Cairo university

Faculty of Computers and Artificial Intelligence

Graduation Project

1st Semester 2020-2021 Project

Survey

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Name** | **E-mail** | **Group** |
| 20170021 | Ahmed Sayed Ibrahim | ahmed111522@gmail.com | CS-DS-1 |
| 20170136 | Atef Magdy Mitwally | atefmagdy12@gmail.com | CS\_IS\_1 |
| 20170053 | Ashraf Samir Ali | ashrafsamer423@gmail.com | CS-DS-1 |
| 20170002 | Ibrahim Ramadan Abdou | ibrahemramadan130@gmail.com | CS-IS-1 |
| 20170022 | Ahmed Sayed Mansour | mmostafaouda@gmail.com | CS\_IS\_1 |

[**Background**](#_qu5hijx9gn11) **3**

[Machine Learning](#_nimspvu43tr) 3

[Types of ML and Techniques in every one](#_eo23odc153nn) 3

[Supervised Learning](#_h8v3jd2mxqdv) 3

[Classification](#_f400y6a8wxb0) 4

[Regression](#_x8s3k0idlyop) 5

[Unsupervised Learning](#_xmfif51wv8za) 6

[Clustering](#_3lq0qtguwy1j) 6

[Reinforcement Learning](#_cnb787fcnjy3) 8

[Usages of each type](#_5apvr8myjoyj) 8

[Supervised Learning](#_lcczwfah99or) 8

[Unsupervised Learning](#_f4kfptvkjrvp) 9

[Reinforcement Learning](#_puduqotdmia8) 10

[**Classification**](#_8ele1cc2cizg) **10**

[Classification In Machine Learning](#_vxdi36oynop8) 10

[Text analysis can be performed at:](#_5w8lm0e1tvuv) 10

[How Does Automatic Document Classification Work?](#_9z00e48uwnv4) 11

[Different approaches to document classification you can adopt:](#_u9da9sj8611k) 11

[Classification Types:](#_rr9f6t4why4u) 12

[Types of Classification Algorithms:](#_uuiwbyekl9ve) 12

[**Our project(Multi label Document Classification)**](#_8ele1cc2cizg) **14**

[Choosing what kind of classifier to use:](#_2lxmswe69fd) 14

[Large and difficult category taxonomies](#_qip4u71gola9) 15

[Features for text](#_gkckdwpzpnoo) 15

[Document zones in text classification](#_y4h7o7rrdiab) 15

[**References**](#_21w9qxzid5vd) **16**

# Background

## Machine Learning

Machine learning is a part of artificial intelligence (AI) that gives systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn.

The learning process begins with looking at the data, for example, direct experience, or instruction, to look at patterns in data and make better decisions in the future based on the examples we provide. The main purpose is to allow computers to read automatically without human intervention or assistance and to correct actions accordingly.

However, using old machine learning algorithms, text is considered a sequence of key words; instead, a method based on semantic analysis mimics a person's ability to understand the meaning of a text.

## Types of ML and Techniques in every one

### Supervised Learning

Supervised learning solves problems that involve using a model to map between input examples and the desired output. Models are fed with training data that contains inputs and outputs and used to make predictions on test sets where only the inputs are provided and the outputs from the model are compared to the target variables and used to estimate the skill of the model.

There are two techniques in supervised learning problems:

