**Aim / Intro**

Aim and why the study I important

Naïve Bayes is widely used as an open source spam detection filter due to the simplicity in its implementation. This study aims to evaluate its accuracy in a Java implementation constructed specifically for this study, comparing it with other spam detection methods implemented in Weka (X). While it is widely used there are many variations of Naïve Bayes where the implementation itself differs as well as the pre-processing selection stage. These variations are an important area of study in terms of how its performance may be improved (or decline) and then compared to other spam detection methods. \*\*Stretching\*\*

**Data Pre-processing**

Describe what we did, include number of words for each corpus (before and after removing stop words). List top 100 words for each corpus and show their DF score. Does the selection make sense given the task? Are the two lists different?

**Subject vs Body**

|  |  |
| --- | --- |
| **Corpus:** Subject | |
|  | **Accuracy [%]** |
| **ZeroR** |  |
| **OneR** |  |
| **1-NN** |  |
| **3-NN** |  |
| **NB** |  |
| **DT** |  |
| **MLP** |  |
| **MyNB** |  |
|  | |
| **Corpus:** Body | |
|  | **Accuracy [%]** |
| **ZeroR** |  |
| **OneR** |  |
| **1-NN** |  |
| **3-NN** |  |
| **NB** |  |
| **DT** |  |
| **MLP** |  |
| **MyNB** |  |

Compare the performance of the two, compare our NB with Weka’s NB and anything else important

**Challenge Results**

Describe what we did and what the results are (accuracy) and discuss

**Conclusion**

Summarise main findings and suggest further work if possible

**Reflection**

What was the most important learning to you?