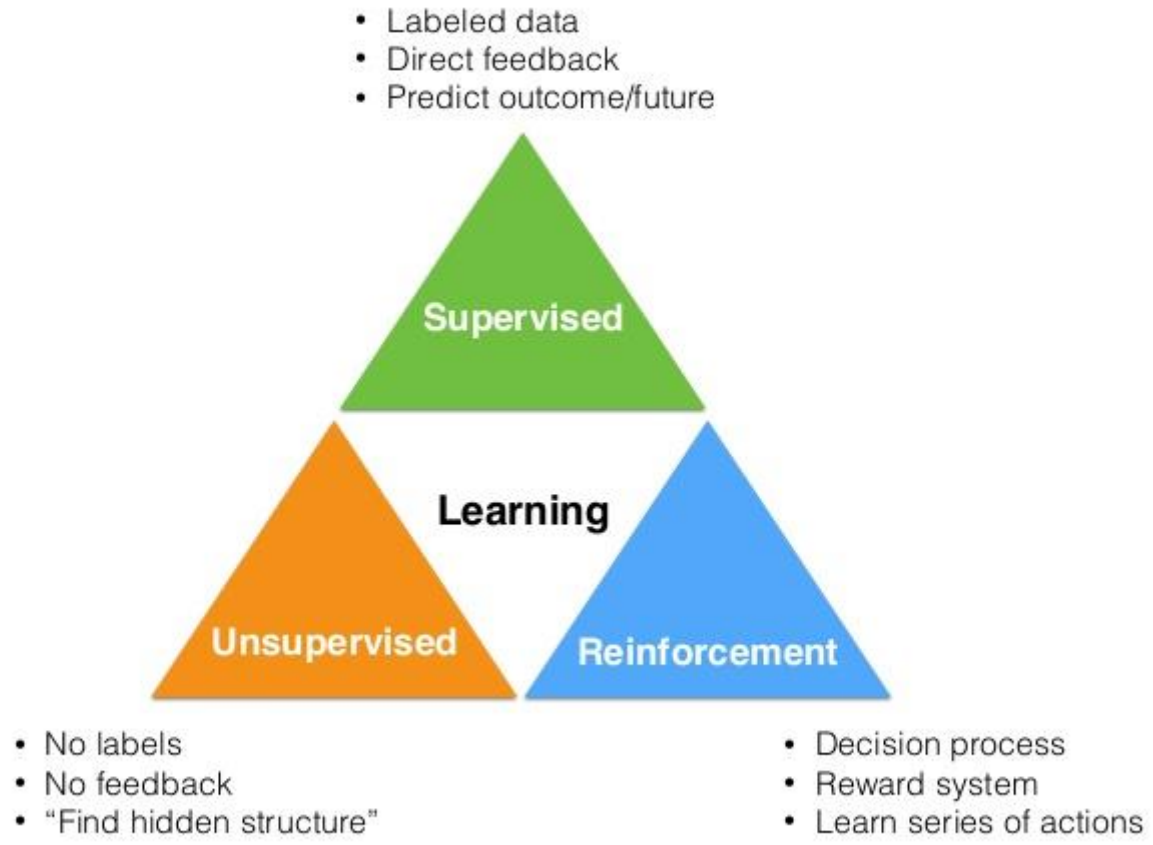
- Machines imitating and adapting human like behavior.