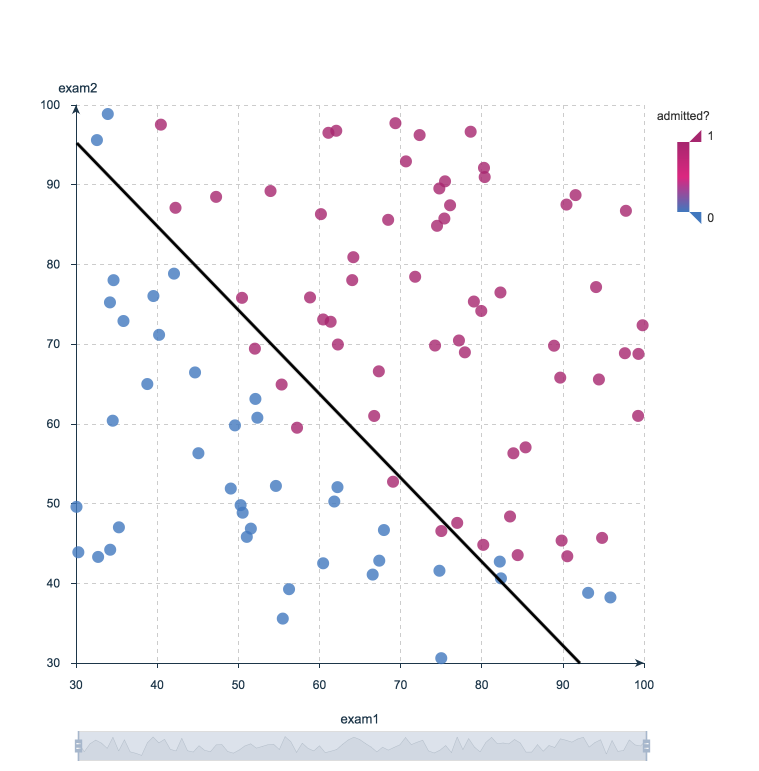
#### Classification

Classification problems predict class value. For example we can predict if the customer will buy a product or not. The output is a discrete value either yes or no. classification problem not limited for only 2 classes we can extend the model to as many classes as we want in our problem. For example we can train a model with a dataset consisting of images that contains cat or dog or neither, and the model task is to assign the output to one of this discrete value either the image contains a cat , 2) the image contains a dog, or 3) the image contains neither a cat nor a dog.

The simplest classification algorithm is logistic regression — which makes it sounds like a regression method, but it’s not. Logistic regression estimates the probability of an occurrence of an event based on one or more inputs.

For instance, a logistic regression can take as inputs two exam scores for a student in order to estimate the probability that the student will get admitted to a particular college. Because the estimate is a probability, the output is a number between 0 and 1, where 1 represents complete certainty. For the student, if the estimated probability is greater than 0.5, then we predict that he or she will be admitted. If the estimated probability is less than 0.5, we predict he or she will be refused.

The chart below plots the scores of previous students along with whether they were admitted. Logistic regression allows us to draw a line that represents the decision boundary.



Logistic Regression Decision Boundary: Admitted to College or Not?

#### Regression

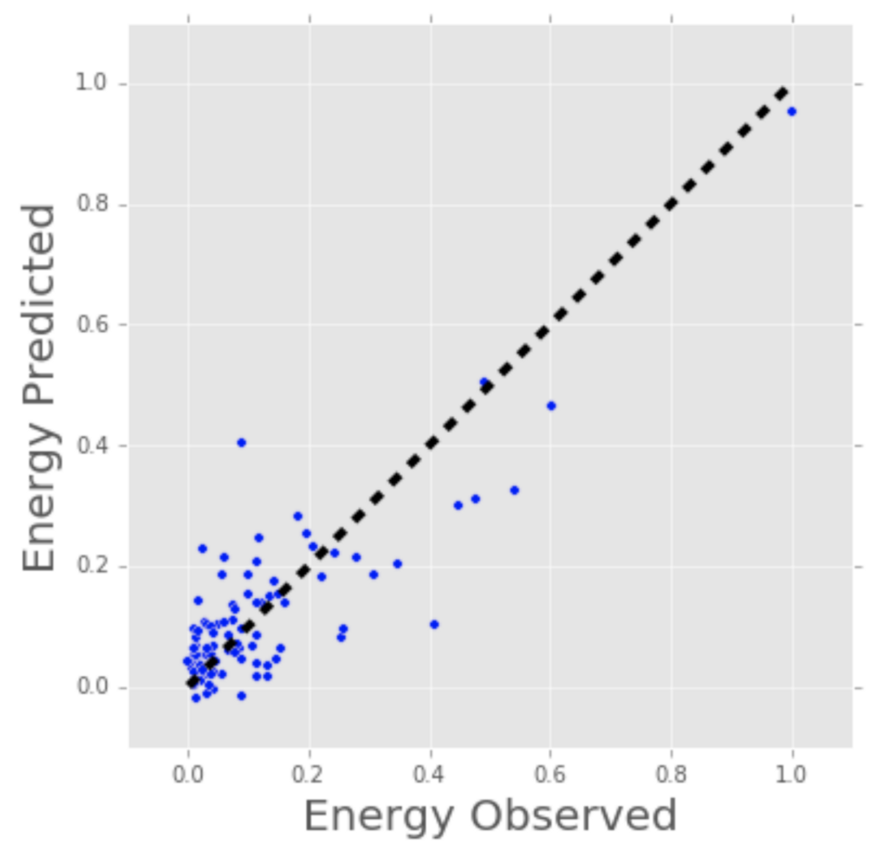
Regression methods belong to the category of supervised Machine Learning. They help to predict a particular continuous numerical value based on a dataset, for example predicting the price of a house based on previous pricing data for similar houses.

