Ecole Supérieure de Technologie

Essaouira

Sentiment Analysis Report

IDSD

By:

Salah Eddine ZKARA

Supervisor:

Pr. Azidine GUEZZAZ

Academic year:

2019/2020

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to Professor Azidine Guezzaz my research supervisor, for his patient guidance, enthusiastic encouragement, and useful critiques of this research work.

My grateful thanks are also extended to all my Professors during this year for their serious work and encouragement.

Finally, I wish to thank my parents for their support and encouragement throughout my study.

Table of Contents

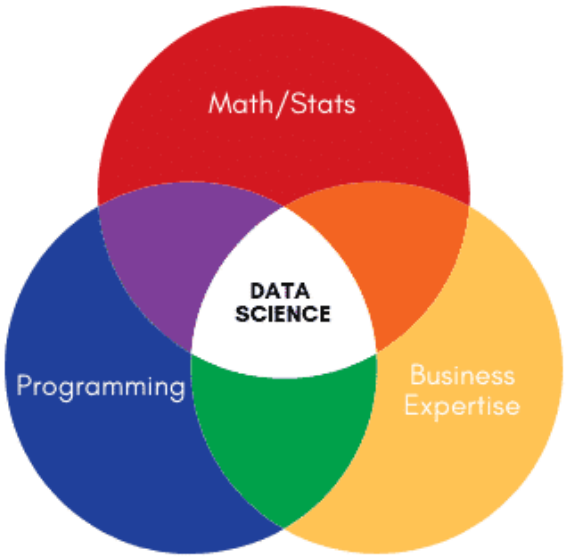
Introduction

Sentiment analysis is the automated process of determining whether a text expresses a positive or negative opinion about a product or topic. By using sentiment analysis, companies do not have to spend endless hours tagging customer data such as survey responses, reviews, support tickets, and social media comments.

Sentiment analysis helps companies monitor their brand reputation on social media, gain insights from customer feedback, and much more!

In this report, we will go into more details about how to perform Sentiment Analysis with Machine Learning *(naive bayes classifier)*.

I- What is Data Science?



First, data science is a scientific method of providing actionable intelligence from data using mathematics, statistics, programming, and business expertise. Like any scientific method, it involves gathering data, identifying a problem, forming a hypothesis, and running tests. More specifically, data scientists follow a process of gathering and cleaning data (wrangling), investigation (exploratory data analysis), building automation using machine learning (feature engineering, model development, and deployment), delivering results (visualizations, reporting, storytelling), and maintenance. Practitioners typically spend 70-80% of their time in the wrangling/exploration, 20% on machine learning models, and the rest in maintenance. Most importantly, this whole process should result in a valuable action or insight for the end-user, i.e. a business or customer!

1- Statistics Descriptive

- **Descriptive statistics** are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.
- **Descriptive Statistics** are used to present quantitative descriptions in a manageable form. In a research study we may have lots of measures. Or we may measure a large number of people on any measure. Descriptive statistics help us to simplify large amounts of data in a sensible way.

2- Statistics Inferential

- **Statistics Inferential** we are trying to reach conclusions that extend beyond the immediate data alone. For instance, we use inferential statistics to try to infer from the sample data what the population might think. Or we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study. Thus, we use inferential statistics to make inferences from our data to more general conditions

3- Data Mining

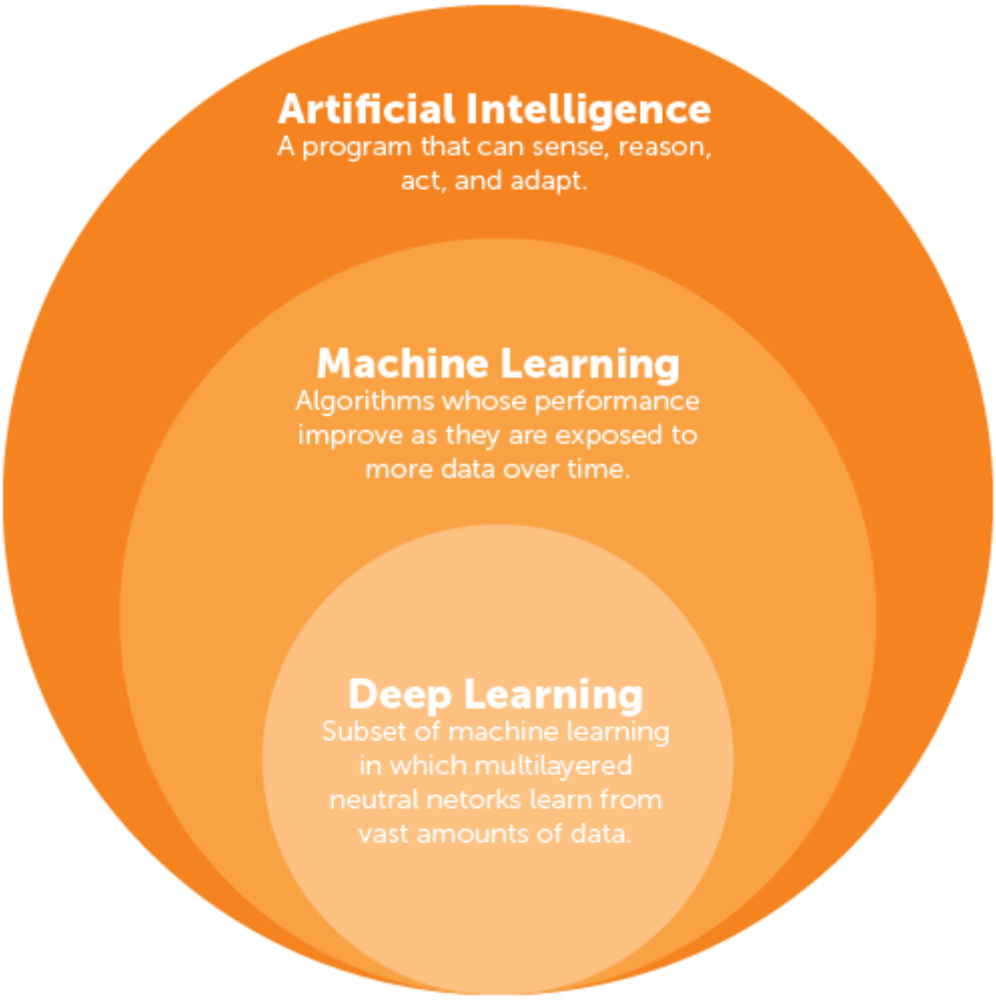
Data Mining is **the process of finding anomalies, patterns, and correlations within large data sets to predict outcomes**. Using a broad range of techniques, one can use this information to increase revenues, cut costs, improve customer relationships, reduce risks and more. **Thus, Data Mining is about finding the trends in a data set. Moreover, using these trends to identify patterns**.

