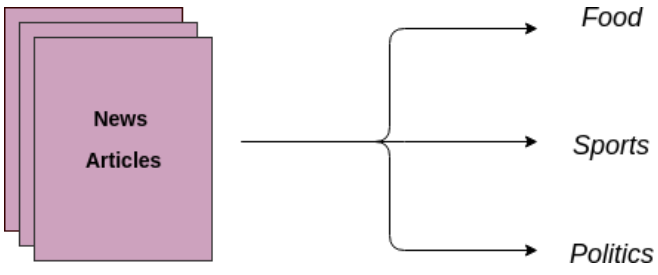
Text Categorization

Text categorization (TC – also known as text classification, or topic spotting) is the task of automatically sorting a set of documents into categories (or classes, or topics) from a predefined set. This task, that falls at the crossroads of information retrieval (IR) and machine learning (ML), has witnessed a booming interest in the last ten years from researchers and developers alike.

This task has several applications, including automated indexing of scientific articles, filing patents into patent directories, selective dissemination of information to information consumers, automated population of hierarchical catalogues of Web resources, spam filtering, identification of document genre, authorship attribution, survey coding, and even automated essay grading.

TC may be formalized as the task of approximating the unknown target function $\Phi : D \times C \rightarrow \{T,F\}$ (that describes how documents ought to be classified) by means of a function $\Phi : D \times C \rightarrow \{T,F\}$ called the classifier, where $C = \{c1, \ldots, c|C|\}$ is a predefined set of categories and D is a (possibly infinite) set of documents. If $\Phi(dj, ci) = T$, then dj is called a positive example (or a member) of ci, while if $\Phi(dj, ci) = F$ it is called a negative example of ci.



Stages of Text Categorization

1) Document Indexing

Document indexing denotes the activity of mapping a document dj into a compact representation of its content that can be directly interpreted

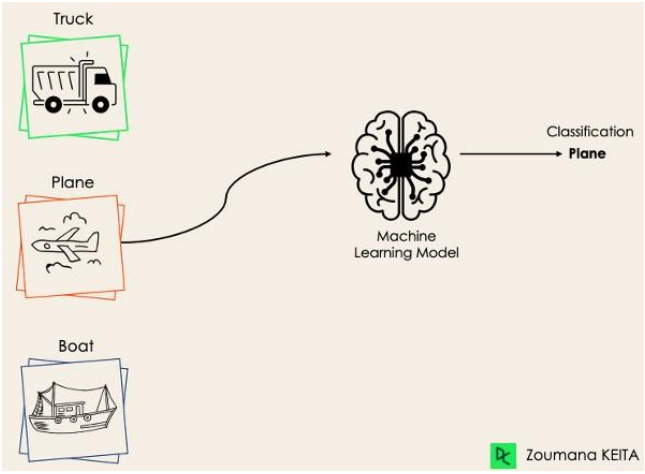
(i) by a classifier building algorithm and

(ii)by a classifier, once it has been built.

	Reference 1	Reference 2	Reference 3	Reference 4
Legal Documents	Client surname	Legal matter	Matter number	Retention date
Medical Records	Hospital/patient number	Surname, first name	Date of birth	Retention date
Business Documents	Invoice number	Client name	Goods	Retention date

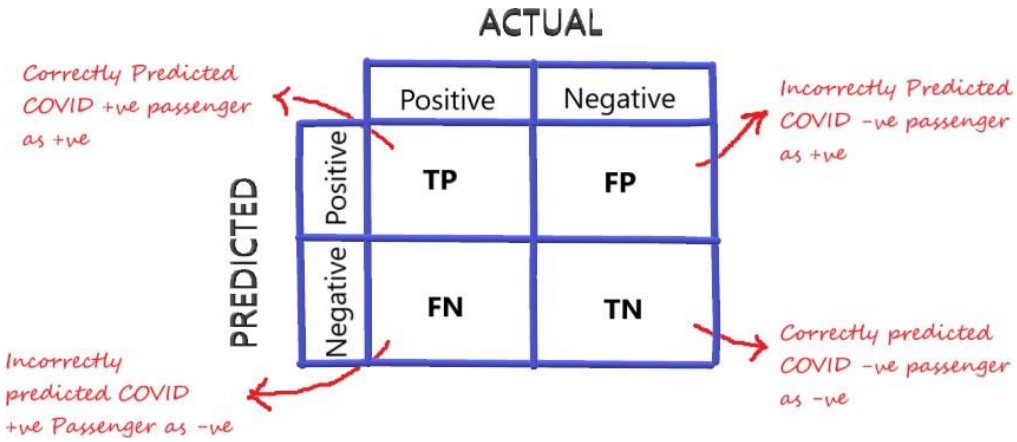
2) Classifier Learning

A text classifier for ci is automatically generated by a general inductive process (the learner) which, by observing the characteristics of a set of documents pre-classified under ci or ci, gleans the characteristics that a new unseen document should have in order to belong to ci



3) Classifier Evaluation

Training efficiency (i.e. average time required to build a classifier Φi), as well as classification efficiency (i.e. average time required to classify a document by means of Φi), and effectiveness (i.e. average correctness of Φi's classification behaviour) are all legitimate measures of success for a learner.

