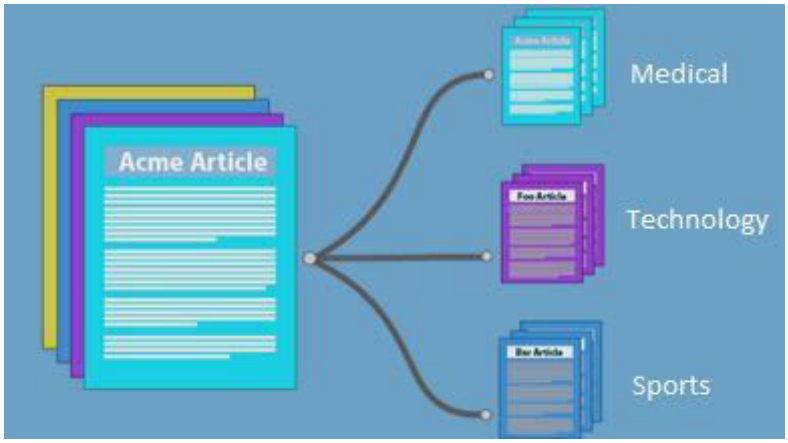
Text Classification or Categorization

- It has been investigated by many researchers over more than past 2 decades. Due to the extreme increase in online textual information, e.g. Email messages, online news, web pages, as well as a huge number of resources for scientific online abstracts such as MEDLINE, there is an ever-growing demand for Text Classification.
- It is an interesting questions how to achieve high performance in the task of assigning multiple topics to documents in a targeted domain and how to make the most of the multi-topical features of the documents.



- Text categorization is the grouping of documents into a fixed number of predefined classes.
- Each document can be in multiple, exactly one, or no category at all.
- Using machine learning, the objective is to classifiers from examples which perform the category tasks automatically. This is a supervised learning problem. Since categories may overlap, each category is treated as a separate group.

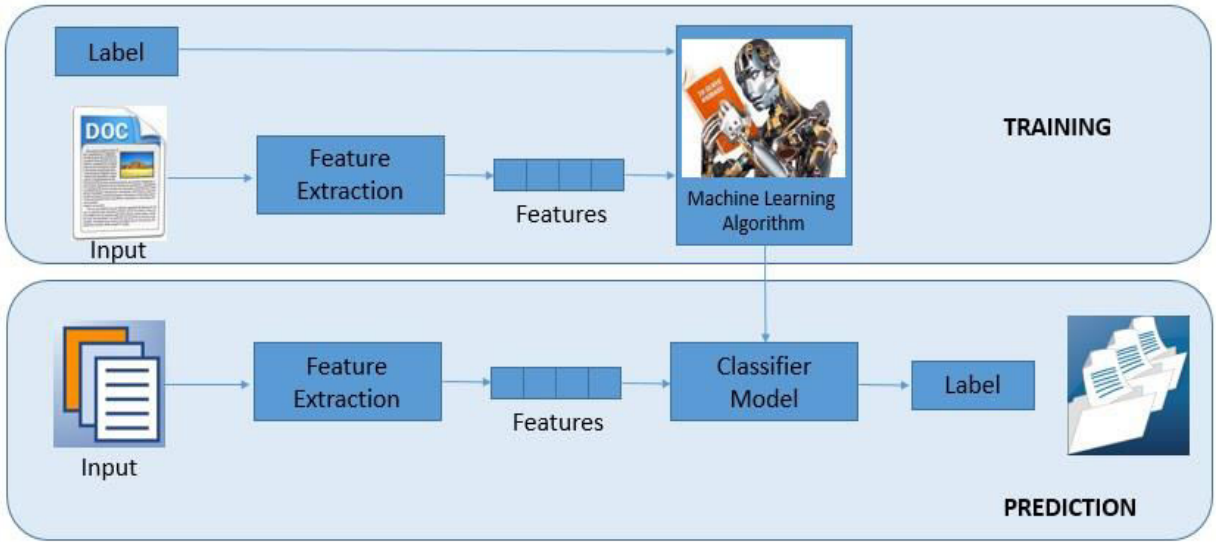
Supervised Learning for Text Classification.

PART-I: **Training**

1. During training, a feature extractor is used to transform each input value to a feature set.
2. These feature sets, which capture the basic information about each input that should be used to categorize it.
3. Pairs of feature sets and labels are fed into the machine learning algorithm to produce a model.

PART-II: Prediction

1. During prediction, the same feature extractor is used to transform unobserved inputs to feature sets. These feature sets are then fed into the model, which produces predicted labels.



Process flow for the Text classification:
For performing the Text classification using Naïve Bayes includes,

Step 1: Reading the Data from .csv file.
Step 2: Divide the dataset into two parts as training dataset and testing dataset.
Step 3: Create a corpus for training dataset and testing dataset.
Step 4: Performing the Data processing transformation on the training dataset and testing datasets.

1. Transform characters to lower case.
2. Converting to Plain Text Document.
3. Remove punctuation marks.
4. Remove digits from the documents.
5. Remove from the documents words which we find redundant for text mining (e.g. Pronouns, conjunctions). We set this words as stopwords("English") which is a built-in list for English language.
6. Remove extra whitespaces from the documents.

Step 5: Now create the "Term document matrix". It describes the frequency of each term in each document in the corpus and performs the transposition of it.
Step 6: Train Naïve Bayes model using transposed "Term document matrix" data and Target class vector.
Step 7: Apply the prediction on generated model for testing dataset.

Naive Bayes Classifier:

- The Naive Bayes classifier is a simple probabilistic classifier which is based on Bayes theorem with strong and naive independence assumptions.
- It is one of the most basic text classification techniques with various applications in email spam detection, document categorization, language detection, sentiment detection and automatic medical diagnosis.
- It is one of the most basic text classification techniques used in various applications. It is highly scalable.