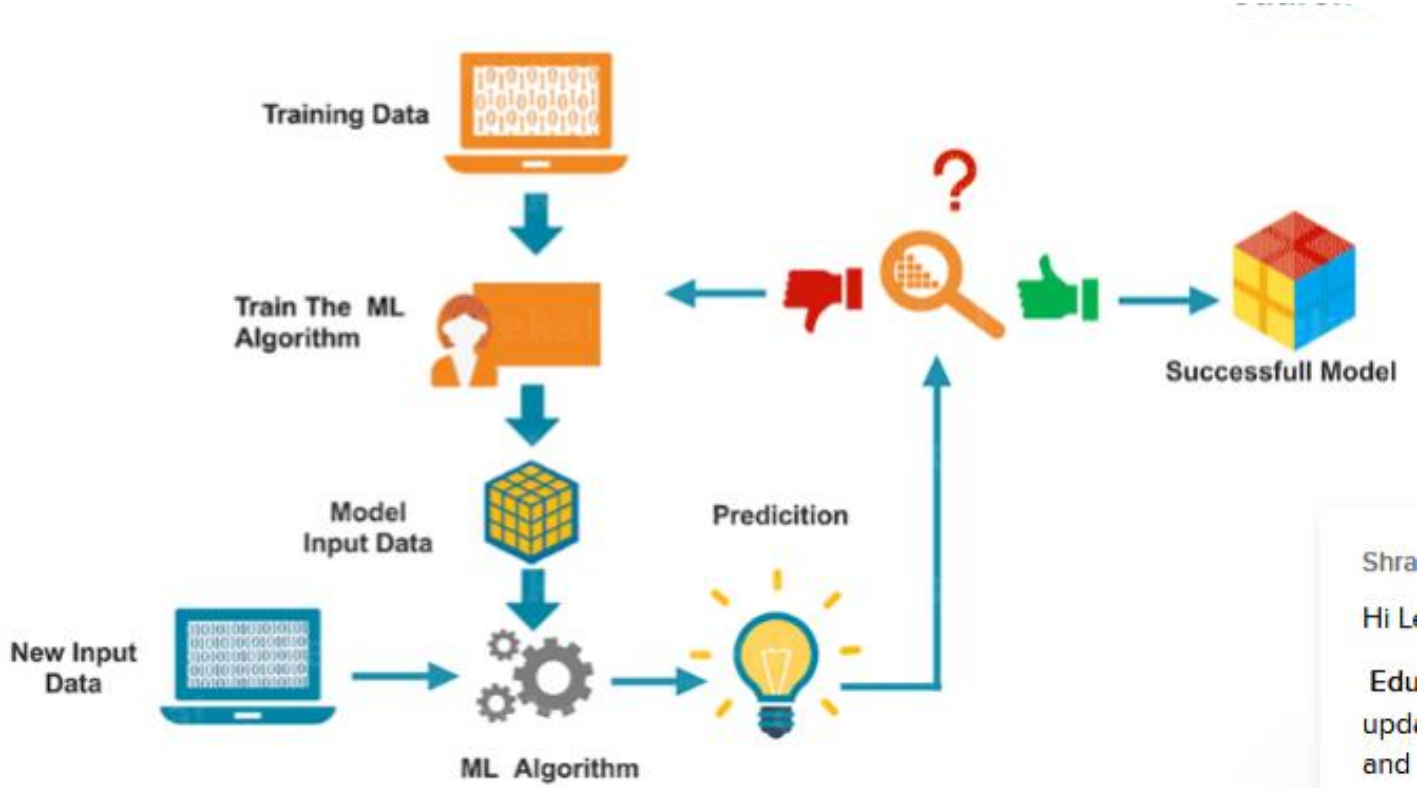
Machine Learning - Types

- Supervised Learning – Train Me!
- Unsupervised Learning – I am self sufficient in learning
- Reinforcement Learning – My life My rules! (Hit & Trial)

Machine Learning - Types



Machine Learning - Types



Supervised Learning

- Finds patterns (and develops predictive models) using both, input data and output data.
- All Supervised Learning techniques are a form of either
- Classification or
- Regression.

