COMP 6721

Applied Artificial Intelligence

Instructor: Dr. René Witte

Team name: HA-G07

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Student name** | **Student ID** | **Email address** |
| **1** | Jil Mehta | 40089806 | mehtajil131@gmail.com |
| **2** | Sarvesh Vora | 40081458 | vorasarvesh99@gmail.com |

Date: April 25, 2020

The below matrices were generated using our program by maintaining following counters,

TruePositive (TP): when actual class is Ham and model predicts as Ham

TrueNegative(TN): when actual class is Ham and model predicts as Spam

FalsePositive(FP): when actual class is Spam and model predicts as Spam

FalseNegative(FN): when actual class is Spam and model predicts as Ham

The above counters are put into the matrices according to their description.

Ham Matrix:

|  |  |  |
| --- | --- | --- |
| **Model predicts** | **Email is Ham in real** | **Email is Spam in real** |
| **Email is Ham** | TP = 394 | FP = 6 |
| **Email is Spam** | FN = 64 | TN = 336 |

Spam Matrix:

|  |  |  |
| --- | --- | --- |
| **Model predicts** | **Email is Spam in real** | **Email is not Spam in real** |
| **Email is Spam** | TP = 336 | FP = 64 |
| **Email is not Spam** | FN = 6 | TN = 394 |

, , ,

Table of results:

|  |  |  |
| --- | --- | --- |
| **Measures** | **While detecting Ham** | **While detecting Spam** |
| **Accuracy** | 91.25% | 91.25% |
| **Precision** | 98.5% | 84% |
| **Recall** | 86.026% | 98.245% |
| **F1-measure** | 91.84% | 90.5657% |

The precision for finding Ham is high and recall is low whereas for finding spam the recall is high and precision is less, therefore we require F1-measure which allow us to compare our model to find Ham and Spam. Here we can observe that our model performs better when it comes to detecting Ham emails. Because the F1 score of predicting Ham is better than F1 score of predicting Spam. We can improve the performance by filtering the common words like “a”, ”the”, “from”, “received” etc. as they occur very frequently in the documents and thus their continuous presence can drastically impact the classification process. In order to get the actual words which, affect the probability, we can use term frequency and inverse document frequency to determine some of the most commonly occuring words.

References:

<https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>

<https://stackoverflow.com/questions/4880960/how-to-sum-all-the-values-in-a-dictionary>

<https://en.wikipedia.org/wiki/Confusion_matrix>

<https://stats.stackexchange.com/questions/49226/how-to-interpret-f-measure-values>

Lecture slides on Naïve Bayes and NLP with related worksheets.