Data Mining Vs Data Science

- Data Mining is an activity which is a part of a broader Knowledge Discovery in Databases (KDD) Process while Data Science is a field of study just like Applied Mathematics or Computer Science.
- Often Data Science is looked upon in a broad sense while Data Mining is considered a niche.
- Some activities under Data Mining such as statistical analysis, writing data flows and pattern recognition can intersect with Data Science. Hence, Data Mining becomes a subset of Data Science.
- Machine Learning in Data Mining is used more in pattern recognition while in Data Science it has a more general use.

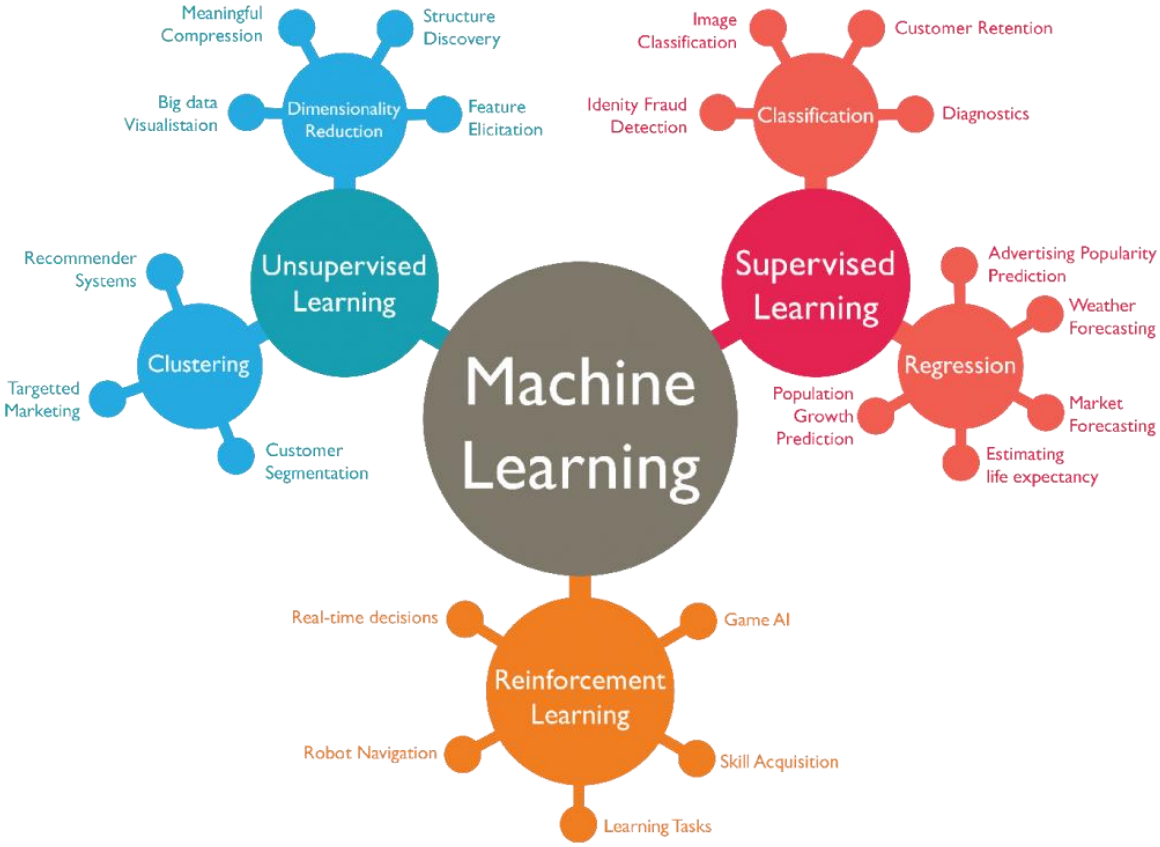
Basis for comparison	Data Mining	Data Science
What is it?	A technique	An area
Focus	Business process	Scientific study
Goal	Make data more usable	Building Data-centric products for an organization
Output	Patterns	Varied
Purpose	Finding trends previously not known	Social analysis, building predictive models, unearthing unknown facts, and more
Vocational Perspective	Someone with a knowledge of navigating across data and statistical understanding can conduct data mining	A person needs to understand Machine Learning, Programming, info-graphic techniques and have the domain knowledge to become a data scientist
Extent	Data mining can be a subset of Data Science as Mining activities are part of the Data Science pipeline	Multidisciplinary – Data Science consists of Data Visualizations, Computational Social Sciences, Statistics, Data Mining, Natural Language Processing, et cetera
Deals with (the type of data)	Mostly structured	All forms of data – structured, semi-structured and unstructured
Other less popular names	Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction	Data-driven Science

4- Machine Learning (overview)



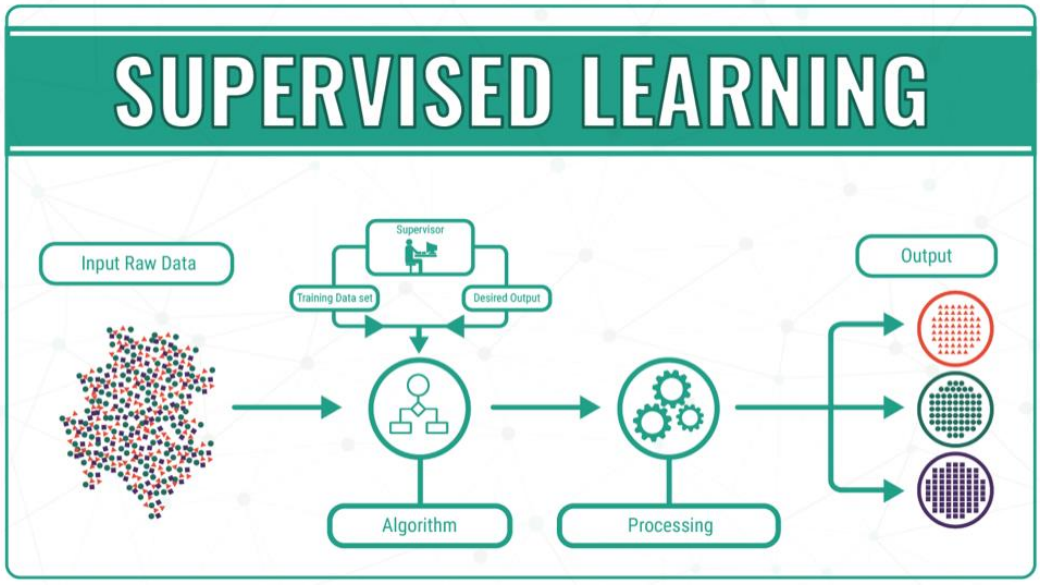
- **Machine learning (ML)** is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks.
- **Machine learning** is closely related to computational statistics, which focuses on making predictions using computers.
- The discipline of machine learning employs various approaches to teach computers to accomplish tasks where no fully satisfactory algorithm is available. In cases where vast numbers of potential answers exist, one approach is to label some of the correct answers as valid. This can then be used as training data for the computer to improve the algorithm(s) it uses to determine correct answers.

