Spam sms Documentation

Classification problem

Modus Operand:

1. Clean the data
   1. remove empty columns
   2. Label Encode spam and ham to 1 & 0 respectively
   3. Checking null values
   4. Checking duplicates and removing them
2. Exploratory Data Analysis(EDA) :-Analyze the data
   1. Showing percentage of spam , not spam in pie chart
   2. We count no of alphabets,words,sentences in each sms and analyze that
   3. Use Histogram,Corr Coefficient to take decision
3. Data preprocessing : Vectorization , other edits

1. Lower Case

2. Tokenization

3. Removing Special Characters

4. Removing stop words and punctuations

5. Stemming - converting words like run running etc to one word

6. Adding this converted text into a new column

7. Creating word-cloud of spam and not spam

1. Model building
2. For vectorizing text we start with bag of number
3. Taking X\_trian,y\_train
4. Taking tfidf vectorizer
5. Implementing naïve bayees Multi Dimensional because precision score is 1
6. Evaluation
7. Improvement
8. Website implement
9. Creating a pipeline
10. Connecting with streamlit