Supervised Learning - Classification

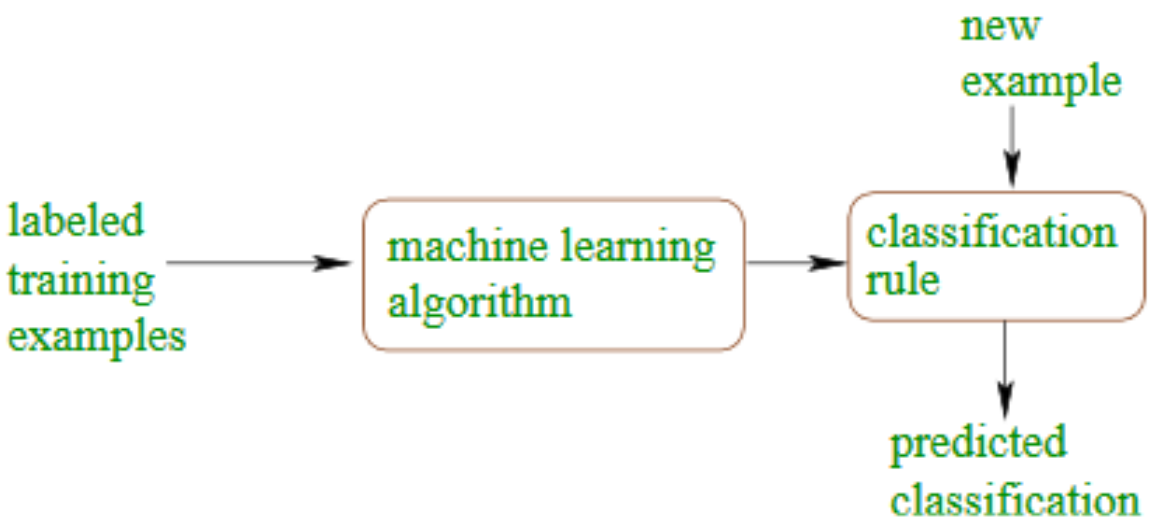
- Classification is used for predicting discrete responses.
- Whether India will WIN or LOSE a Cricket match?
- Whether an email is SPAM or GENUINE?
- WIN, LOSE, SPAM, GENUINE are the predefined classes.
- And output has to fall among these depending on the input

Supervised Learning - Regression

- Regression is used for predicting continuous responses.
- For example:
- Trend in stock market prices, Weather forecast, etc.

Machine Learning

- studies how to automatically learn to make accurate predictions based on past observations
- Classification problems:
  - classify examples into given set of categories