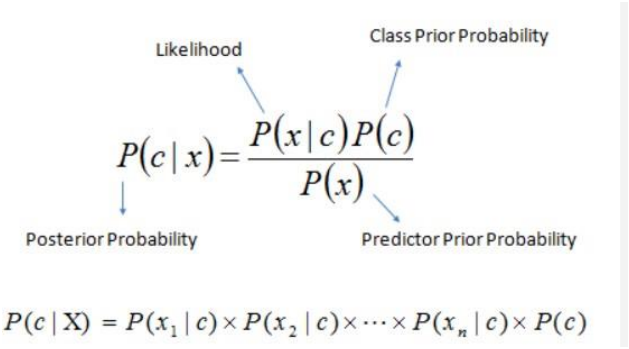


Algorithms:

1. **Naive Bayes Classifier**

   Naive Bayes is the simple algorithm that classifies text based on the probability of occurrence of events. This algorithm is based on the Bayes theorem, which helps in finding the conditional probabilities of events that occurred based on the probabilities of occurrence of each individual event.

	sent	class
0	This is my book	stmt
1	They are novels	stmt
2	have you read this book	question
3	who is the author	question

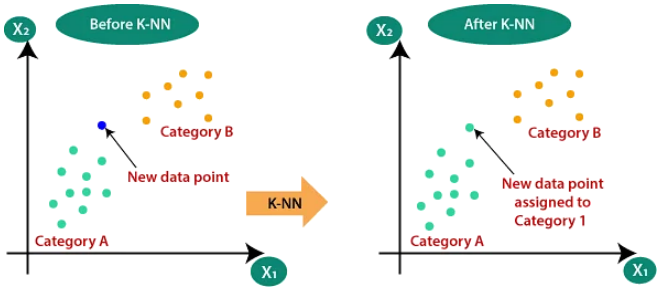


2. KNN

KNN stands for K Nearest Neighbour. A supervised machine learning algorithm classifies the new text by mapping it with the nearest matches in the training data to make predictions. KNN algorithm determines the K nearest neighbors by the closeness and proximity among the training data. The model is trained so that when new data is passed through the model, it can easily match the text to the group or class it belongs to.

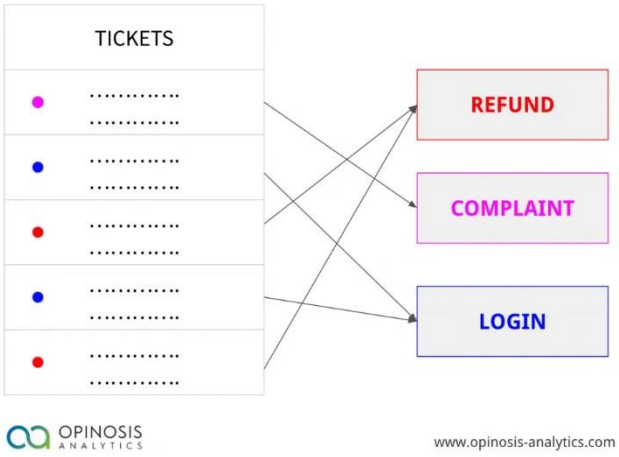
$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

title	label
Which Zodiac Should You Date Based On Your Fav...	clickbait
Guidant announces more defibrillator problems	not-clickbait
Many nations offer material aid to hurricane v...	not-clickbait
Woman Files Complaint Against Bernie Williams	not-clickbait
21 Reasons Why No One Should Have New York Values	clickbait
Washington Train Crash Prompts Safety Warning	not-clickbait

