

ASSIGNMENT-7

SPAM DETECTOR

1) Describe the process you followed to create the model?[3]

I have implemented the Spam Detector model by using Naïve Bayes model in R.

➔ Firstly I have loaded the spam.csv file into workspace which has empty data in the columns 3 to 5 so I have read only first two columns of csv file which contains required data for the model. On this data I have performed following steps for building the model.

- 1) **Tokenization:** Dividing each row of the column containing message into number of documents and tokenizing each word in the document.
- 2) **Document Preprocessing:** Each message in the csv file contains similar words in uppercase and lower case, contains exclamations, whitespaces and different types of punctuations which are to be removed for creating the model.
- 3) **Creating Document Matrix:** The above cleaned and filtered data has to be converted to matrix form where rows contains different types of words in the file and columns contains whether that word is present in particular document or not.(this is a binary matrix having 0/1 as matrix values).
- 4) **Finding frequently occurred words:** From the document matrix we have to find the words that are frequently occurred (which are repeated specific number of times which we mention in the code. I have selected words that are repeated at least 10 times in data).
- 5) **Data partitioning:** 80% of data has been partitioned into training and 20% into testing dataset where training dataset is used to create model using Naïve Bayes and testing is used to predict the accuracy of model. Also at this stage we count the number of times each word is repeated.
- 6) **Running model using Naïve Bayes classification:** Applying Naïve Bayes classification on training dataset and training data partitioned from raw dataset which results in generating conditional probabilities of independent variables.
- 7) **Running the model with test dataset and generating confusion matrix:** The test dataset created is tested with the model and a confusion matrix is generated which gives the accuracy of our model.

2) What data preprocessing steps did you implement for your model?

The data preprocessing steps that I have done in my model are:

- 1) I have deleted all the words which doesn't have the meaning in English literature.
- 2) I have removed the extra spaces in each tokens present in each document.
- 3) I have converted all the words to lowercase because 2 same words one in Upper case and other in lower case are read as two different words which may result in wrong predictions.
- 4) I have deleted the punctuations and numbers in each tokens present in each document or row.

3) In general what were the features of your model?

The model can predict whether a given message is spam/ham based on the messages previously received and generating a classification model using Naïve Bayes or K-nearest algorithms with an accuracy of nearly 98%.

The model is made by classifying 80% data into training dataset and 20% into testing dataset where model is build using training dataset and tested using testing dataset.

4) What is F1 scoring formula and in what circumstances is it better than the accuracy formulae? [1] [2]

A F1 scoring formulae also called as F score conveys the balance between precision and recall which is given by formulae $2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$. In certain circumstances it gives the accuracy of our model. The accuracy formula works best when false positives and false negatives have similar score. When the false negatives and false positives have very different scores it is better to look at precision and accuracy. For example: In our model the false positive score is 4 whereas false negative score is 966 so there is huge difference in these scores, so in this case it is better to look at f1 score of our model than accuracy.

5) What is the F1 score of your model?

The F1 score of my model is **89.21%** (where as the accuracy is 97.39%).

References

- [1] Exsilio solutions, "Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures," [Online]. Available: <https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/>.
- [2] S. Narkhede, "Understanding confusion matrix," [Online]. Available: <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>.
- [3] D. Sangeetha, "SMS-HAM|SPAM?," [Online]. Available: <https://www.kaggle.com/devisangeetha/sms-ham-spam/notebook>.