II- Machine Learning



1- Supervised

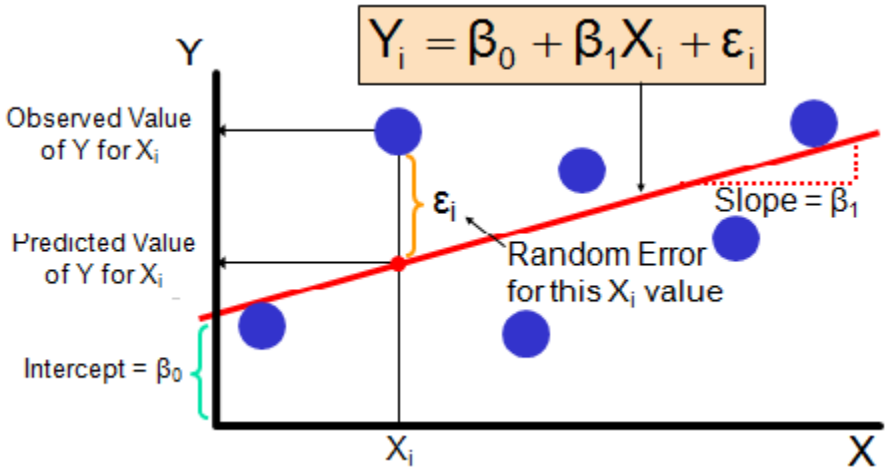
Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples. In supervised learning, each example is a pair consisting of an input object *(typically a vector)* and a desired output value *(also called the supervisory signal)*.



a- Regression

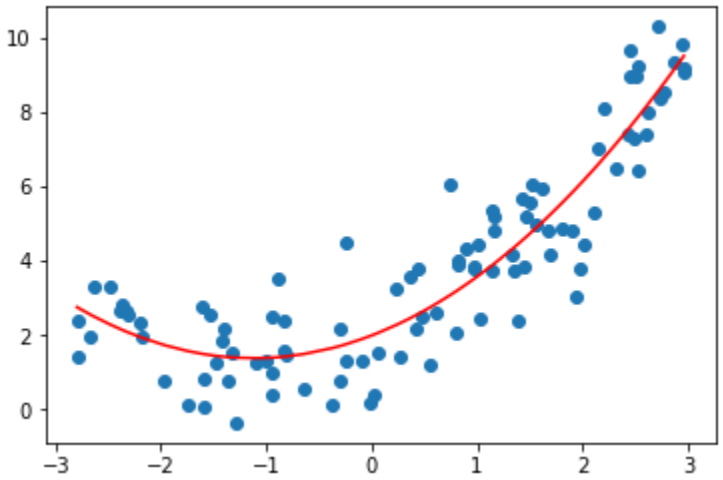
Linear regression:

linear regression is a **linear approach to modeling the relationship between a scalar response** *(or dependent variable)* **and one or more explanatory variables** *(or independent variables)*. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.



Polynomial regression:

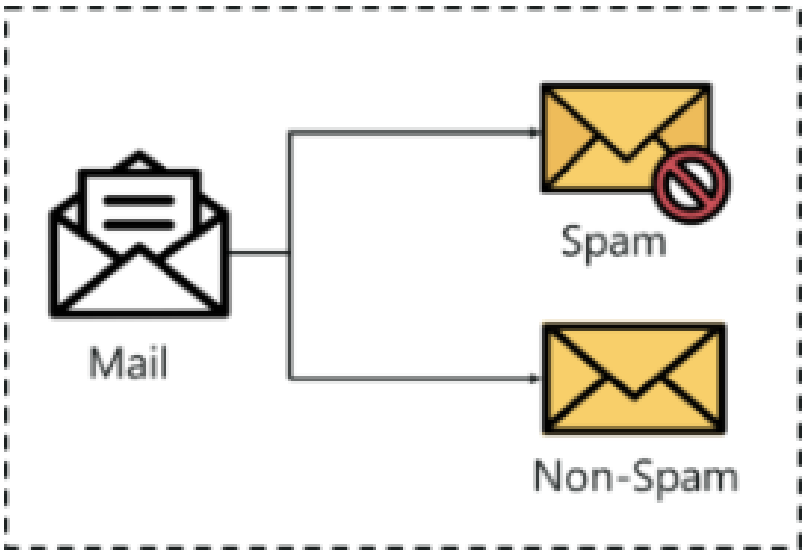
Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an **nth degree polynomial in x**. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y |x). Although polynomial regression fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function E(y | x) is linear in the unknown parameters that are estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression.



b- Classification

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.

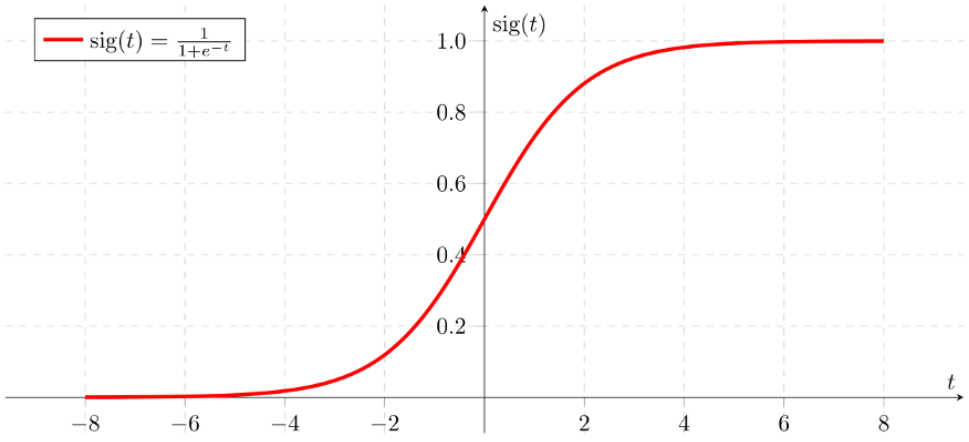
The classification predictive modeling is the task of approximating the mapping function from input variables to discrete output variables. The main goal is to identify which class/category the new data will fall into.