The simplest method is linear regression where we use the line equation (y = m \* x + b) to model a dataset. We train a linear regression model with data pairs (x, y) by calculating the position and slope of a line that minimizes the total distance between all of the dataset points and the predicted line.

A more concrete example of linear regression. A model predicts the energy consumption (in kWh) of certain buildings by collecting together the age of the building, number of stories, square feet and the number of plugged wall equipment. In this example there is more than one input (age, square feet, etc…). The principle was the same as a simple one-to-one linear regression, but in this example the “line” created occurred in multi-dimensional space based on the number of variables.

The plot below shows how well the linear regression model fits the actual energy consumption of the building. Now imagine that you have access to the characteristics of a building (age, square feet, etc…) but you don’t know the energy consumption. In this case, we can use the fitted line to approximate the energy consumption of the particular building.

Note that you can also use linear regression to estimate the weight of each factor that contributes to the final prediction of consumed energy. For example, once you have a formula, you can determine whether age, size, or height is most important.



Linear Regression Model Estimates of Building’s Energy Consumption (kWh).

### 

### 

### Unsupervised Learning

Unsupervised learning describes problems that use a model to describe or extract relationships in data.

Compared to supervised learning, unsupervised learning operates upon only the input data without output knowledge. As such, unsupervised learning does not have a teacher correcting the model, as in the case of supervised learning.

There are many techniques in supervised learning problems, such:

#### Clustering

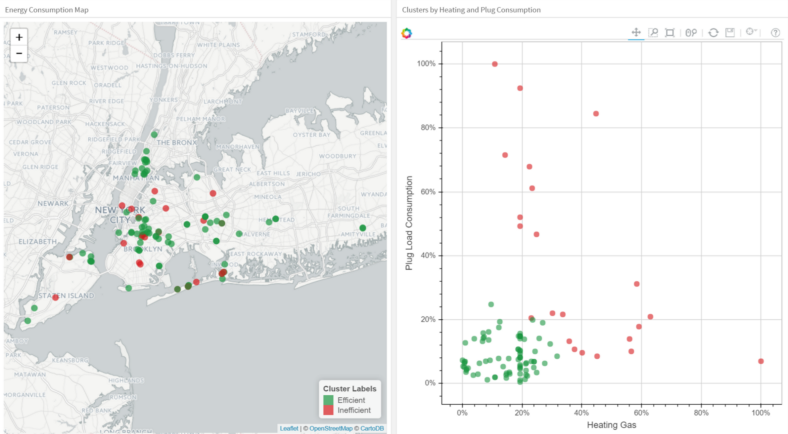
In clustering methods, we are talking about the category of unsupervised ML because their goal is to group or cluster observations that have similar characteristics. Clustering methods don’t use output information for training the model, but instead they let the algorithm define the output. In clustering methods, we can only use visualizations to inspect the quality of the solution.

The most popular clustering method is K-Means, where “K” represents the number of clusters the user chooses to create to train the model on it.

what K-Means does with the data points:

1. Randomly chooses K centers within the data.
2. Assigns each data point to the closest of the randomly created centers.
3. Re-computes the center of each cluster.
4. If centers don’t change (or change very little), the process is finished. Otherwise, we return to step 2. (To prevent ending up in an infinite loop if the centers continue to change, set a maximum number of iterations in advance.)

The next plot applies K-Means to a data set of buildings. Each column in the plot indicates the efficiency for each building. The four measurements are related to air conditioning, plugged-in equipment (microwaves, refrigerators, etc…), domestic gas, and heating gas. We chose K=2 for clustering, which makes it easy to interpret one of the clusters as the group of efficient buildings and the other cluster as the group of inefficient buildings. To the left you see the location of the buildings and to right you see two of the four dimensions we used as inputs: plugged-in equipment and heating gas.