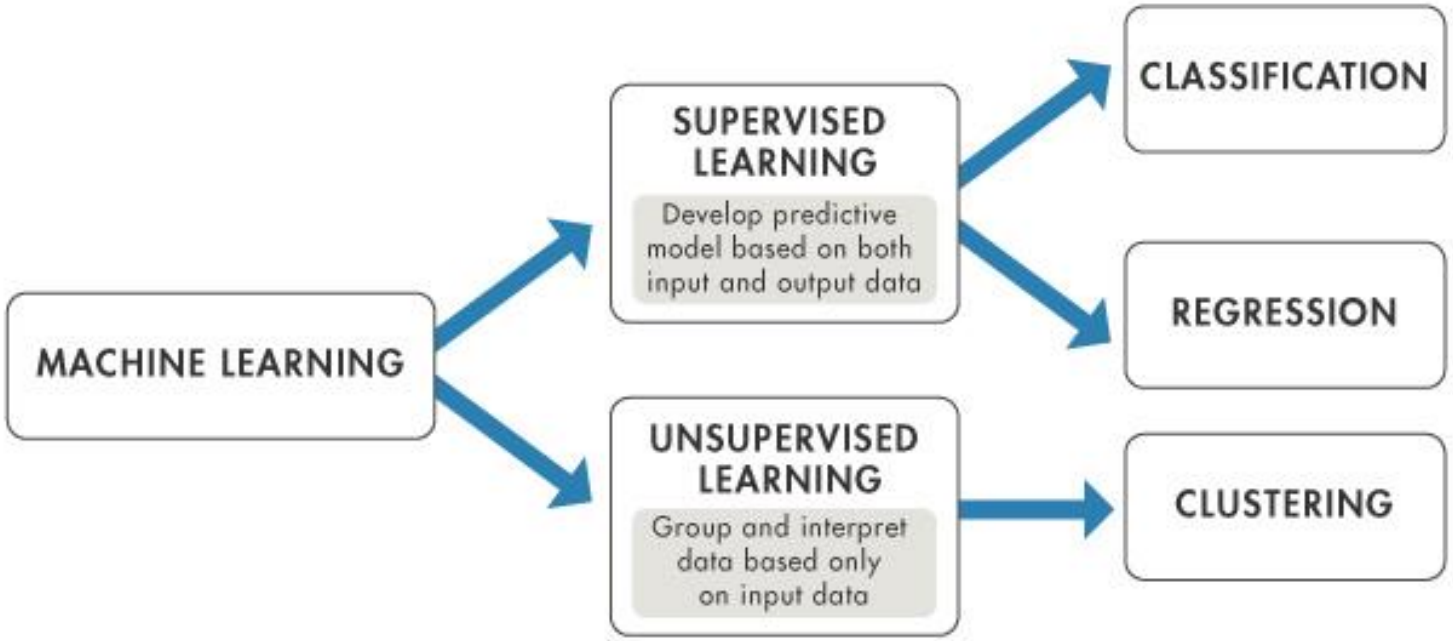


Machine Learning – Primary Goal

- Highly accurate predictions on test data

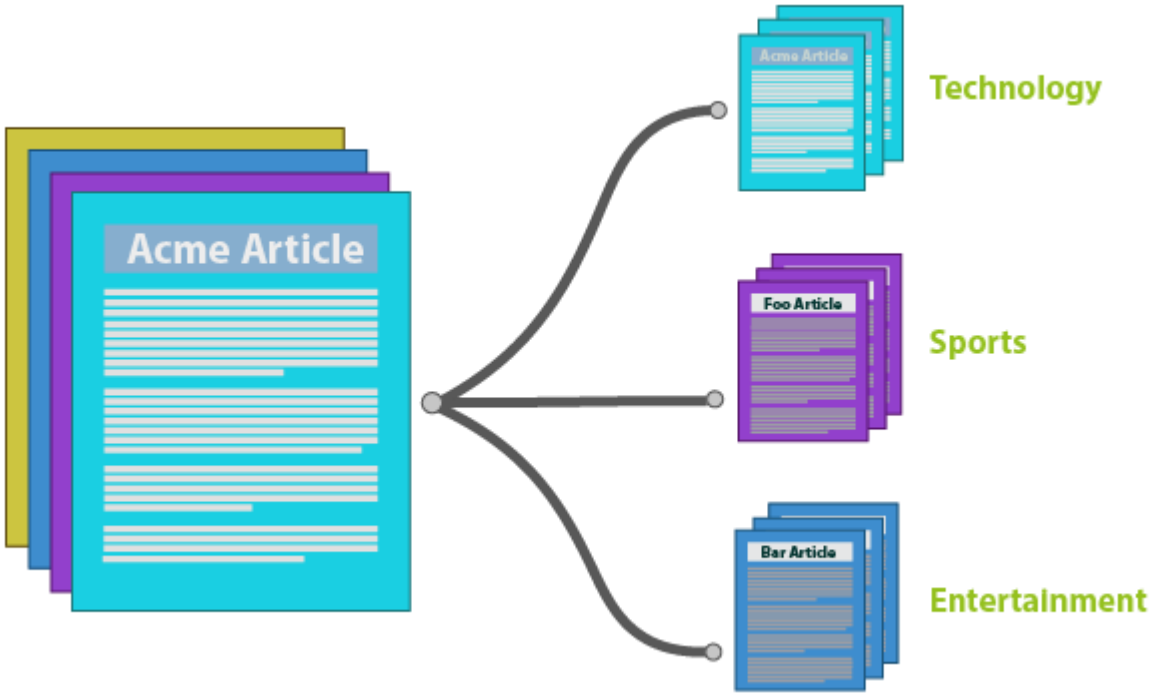
Supervised Learning

- 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


MACHINE LEARNING
SUPERVISED LEARNING
Develop predictive model based on both input and output data
UNSUPERVISED LEARNING
Group and interpret data based only on input data
CLASSIFICATION
REGRESSION
CLUSTERING

Goal - Text Classification

- 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.


Acme Article
Acme Article
Technology
Foo Article
Sports
Bar Article
Entertainment

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!

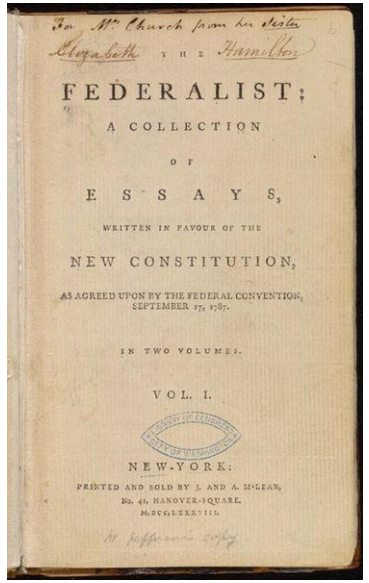
You can now access the latest news by using the link below to login to Stanford University News Forum.