Logistic regression:

Logistic regression is a **statistical model that in its basic form uses a logistic function to model a binary dependent variable**, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probity model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given

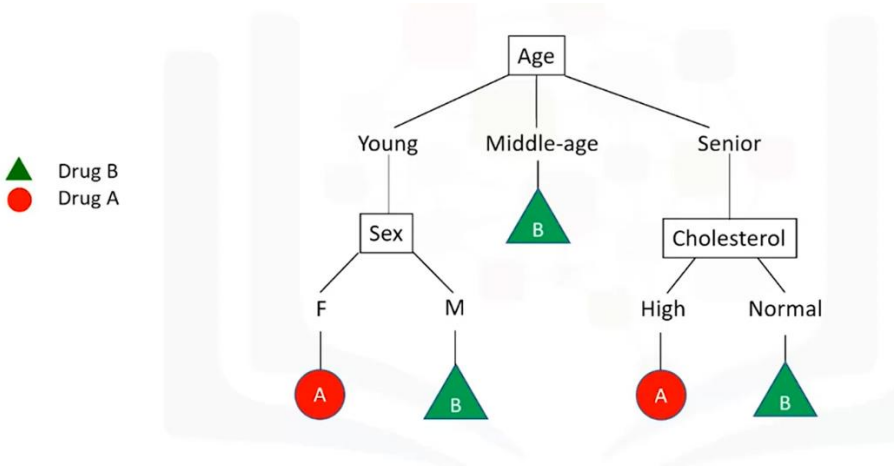
outcome at a constant rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.



Decision Tree:

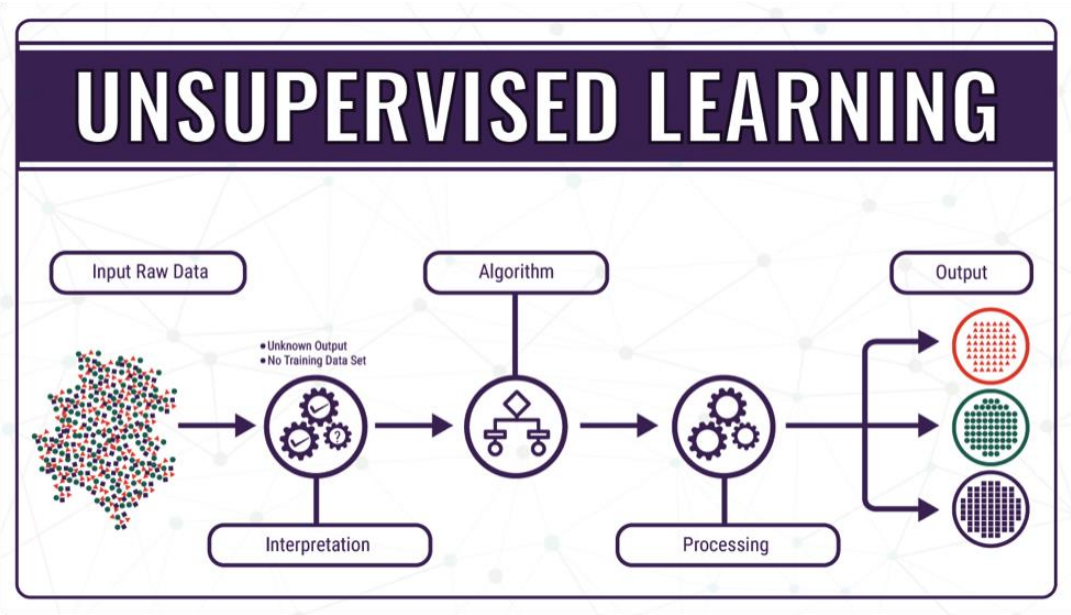
A **decision tree** is a **flowchart-like** structure in which **each internal node represents a "test" on an attribute, each branch represents the outcome of the test,** and **each leaf node represents a class label** *(decision taken after computing all attributes Entropy and Information Gain)*. The paths from root to leaf represent classification rules.

In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.



2- Unsupervised

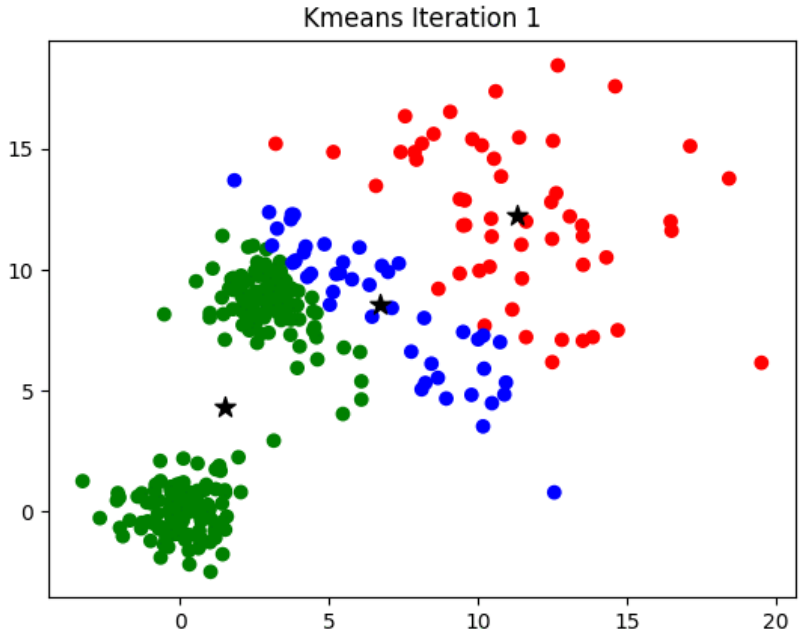
Unsupervised learning is a type of machine learning that looks for previously **undetected patterns in a data set with no pre-existing**. so, when a dataset is provided without labels the model learns useful properties of the structure of the dataset and come with a patterns or conclusions from the unlabeled data.



