Bag of Words

Sentence → S1 → He is a good boy
S2 → She is a good girl
S3 → Boy & Girl are good

Step 1:- Stop Keywords & lower the Sentences :-

Stop Keywords like → He, She, is, a are not much important

the Keywords → Boy, girl, Good are important

Lower the Sentences → Convert uppercase to lowercase

S1	good	boy	
S2	good	girl	
S3	boy	girl	good

Step 2:- Create histogram → write the frequency (count) for the important Keywords

word	frequency
good	3
boy	2
girl	2

Note: When we calculate frequency, it should be in sorted order.

Step 3:- Now apply bag of words (BOW):

		f_1 good	f_2 boy	f_3 girl	O/P
Vectors ⟹ BOW	S_1	1	1	0	
	S_2	1	0	1	
	S_3	1	1	0	

binary bag of words (0 & 1)

When we apply ML Model; f_1, f_2 & f_3 are independent features & O/P is dependent feature and train the ML Model.

Here, good - 1 & Boy - 1; both words have same representation; but we dont know which word gives more importance. for this, we use the technique called TF-IDF (Term Frequency - Inverse Document Frequency)

TF-IDF

→ Term frequency (TF) = $\frac{\text{No. of repeatation of words in Sentence}}{\text{No. of words in Sentence}}$

→ Inverse-Document frequency (IDF) = $\log\left(\frac{\text{No. of Sentences}}{\text{No. of Sentences Containing words}}\right)$

→ Then calculate TF*IDF.

Given:-
S1 → good boy
S2 → good girl
S3 → good boy girl

words	frequency
good	3
boy	2
girl	2

Calculate TF:-

	S1	S2	S3
good	1/2	1/2	1/3
boy	1/2	0	1/3
girl	0	1/2	1/3

Calculate IDF:-

words	IDF
good	$\log(3/3) = 0$
boy	$\log(3/2)$
girl	$\log(3/2)$

Now Calculate: Tf*IDf :-

		f_1	f_2	f_3	o/p
		good	boy	girl	
vectors ⟹ Tf-IDF	S_1	0	$1/2 \log(3/2)$	0	
	S_2	0	0	$1/2 \log(3/2)$	
	S_3	0	$1/3 \log(3/2)$	$\frac{1}{3} \log(3/2)$	

Calculation for above table :-

	f_1 good	f_2 boy	f_3 girl
S_1	$1/2 * 0 = 0$	$\frac{1}{2} * \log(3/2) = \frac{1}{2} \log(3/2)$	$0 * \log(3/2) = 0$
S_2	$1/2 * 0 = 0$	$0 * \log(3/2) = 0$	$1/2 \log(3/2) = 1/2 * \log(3/2)$
S_3	$1/3 \log(3/3) = 0$	$1/3 * \log(3/2) = 1/3 \log(3/2)$	$1/3 \log(3/2) = 1/3 \log(3/2)$

Here, we will get decimals which gives semantic importance to boy, girl words.