Clustering Buildings into Efficient (Green) and Inefficient (Red) Groups.

### Reinforcement Learning

Reinforcement learning describes a type of problems where the model operates in an environment and must learn from the errors that occur.

Reinforcement learning is learning what to do — how to map situations to actions—so as to maximize a numerical reward signal. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them.

The use of an environment means that there is no fixed training dataset, rather a goal or set of goals that an agent is required to achieve, actions they may perform, and feedback about performance toward the goal.

Some machine learning algorithms do not just experience a fixed dataset. For example, reinforcement learning algorithms interact with an environment, so there is a feedback loop between the learning system and its experiences.

## Usages of each type

### Supervised Learning

The most practical machine learning uses this type (supervised learning). Some common sorts of problems built on top of classification and regression include recommendation and statistical prediction respectively.

Some popular samples of supervised machine learning algorithms are:

1. Linear regression for regression problems.
2. Random forest for classification and regression problems.
3. Support vector machines for classification problems.

Other examples.

1. Visual Recognition:

An AI that's learning to spot pedestrians on a street is trained with 2 million short videos of street scenes from self-driving cars. a number of the videos contain no pedestrians in the least while others have up to 25. a spread of learning algorithms are trained on the info with each having access to the right answers. Each algorithm develops a spread of models to spot pedestrians in fast paced scenes . The algorithms are then tested against another set of knowledge to gauge accuracy and precision.

1. Sorting:

A robot is learning to sort garbage using visual identification. It sits all day picking out recyclable items from garbage because it passes on the conveyor belt . it places items like glass, plastic and metal into 12 bins. Each item is labeled with a number on a sticker. Once each day , human experts examine the bins and inform the robot which items were improperly sorted. The robot uses this feedback to enhance .

1. Decision Support:

An AI is learning to estimate investing risk. it's fed an outsized number of trades that real investors made and asked to estimate a risk/reward ratio for every trade supported company fundamentals, price and other factors like volume. The estimated risk/reward ratio is then compared to the historical results of the trade at a spread of your time intervals like each day or year.

### Unsupervised Learning

The goal for unsupervised learning is to model the underlying structure or distribution within the data so as to find out more about the data.

Some popular samples of unsupervised learning algorithms are:

* k-means for clustering problems.
* Apriori algorithm for association rule learning problems.

Other examples.

1. Human Behavior:

A learner that processes highly developed visual and speech recognition capabilities could watch an outsized number of television shows to find out about human behavior. For instance , a learner could be ready to build a model that detects when people are smiling and supports the correlation of facial patterns and words like “what are you smiling about?”

1. Robotics:

A highly developed AI that is a housekeeping robot develops a theory that there's usually dust under a settee . Each week, the idea is confirmed because the robot often finds dust under sofas. Nobody explicitly tells the robot the idea is correct but it's ready to develop confidence in it nonetheless.

### Reinforcement Learning

There are many industrial applications of reinforcement learning and it includes an incredibly wide chain of problems with a big impact, including:

* Personalized dynamic recommender systems
* Personalized multi-channel marketing
* Automated ad bidding and buying
* Personalized medication dosing
* Dynamic resource allocation in wind farms, HVAC systems, and computing clusters
* Automated calibration of engines and other machines
* Robotic control
* Supply chain optimization

# 

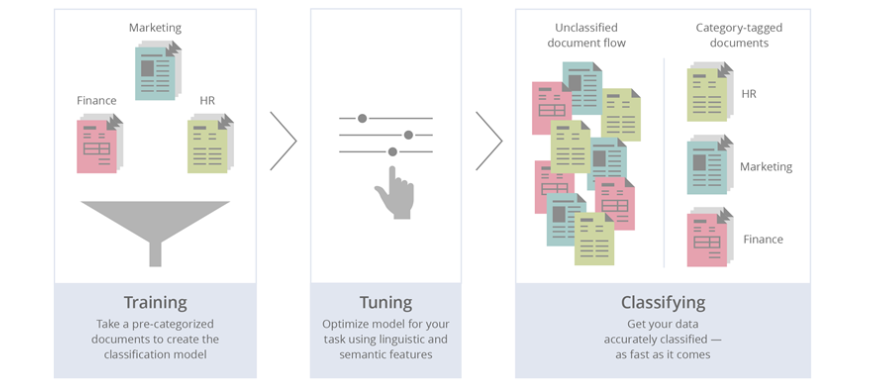
## Classification

### Classification In Machine Learning

Classification is a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.