a- Clustering

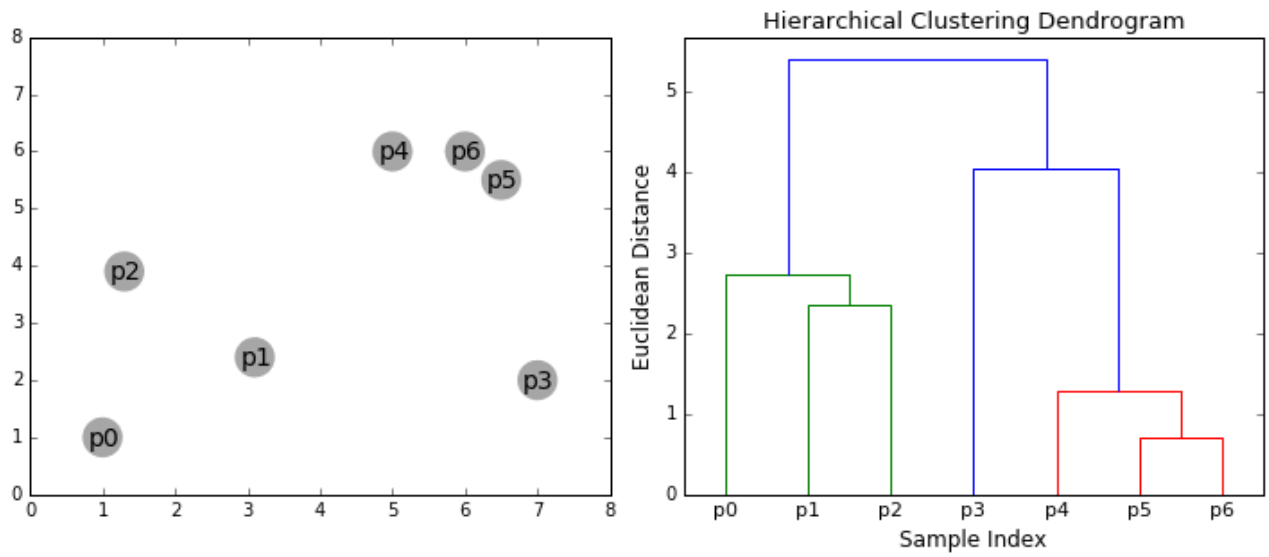
k-means:

k-means clustering is a method of **vector quantization**, originally from signal processing, that aims to **partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean** *(cluster centers or cluster centroid)*, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. It is popular for cluster analysis in data mining. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.



hierarchical:

Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly like each other.



3- Natural Language Processing (NLP)

Natural Language Processing *(NLP)* is a **subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language**, in particular how to program computers to process and analyze large amounts of natural language data.

Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation.

III- Sentiment Analysis

1- Principal Data Cleaning Process (NLTK)

The **Natural Language ToolKit**, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical Natural Language Processing *(NLP)* for English written in the Python programming language. NLTK includes graphical demonstrations and sample data. It is accompanied by a book that explains the underlying concepts behind the language processing tasks supported by the toolkit.

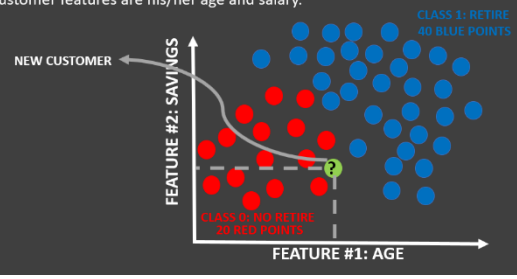
In this project, we are going to use **re** *(regular expression)* and NLTK libraries to perform data cleaning.

2- Algorithms Implemented

Naive Bayes:

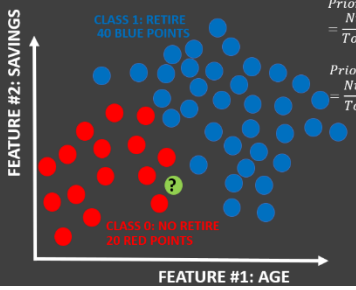
NAÏVE BAYES: INTUITION

- Naïve Bayes is a classification technique based on Bayes' Theorem.
- Let's assume that you are data scientist working major bank in NYC and you want to classify a new client as eligible to retire or not.
- Customer features are his/her age and salary.



NAÏVE BAYES: 1. PRIOR PROBABILITY

- Points can be classified as RED or BLUE and our task is to classify a new point to RED or BLUE.
- Prior Probability: Since we have more BLUE compared to RED, we can assume that our new point is twice as likely to be BLUE than RED.

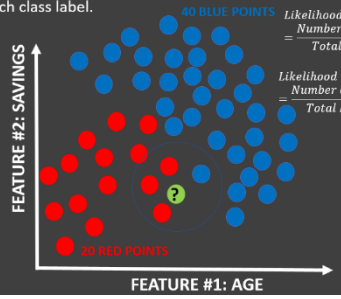


$$\text{Prior Probability for RED} = \frac{\text{Number of RED Points}}{\text{Total Number of Points}} = \frac{20}{60}$$

$$\text{Prior Probability for BLUE} = \frac{\text{Number of BLUE Points}}{\text{Total Number of Points}} = \frac{40}{60}$$

NAÏVE BAYES: 2. LIKELIHOOD

- For the new point, if there are more BLUE points in its vicinity, it is more likely that the new point will be classified as BLUE.
- So we draw a circle around the point, then we calculate the number of points in the circle belonging to each class label.

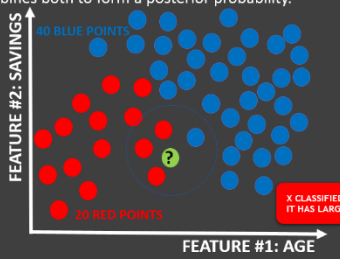


$$\text{Likelihood of X being RED} = \frac{\text{Number of RED Points in vicinity}}{\text{Total Number of RED Points}} = \frac{3}{20}$$

$$\text{Likelihood of X being BLUE} = \frac{\text{Number of BLUE Points in vicinity}}{\text{Total Number of BLUE Points}} = \frac{1}{40}$$

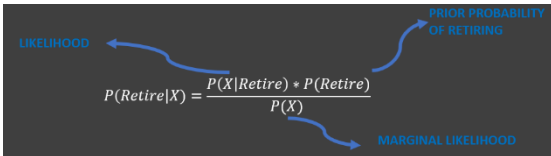
NAÏVE BAYES: 3. POSTERIOR PROBABILITY

- Let's combine prior probability and likelihood to create a posterior probability.
- Prior probabilities: suggests that X may be classified as BLUE Because there are 2x as much blue points.
- Likelihood: suggests that X is RED because there are more RED points in the vicinity of X.
- Bayes' Rule combines both to form a posterior probability.



