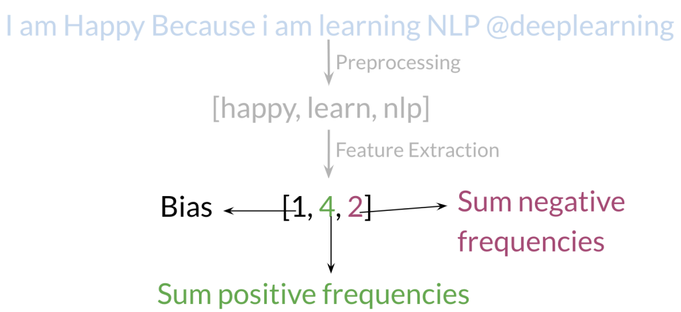
Putting it all together

Over all , you start with a given text, you perform preprocessing, then you do feature extraction to convert text into numerical representation as follows:



Your X becomes of dimension $(m, 3)$ as follows.

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_1^{(1)} & X_2^{(1)} \\ 1 & X_1^{(2)} & X_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & X_1^{(m)} & X_2^{(m)} \end{bmatrix}$$

When implementing it with code, it becomes as follows:

```
freqs = build_freqs(tweets,labels) #Build frequencies dictionary
X = np.zeros((m,3)) #Initialize matrix X
for i in range(m): #For every tweet
    p_tweet = process_tweet(tweets[i]) #Process tweet
    X[i,:] = extract_features(p_tweet,freqs) #Extract Features
```

You can see in the last step you are storing the extracted features as rows in your X matrix and you have m of these examples.

Mark as completed

Like Dislike Report an issue