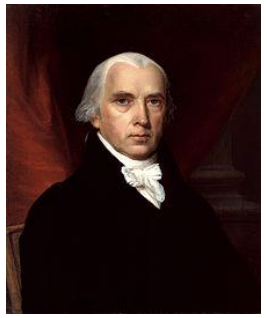
http://www.123contactform.com/contact-form-StanfordNew1-236335.html

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.

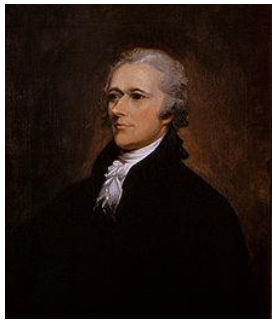
Who wrote which Federalist papers?

- 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

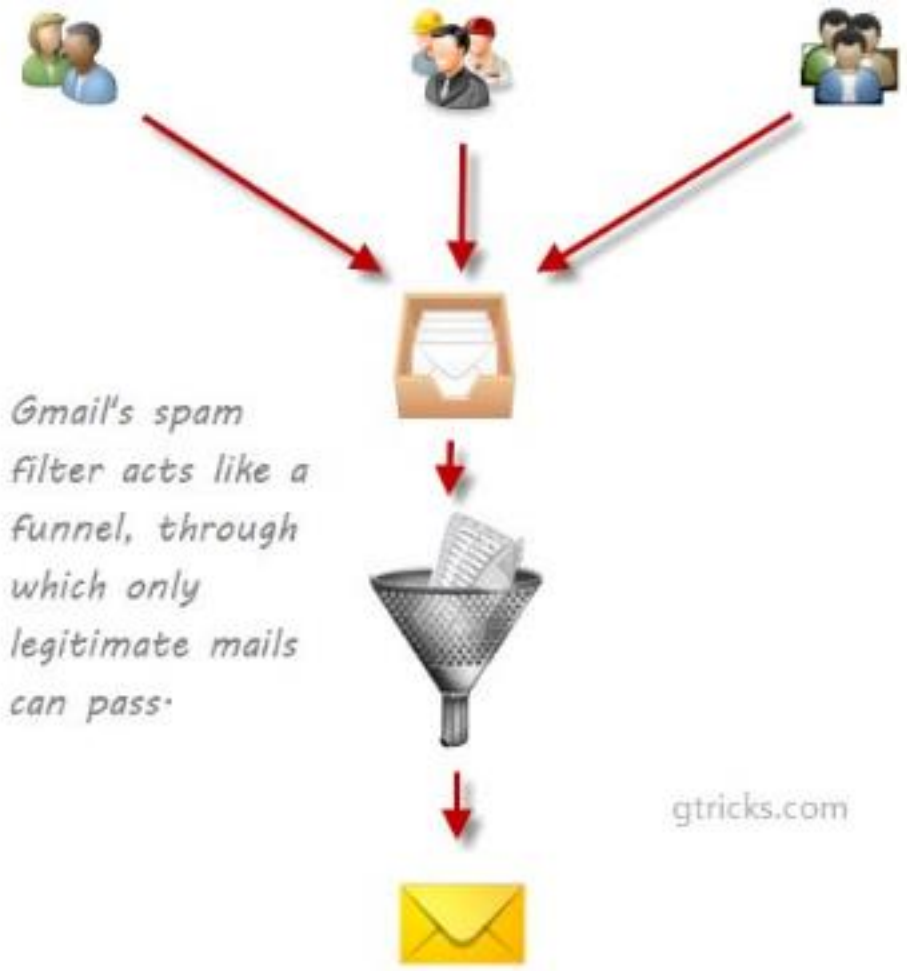
Brain and Cognition

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

?

MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …


Gmail's spam filter acts like a funnel, through which only legitimate mails can pass.
gtricks.com

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 $C = \{c_1, c_2, \ldots, c_J\}$

- *Output*: a predicted class $c \in C$

Classification Methods: Supervised Machine Learning

- *Input:*
  - a document *d*
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
  - A training set of *m* hand-labeled documents $(d_1,c_1),\ldots,(d_m,c_m)$
- *Output:*
  - a learned classifier $\gamma: d \rightarrow c$

Classification Methods: Supervised Machine Learning

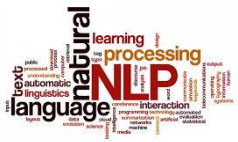
Any kind of classifier

- 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.