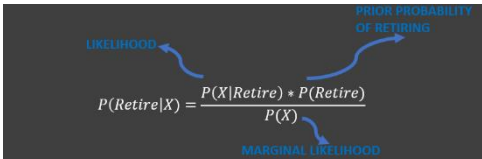
$$\text{Posterior Probability of X being RED} = \text{Prior Probability of RED} * \text{Likelihood of X being RED} = \frac{20}{60} * \frac{3}{20} = \frac{1}{20}$$

$$\text{Posterior Probability of X being BLUE} = \text{Prior Probability of BLUE} * \text{Likelihood of X being BLUE} = \frac{40}{60} * \frac{1}{40} = \frac{1}{60}$$



$$P(Retire|X) = \frac{P(X|Retire) * P(Retire)}{P(X)}$$

- **Naïve Bayes is a classification technique based on Bayes' Theorem.**
- X: New Customer's features; age and savings
- $P(Retire|X)$: probability of customer retiring given his/her features, such as age and savings
- $P(Retire)$: Prior probability of retiring, without any prior knowledge
- $P(X|Retire)$: likelihood
- $P(X)$: Marginal likelihood, the probability of any point added lies into the circle



$$P(Retire|X) = \frac{P(X|Retire) * P(Retire)}{P(X)}$$

- $P(Retire) = \frac{\text{\# of Retiring}}{\text{Total points}} = 40/60$
- $P(X|Retire) = \frac{\text{\# of smilar observations for retiring}}{\text{Total \# retiring}} = 1/40$
- $P(X) = \frac{\text{\# of Similar observations}}{\text{Total \# Points}} = 4/60$
- $P(Retire|X) = \frac{\frac{40}{60} * \frac{1}{40}}{\frac{4}{60}} = \frac{1/60}{4/60} = 0.25$

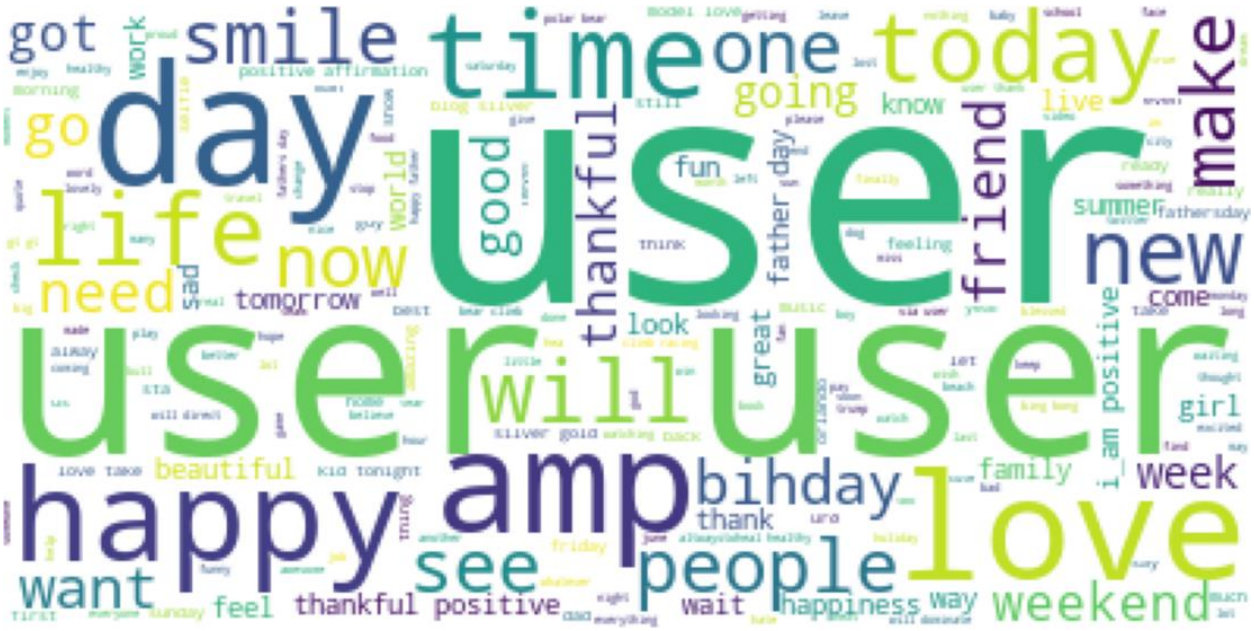
IV- Project Implementation

To implement our project, we are going to use **jupyter notebook** to **Read** Data, **Visualize** Data, **Clean** Data then **train** and **test** our model. *(CSV downloaded from Kaggle)*

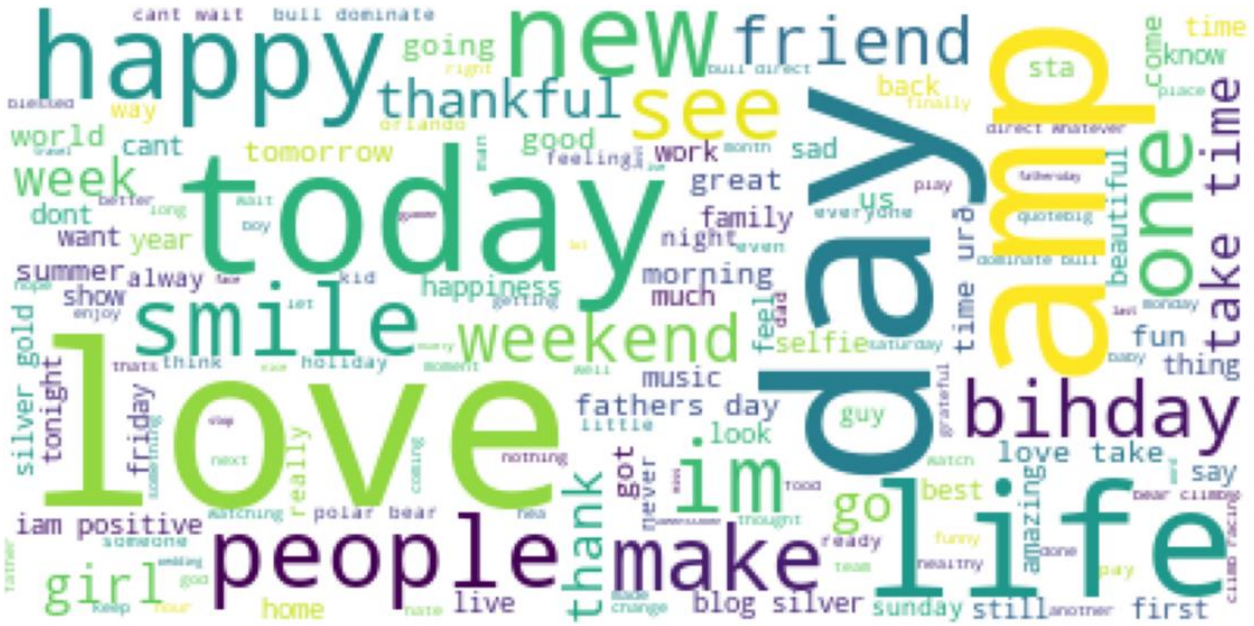
1- Data Cleaning

Data Cleaning or Data Wrangling is an essential Process to remove unnecessary words and punctuation that will add complexity and affect our model negatively.