### Text analysis can be performed at:

1. Document-level:you will obtain relevant information for a full document. Paragraph level: obtains the most important categories of just one paragraph.
2. Sentence level: obtains relevant information of a single sentence.
3. Sub-sentence level: obtains relevant information of sub-expressions within sentences (also known as opinion units). This is particularly useful when there are ambiguous sentences that mention multiple topics.



How document classification works ?

### How Does Automatic Document Classification Work?

This is a process fueled by Natural Language Processing (NLP), by which algorithms automatically assign one or more categories to your text-based documents such as articles, emails, or survey responses. Using machine learning models is faster, more scalable, and less biased than manual classification because machines never get tired, bored, or change their criteria over time.

### Different approaches to document classification you can adopt:

1. Supervised: In this method, machine learning models need you to manually tag a number of texts before they can start making predictions on their own. So, this means that first you will have to define a set of tags (let’s say, Customer Service, Usability, Pricing) that you will later use to classify your documents by hand before the model can do it on its own. From these examples, the model will learn to make associations between the texts and the expected tags. For example, a customer review that says “the software is quite expensive” needs to be tagged as Pricing. The number of texts you classify will also influence the confidence of the model.
2. Unsupervised: With this method, documents containing similar words or sentences will be grouped together by a classifier without any prior training. For example, the words RAM, SSD, or Printer in customer reviews would be recognized as sharing similar qualities and grouped within the same cluster.
3. Rules-based: As its name indicates, this method is based on linguistic rules that give instructions to the model, which will automatically tag your texts following these patterns. These rules are based on morphology, lexis, syntax, semantics, and phonology.

### Classification Types:

1. Binary classification means there are two classes to work with that relate to one another as true and false.
2. Multiclass classification helps us to sort all inputs to many classes.
3. Multi-label classification is applied when one input can belong to more than one class.
4. Imbalanced classification is used for fraud detection software and medical diagnosis.

### Types of Classification Algorithms:

1. **Logistic regression:**In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function.

**Advantages:** Logistic regression is designed for this purpose (classification), and is most useful for understanding the influence of several independent variables on a single outcome variable.

**Disadvantages:** Works only when the predicted variable is binary, assumes all predictors are independent of each other and assumes data is free of missing values.

1. **Naive Bayes algorithm** based on Bayes’ theorem with the assumption of independence between every pair of features. Naive Bayes classifiers work well in many real-world situations such as document classification and spam filtering.

**Advantages:** This algorithm requires a small amount of training data to estimate the necessary parameters. Naive Bayes classifiers are extremely fast compared to more sophisticated methods.

**Disadvantages:** Naive Bayes is known to be a bad estimator.

1. **Stochastic gradient descent** is a simple and very efficient approach to fit linear models. It is particularly useful when the number of samples is very large. It supports different loss functions and penalties for classification.

**Advantages:** Efficiency and ease of implementation.

**Disadvantages:** Requires a number of hyper-parameters and it is sensitive to feature scaling.

1. **K-Nearest Neighbours** based classification is a type of lazy learning as it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the k nearest neighbours of each point.

**Advantages:** This algorithm is simple to implement, robust to noisy training data, and effective if training data is large.

**Disadvantages:** Need to determine the value of K and the computation cost is high as it needs to compute the distance of each instance to all the training samples.

1. **Decision Tree:**Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data.

**Advantages:** Decision Tree is simple to understand and visualise, requires little data preparation, and can handle both numerical and categorical data.

**Disadvantages:** Decision tree can create complex trees that do not generalise well, and decision trees can be unstable because small variations in the data might result in a completely different tree being generated.

1. **Random forest classifier** is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.

**Advantages:** Reduction in over-fitting and random forest classifiers is more accurate than decision trees in most cases.

**Disadvantages:** Slow real time prediction, difficult to implement, and complex algorithm.

1. **Support vector machine** is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

**Advantages:** Effective in high dimensional spaces and uses a subset of training points in the decision function so it is also memory efficient.