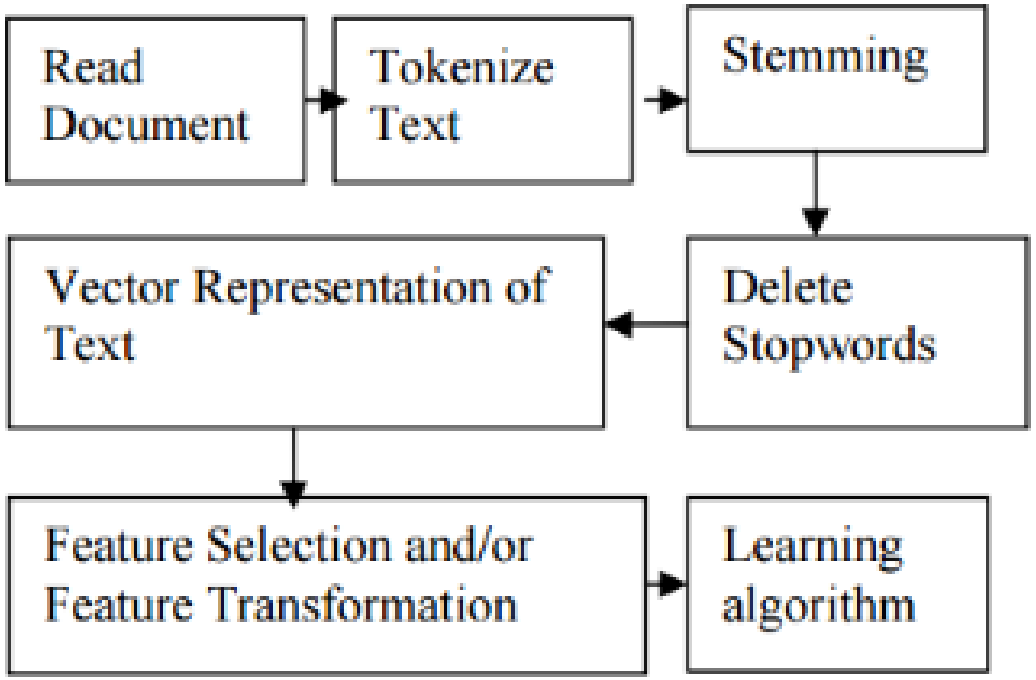


Machine Learning - Types

- 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.

Text classification - 2 Sections

- pre-processing and
- classification.



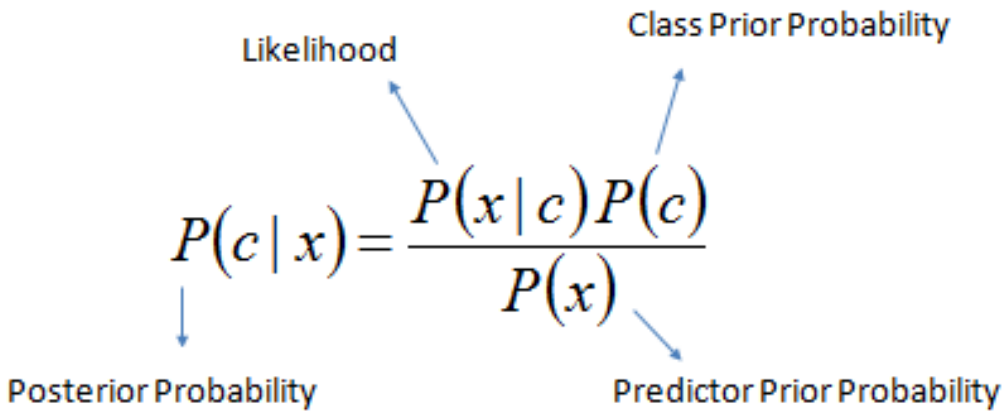
Text Classification

- 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.

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.

Algorithm



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$