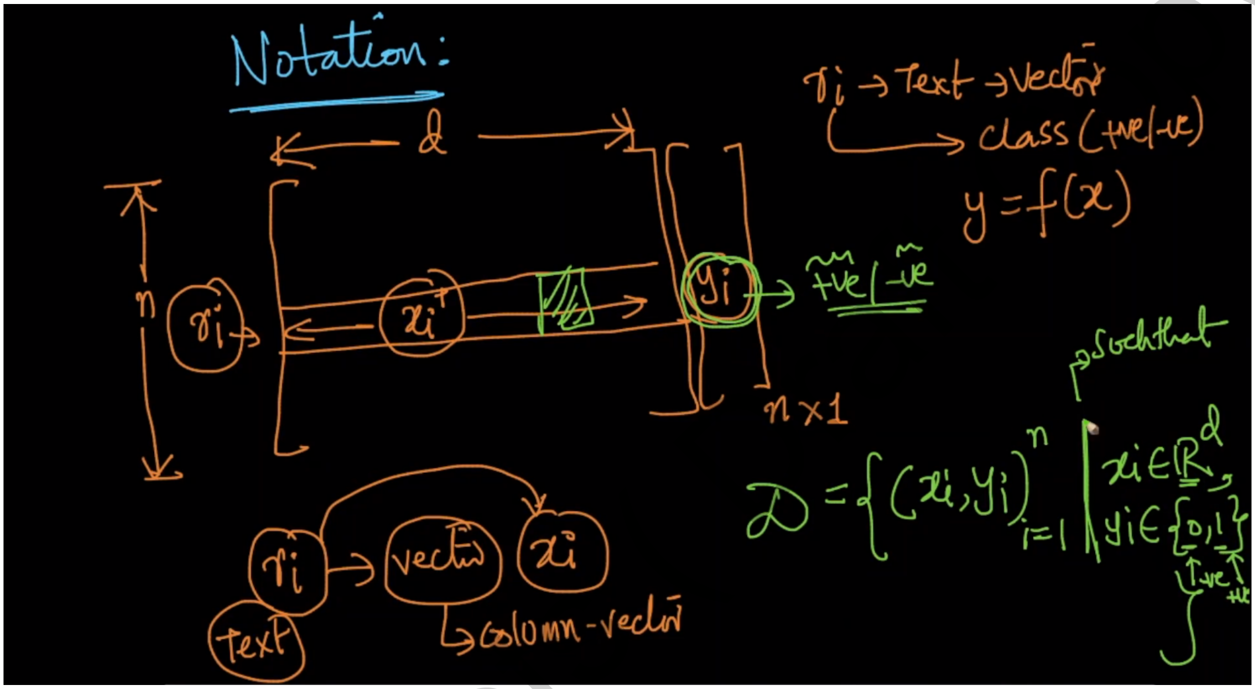
29.2 Data Matrix Notation

We have the reviews in Amazon Fine Food Reviews Dataset and let each review 'r_i' be represented in a vector form 'x_i' and the corresponding output class may be denoted by 'y_i'. Then each data point is represented in data matrix format as shown below.


Notation:
d
n
r_i
x_i^T
y_i
+ve/-ve
$n \times 1$
$r_i \rightarrow$ Text $\rightarrow$ vector
$\rightarrow$ class (+ve/-ve)
$y = f(x)$
such that
$D = \{(x_i, y_i)\}_{i=1}^{n} \mid x_i \in \mathbb{R}^d, y_i \in \{0, 1\}$
+ve
-ve
$r_i \rightarrow$ vector $\rightarrow x_i$
text
column-vector

Here the rows represent the data points and the columns represent the dimensions/features. Let us assume we have 'n' data points, and each data point has components in 'd' dimensions. Then the dataset is represented mathematically in the form of a set as **$D = \{(x_i, y_i)_{i=1}^{n} | x_i \in R^d, y_i \in \{0,1\}\}$**

Here we are converting the classes from 'Negative' and 'Positive' format to 0-1 format, because linear algebra couldn't understand the terms 'Positive' and 'Negative'.

Note: As there are 'd' dimensions in every data point 'x_i', we have used the notation **$x_i \in R^d$**. As 'y_i' takes only two values (ie., 0 and 1), we have used the notation **$y_i \in \{0,1\}$**.As there are only 2 classes in the dataset, we call this problem a **binary classification** problem.

Note: If we consider the MNIST dataset, there are 60000 data points, each data point is a 784-dimensional vector and the class labels are the numbers 0 to 9, then the MNIST dataset is represented mathematically in the form of a set as

$D = \{(x_i, y_i)_{i=1}^{60000} | x_i \in R^{784}, y_i \in \{0,1,2,3,4,5,6,7,8,9\}\}$

As there are 10 classes in the MNIST dataset, we call this problem a **10-class classification** (or) a **multi-class classification** problem.