**Disadvantages:** The algorithm does not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

# Our project(Multi label Document Classification)

**Description:** A web application large Scale document classifier in scientific fields such as math, statistics, electrical engineering, quantitative biology, and economics.

**Project type:** MultiClass and MultiLabel.

**Algorithms:** Naive-Bayes, SVM, Linear SVM, Random forest, Nearest Neighbors, Decision Tree and Neural Network.

**Tools:** python , jupyter, pandas,tensorflow, ReactJs .

Difficulties and Solutions:

## Choosing what kind of classifier to use:

The first question to ask is how much training data is there currently available? None? Very little? Quite a lot? Or a huge amount, growing every day? Often one of the biggest practical challenges in fielding a machine learning classifier in real applications is creating or obtaining enough training data. For many problems and algorithms, hundreds or thousands of examples from each class are required to produce a high performance classifier and many real world contexts involve large sets of categories. We will initially assume that the classifier is needed as soon as possible; if a lot of time is available for implementation, much of it might be spent on assembling data resources.

Solutions:

* If you have fairly little data and you are going to train a supervised classifier, then machine learning theory says you should stick to a classifier with high bias.
* If there is a reasonable amount of labeled data, then you are in the perfect position to use everything that we have presented about text classification.
* If a huge amount of data is available, then the choice of classifier probably has little effect on your results and the best choice may be unclear.  
  It may be best to choose a classifier based on the scalability of training or even runtime efficiency.

### 

## Large and difficult category taxonomies

If a text classification problem consists of a small number of well-separated categories, then many classification algorithms are likely to work well. But many real classification problems consist of a very large number of often very similar categories.

Solutions:

* A general result in machine learning is that you can always get a small boost in classification accuracy by combining multiple classifiers, provided only that the mistakes that they make are at least somewhat independent.

## Features for text

Classification problems will often contain large numbers of terms which can be conveniently grouped, and which have a similar vote in text classification problems.  
For instance, this would be the case if the keyword *ethnic* was most indicative of the categories *food* and *arts*, the keyword *cleansing* was most indicative of the category *home*, but the collocation *ethnic cleansing* instead indicates the category *world news*.

Solutions:

* Do techniques like stemming and lowercasing (vocabulary) help for text classification? As always, the ultimate test is empirical evaluations conducted on an appropriate test collection. But it is nevertheless useful to note that such techniques have a more restricted chance of being useful for classification.

## Document zones in text classification

documents usually have zones, such as mail message headers like the subject and author, or the title and keywords of a research article. Text classifiers can usually gain from making use of these zones during training and classification.

4.1 . **Upweighting document zones.**

In text classification problems, you can frequently get a nice boost to effectiveness by differentially weighting contributions from different document zones. Often, upweighting title words is particularly effective.

you can get value from upweighting the first sentence of a document.

4.2.**Separate feature spaces for document zones.**

An alternative strategy is to have a completely separate set of features and corresponding parameters for words occurring in different zones.

It is quite uncommon for words to have different preferences when appearing in different zones; it is mainly the strength of their vote that should be adjusted. Nevertheless, ultimately this is a contingent result, depending on the nature and quantity of the training data.

4.3.**Connections to text summarization.**

Adopted the limited goal of extracting and assembling pieces of the original text that are judged to be central based on features of sentences that consider the sentence's position and content.

Based on text summarization research, they consider using :

(i) only the title.

(ii) only the first paragraph.

(iii) only the paragraph with the most title words or keywords.

(iv) the first two paragraphs or the first and last paragraph.

(v) all sentences with a minimum number of title words or keywords.

In general, these positional feature selection methods produced as good results as mutual information

# References

1. Jason Brownlee, August 20, 2020   
   <https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/>
2. John Spacy, May 03, 2017

<https://simplicable.com/new/unsupervised-learning>

1. Page 1, Reinforcement Learning: An Introduction, 2nd edition, 2018.
2. Page 105, Deep Learning, 2016.
3. <https://analyticsindiamag.com/7-types-classification-algorithms/>
4. <https://monkeylearn.com/blog/classification-algorithms/>
5. <https://nlp.stanford.edu/IR-book/html/htmledition/issues-in-the-classification-of-text-documents-1.html>
6. <https://nlp.stanford.edu/IR-book/html/htmledition/issues-in-the-classification-of-text-documents-1.html>