Posterior probability:

$$P(b_i|C) = \prod_{t=1}^{|V|} [b_{it} P(w_t \mid C) + (1-b_{it}) (1-P(w_t \mid C))]$$

Likelihood:

$$\hat{P}(w_t|C=k) = \frac{nk(wt)}{Nk}$$

Prior Probability:

$$\hat{P}(C=k) = \frac{Nk}{N}$$

Application of Text Classification:

1. Sort journals and abstracts by subject groups (e.g., MEDLINE, etc.).
2. Spam filtering, a process which tries to discriminate E-mail spam messages from authentic emails.
3. Language identification, automatically determining the linguistic of a text.
4. Sentiment analysis, determining the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document.
5. Article triage, selecting articles that are relevant for manual literature curation.

EXAMPLE:

- Consider a set of documents, each of which is related either to **Sports (S) or to Informatics (I).**
- Given a training set of 11 documents, we would like to estimate a Naive Bayes classifier, using the Bernoulli document model, to classify unlabelled documents as S or I.

We define a vocabulary of eight words,

V= {w1=goal, w2=tutor, w3=variance, w4=speed, w5=drink, w6=defence, w7=performance, w8=field}

- Thus, each document is represented as an 8-dimensional binary vector.

- The training data is presented below as a matrix for each class, in which each row represents an 8-dimensional document vector

$$B_{Sport} = \begin{matrix} 1\,0\,0\,0\,1\,1\,1\,1 \\ 0\,0\,1\,0\,1\,1\,0\,0 \\ 0\,1\,0\,1\,0\,1\,1\,0 \\ 1\,0\,0\,1\,0\,1\,0\,1 \\ 1\,0\,0\,0\,1\,0\,1\,1 \\ 0\,0\,1\,1\,0\,0\,1\,1 \end{matrix}$$

$$B_{Inf=} \begin{matrix} 0\,1\,1\,0\,0\,0\,1\,0 \\ 1\,1\,0\,1\,0\,0\,1\,1 \\ 0\,1\,1\,0\,0\,1\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,1\,0\,1\,0\,1\,0 \end{matrix}$$

Classify the following into Sports or Informatics using a Naive Bayes classifier.
1. $b_1 = (1,0,0,1,1,1,0,1)^T$
2. $b_2 = (0,1,1,0,1,0,1,0)^T$

Solution:
The total number of documents in the training set

$$N = 11;\ N_S = 6,\ N_I = 5$$

Estimation of the prior probabilities from the training data as:

$$P(S) = 6/11;\quad P(I) = 5/11$$

The word counts in the training data are:

$n_S(w_1) = 3$ $n_S(w_2) = 1$
$n_S(w_3) = 2$ $n_S(w_4) = 3$
$n_S(w_5) = 3$ $n_S(w_6) = 4$
$n_S(w_7) = 4$ $n_S(w_9) = 4$

$n_I(w_1) = 1$ $n_I(w_2) = 3$
$n_I(w_3) = 3$ $n_I(w_4) = 1$
$n_I(w_5) = 1$ $n_I(w_6) = 1$
$n_I(w_7) = 3$ $n_I(w_8) = 1$

Estimate the word likelihoods:
And for class S:

$P(w_1|S) = 1/2$ $P(w_2|S) = 1/6$
$P(w_3|S) = 1/3$ $P(w_4|S) = 1/2$
$P(w_5|S) = 1/2$ $P(w_6|S) = 2/3$
$P(w_7|S) = 2/3$ $P(w_8|S) = 2/3$

And for class I:

$P(w_1|I) = 1/5$ $P(w_2|I) = 3/5$
$P(w_3|I) = 3/5$ $P(w_4|I) = 1/5$
$P(w_5|I) = 1/5$ $P(w_6|I) = 1/5$
$P(w_7|I) = 3/5$ $P(w_8|I) = 1/5$

The posterior probabilities to classify test vectors:

$b_1 = (1,0,0,1,1,1,0,1)^T$

$$P(S|b_1) \propto P(S)\prod_{t=1}^{8} [b_{1t} P(w_t|S) + (1-b_{1t})(1-P(w_t|S))]$$
$$\propto 6/11(1/2 \times 5/6 \times 2/3 \times 1/2 \times 1/2 \times 2/3 \times 1/3 \times 2/3)$$
$$= 5/891$$
$$= 5.6 \times 10^{-3}$$

$P(I|b_1) \propto P(I)\prod_{t=1}^{8}[b_{1t}P(w_t|I)+(1-b_{1t})(1-P(w_t|I))]$

$\propto 5/11(1/5\times2/5\times2/5\times1/5\times1/5\times1/5\times2/5\times1/5)$
$=8/859375$
$= 9.3\times10^{-6}$

Classify this document S

b_2= (0, 1, 1, 0, 1, 0, 1, 0)^T

$P(S|b_2) \propto P(S)\prod_{t=1}^{8}[b_{2t}P(w_t|S)+(1-b_{2t})(1-P(w_t|S))]$

$\propto 6/11(1/2\times1/6\times1/3\times1/2\times1/2\times1/3\times2/3\times1/3)$
$=12/14256$
$=8.4\times 10^{-4}$

$P(I|b_2) \propto P(I)\prod_{t=1}^{8}[b_{2t}P(w_t|I)+(1-b_{2t})(1-P(w_t|I))]$

$\propto 5/11(4/5\times3/5\times3/5\times4/5\times1/5\times4/5\times3/5\times4/5)$
$=34560/4296875$
$= 8.0\times10^{-3}$

Classify as I.