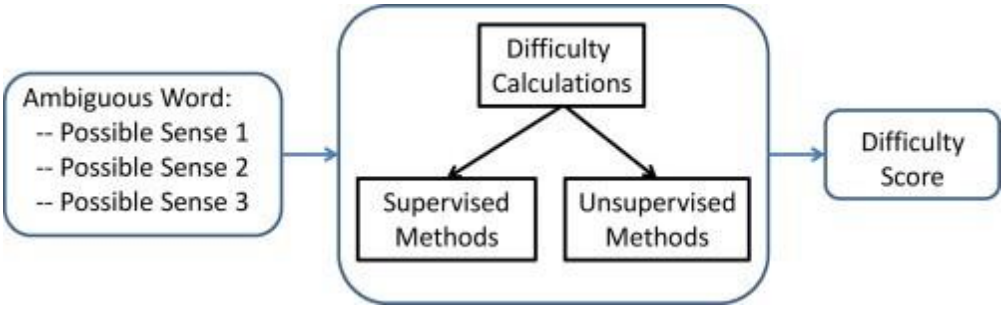


Applications:

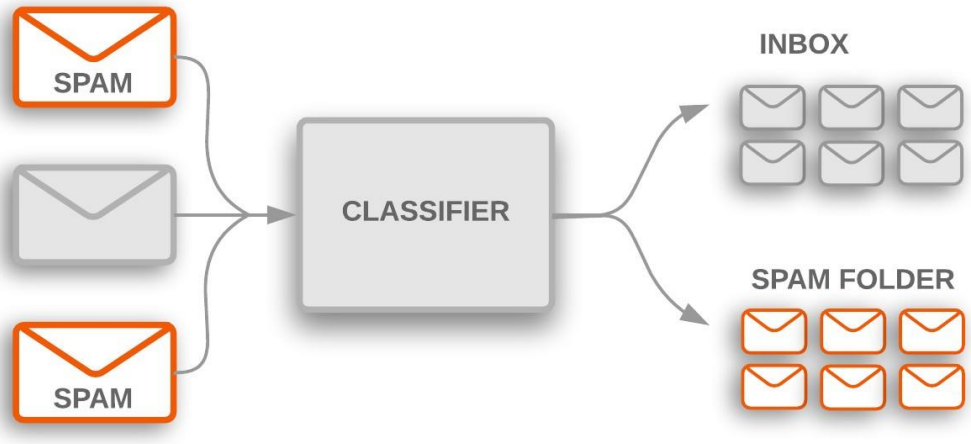
- **Document organization** Indexing with a controlled vocabulary is an instance of the general problem of document base organization. In general, many other issues pertaining to document organization and filing, be it for purposes of the personal organization or structuring of a corporate document base, may be addressed by TC techniques. For instance, at the offices of a newspaper, it might be necessary to classify all past articles in order to ease future retrieval in the case of new events related to the ones described by the past articles. Possible categories might be Home News, International, Money, Lifestyles, and Fashion, but also finer-grained ones such as those organized based on special events.



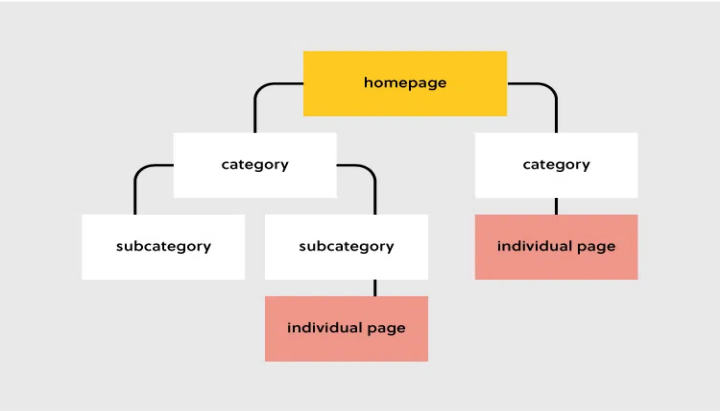
- **Word sense disambiguation** (WSD) is the activity of finding, given the occurrence in a text of an ambiguous (i.e. polysemous or homonymous) word, the sense of this particular word occurrence. For instance, banks may have (at least) two different senses in English, as in the Bank of England (a financial institution) or the bank of River Thames (seashore).



- **Spam filtering** Filtering spam (i.e. unsolicited bulk e-mail) is a task of increased applicative interest that lies at the crossroads between filtering and genre classification. In fact, it has the dynamical character of other filtering applications, such as e-mail filtering, and it cuts across different topics, as genre classification.



- **Hierarchical categorization of Web pages** Web documents are catalogued in this way, rather than issuing a query to a general-purpose Web search engine a searcher may find it easier to first navigate in the hierarchy of categories and then restrict a search to a particular category of interest.Classifying Web pages automatically has obvious advantages, since the manual categorization of a



large enough subset of the Web is unfeasible.