TF-IDF

Word2Vec

O.H.E

Vectorization

Feature Extraction

B.O.W

GloVe

fastText

What is Feature Extraction?

Example: Meals price prediction

Features

- No. of sides
- Limited/Unlimited?
- Veg/Non Veg?
- Location of the restaurant

Your paragraph text

What is Vectorization

Vectorization in NLP is the process of converting text data into numerical vectors that can be processed by machine learning algorithms.

	d1	d2	d3	d4	d5	d6	d7
dog →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
puppy →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
cat →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8

Types of Vectorization

Sparse Vectorization:

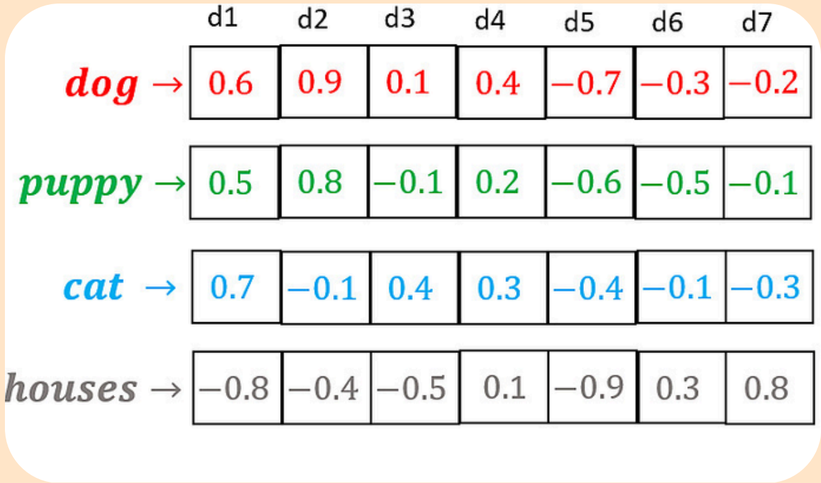
Sparse Vectorization represents data with many zeros, storing only non-zero values and their positions (e.g., Bag of Words, TF-IDF).

Dense Vectorization:

Dense Vectorization uses fixed-size, low-dimensional vectors with mostly non-zero values, capturing context and semantics (e.g., Word2Vec,fastText, BERT).

Key Difference between Sparse and Dense Vectors

- **Dense Vectors capture relationships, Sparse vectors dont**



	d1	d2	d3	d4	d5	d6	d7
dog →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
puppy →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
cat →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8

Bag Of Words

A Text representation technique treats a document as an unordered collection of words, focusing on their frequency and disregarding word order

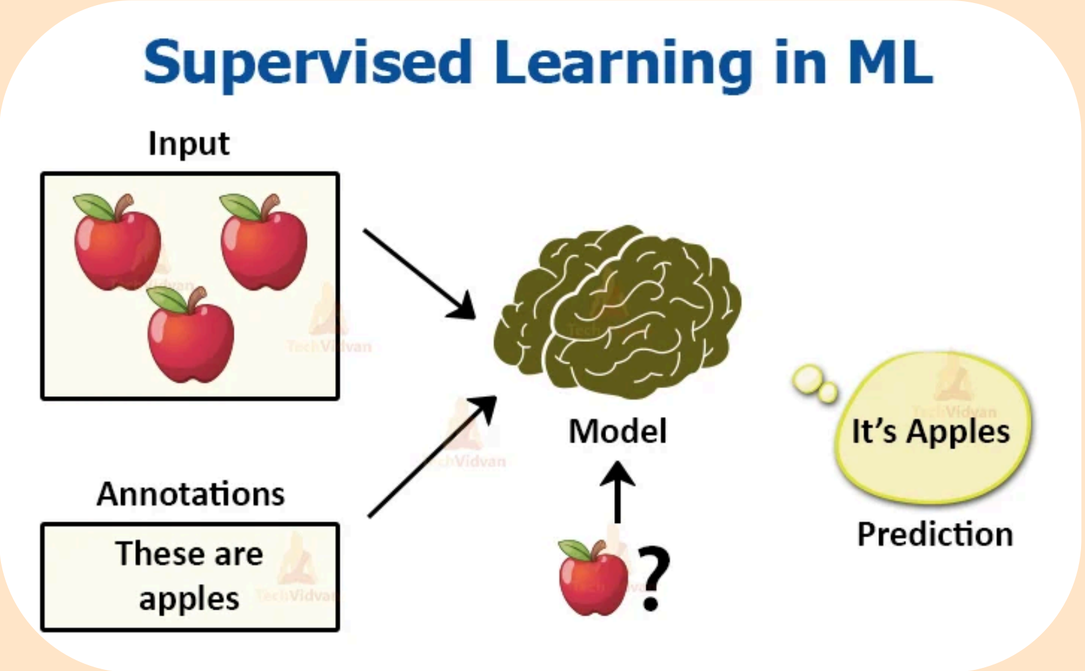
Bag Of Words

	claim	free	help	health	hurry	so	play	money	up	volleyball
play volleyball	0	0	0	0	0	0	**1**	0	0	**1**
Claim FREE money. Hurry up!	**1**	**1**	0	0	**1**	0	0	**1**	**1**	0
Playing helps health so play!	0	0	**1**	**1**	0	**1**	**2**	0	0	0

Supervised Learning

X - Images of an apple

y - Annotations/labels



Dense Vectors

	battle	horse	king	man	queen	..	woman
authority	0	0.01	1	0.2	1	...	0.2
event	1	0	0	0	0	...	0
has tail?	0	1	0	0	0	...	0
rich	0	0.1	1	0.3	1	...	0.2
gender	0	1	-1	-1	1	...	1