Positive Word cloud before data cleaning:

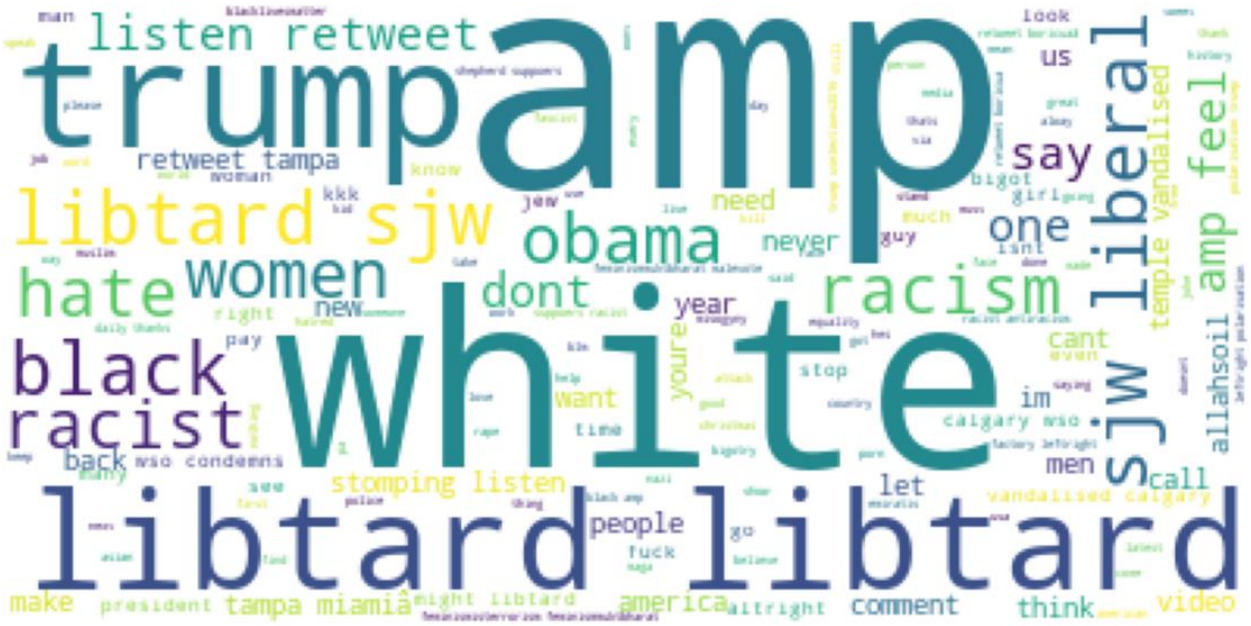


Positive Word cloud after data cleaning:



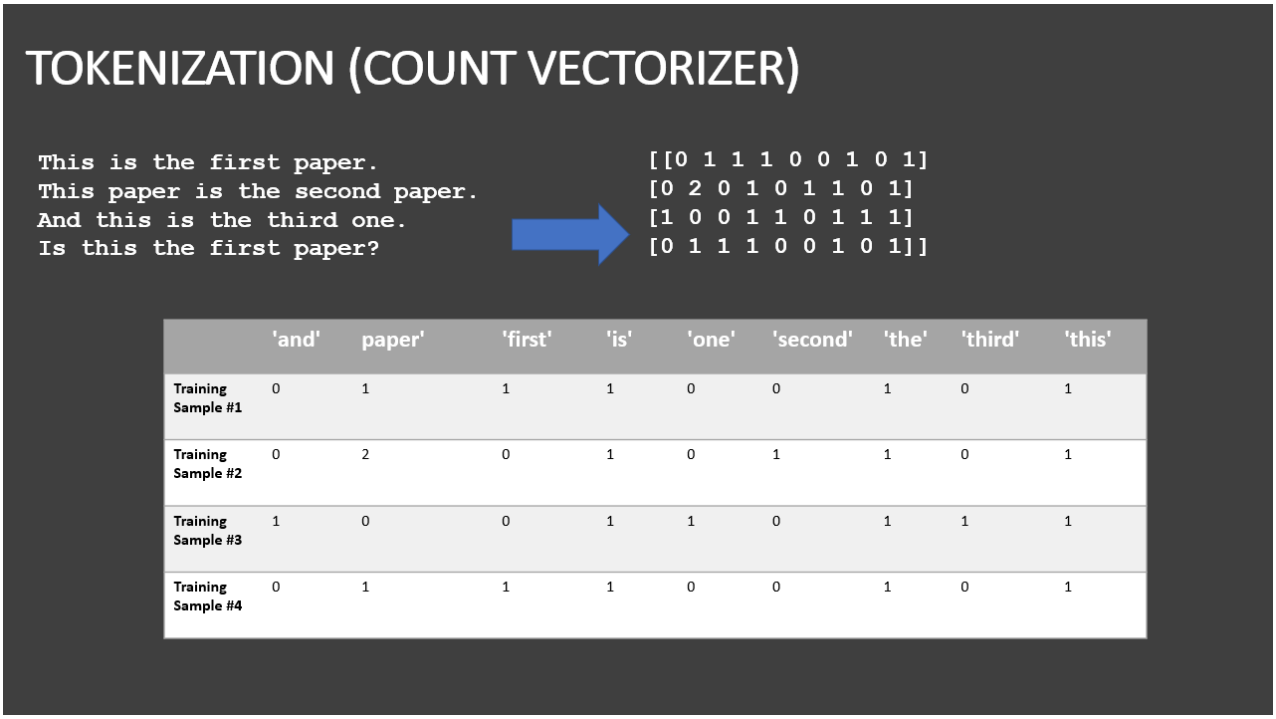
Negative Word Cloud before data cleaning:

Negative Word Cloud after data cleaning:



2- Tokenization

In order to represent words numerically we are going to use Count Vectorization or Tokenization, so to do that we are going to use **CountVectorizer** function from **sklearn**



	'and'	paper'	'first'	'is'	'one'	'second'	'the'	'third'	'this'
Training Sample #1	0	1	1	1	0	0	1	0	1
Training Sample #2	0	2	0	1	0	1	1	0	1
Training Sample #3	1	0	0	1	1	0	1	1	1
Training Sample #4	0	1	1	1	0	0	1	0	1

3- Split Data

To train our model we are going to use 80% from our data, and 20% to test our model. in order to do that we are going to use **train_test_split** function from **sklearn.**

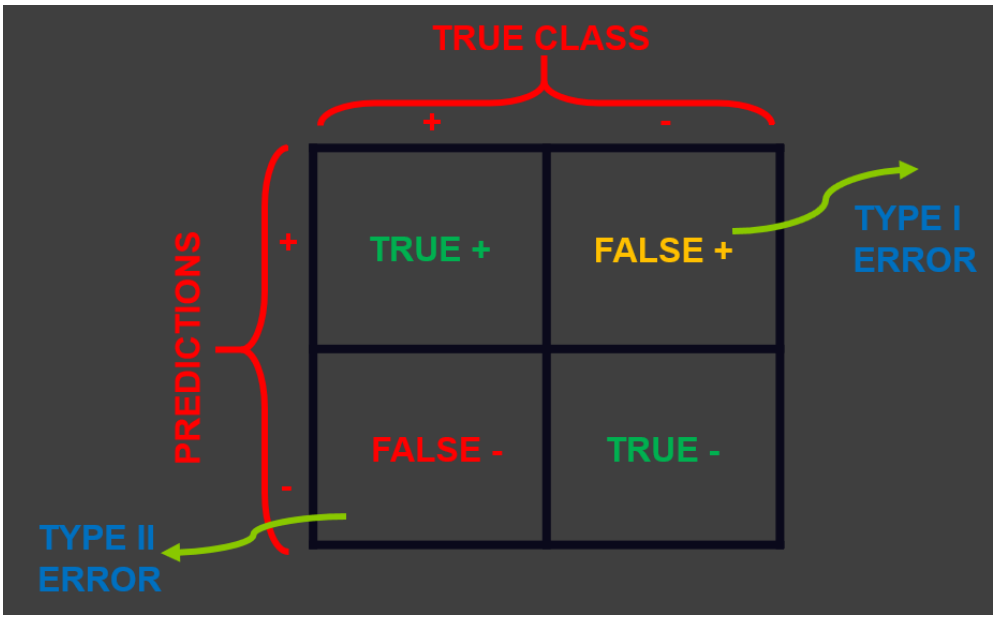
4- Model Training

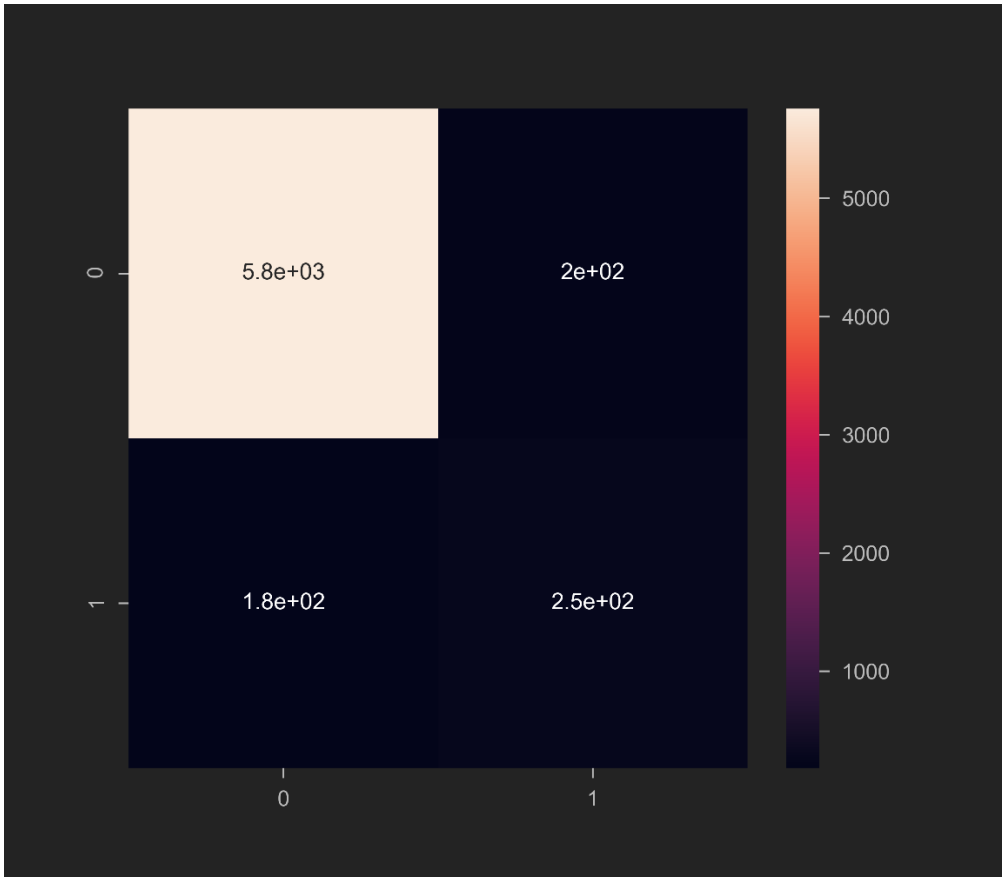
To train our model with NB algorithm we use **MultinomialNB** from **sklearn.**

```
from sklearn.naive_bayes import MultinomialNB

NB_classifier = MultinomialNB()
NB_classifier.fit(X_train, y_train)
```

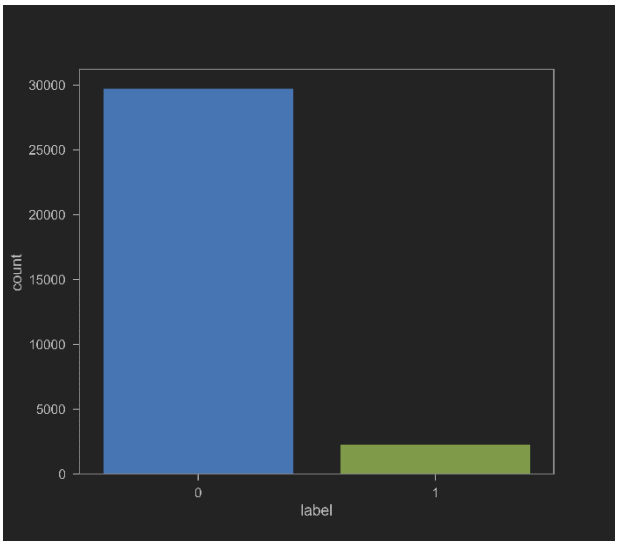
5-Confusion Matrix





6-Classification Report

	precision	recall	f1-score	support
0	0.97	0.97	0.97	5962
1	0.55	0.58	0.56	431
accuracy			0.94	6393
macro avg	0.76	0.77	0.77	6393
weighted avg	0.94	0.94	0.94	6393



(poor accuracy on negative sentences and that because of lack of diversity in the dataset between negative and positive sentences)

7-Heroku Deployment

This is an additional step to get our model to **Production.**

First of all, we are going to **dump** our Model using **pickle** to avoid training another time.

Also, we are going to build a website template using *HTML / CSS / JavaScript*

To get our ML model work we are going to use **Flask**

Deployment Repository

Give it a Try

Conclusion

To sum up, sentiment analysis can improve your business in many ways, from preventing a PR crisis to understanding how your customers feel about your product or service.

Classifying by hand is no longer scalable for businesses. They need fast, accurate, and efficient automated systems that can deliver new insights and empower teams.

Finally, this project gave me the opportunity to apply my skills acquired this year and showed me how to manage my time to get this project done. As a student who is passionate with Data Science and Machine Learning, I give a huge thanks again to my supervisor who taught me to do this kind of projects.