**Result:**

The whole dataset is divided into 30% training set, 20% validation set and 50% application set. Both training set and validation set is random shuffled, while application set is time series. The best model is selected from the validation set, and is used to be applied to the application set.

Table 1 and table 2 displays the number and proportion of each categories in training and validation set.

|  |  |  |
| --- | --- | --- |
| Training Set | | |
|  | **Count** | **Proportion** |
| Technology | 1421 | 29.7% |
| Entertainment & Arts | 872 | 18.2% |
| Business | 1941 | 40.5% |
| Health | 227 | 4.7% |
| Science & Environment | 327 | 6.8% |

**Table 1. Number and proportion of each categories in training set.**

|  |  |  |
| --- | --- | --- |
| Validation Set | | |
|  | **Count** | **Proportion** |
| Technology | 960 | 30.1% |
| Entertainment & Arts | 591 | 18.5% |
| Business | 1262 | 39.5% |
| Health | 171 | 5.4% |
| Science & Environment | 209 | 6.6% |

**Table 2. Number and proportion of each categories in validation set.**

**Effectiveness**

We evaluate the effectiveness of the models based on its precision, recall and f1-score. Since the data is imbalanced and categories with smaller proportion should not be omitted, we consider both micro average and macro average.

The difference between micro and macro average is that micro average gives equal weight on each article, while macro average gives equal weights on each category. A model with high micro average and low macro average performs bad on categories with small proportion. A good model in text classification should perform well on both averages.

Table 3, 4 and 5 are the effectiveness of Naive Bayes, SVM and softmax regression.

|  |  |  |  |
| --- | --- | --- | --- |
| Naïve Bayes | | | |
|  | **precision** | **recall** | **f1score** |
| MicroAvg | 0.861 | 0.848 | 0.836 |
| MacroAvg | 0.907 | 0.698 | 0.741 |

**Table 3. Effectiveness of Naïve Bayes.**

|  |  |  |  |
| --- | --- | --- | --- |
| Support Vector Machine | | | |
|  | **precision** | **recall** | **f1score** |
| MicroAvg | 0.934 | 0.932 | 0.932 |
| MacroAvg | 0.930 | 0.927 | 0.931 |

**Table 4. Effectiveness of SVM.**

|  |  |  |  |
| --- | --- | --- | --- |
| Softmax Regression | | | |
|  | **precision** | **recall** | **f1score** |
| MicroAvg | 0.932 | 0.931 | 0.931 |
| MacroAvg | 0.930 | 0.933 | 0.932 |

**Table 5. Effectiveness of Softmax regression.**

**Coefficients**

Besides effectiveness, it is also possible to interpret the result of Naïve Bayes, SVM and softmax regression by considering its coefficients. Words with higher coefficients are more important predictors for predicting the category of an article. Table 6, 7, 8, 9 and 10 displays the words with top 5 highest coefficients for each category in Naïve Bayes, SVM and softmax regression. Be reminded that these words are preprocessed with stemming.

|  |  |  |  |
| --- | --- | --- | --- |
| Business | | | |
| No. | **Naïve Bayes** | **SVM** | **Softmax** |
| 1 | bank | busi | busi |
| 2 | compani | bank | bank |
| 3 | busi | compani | compani |
| 4 | market | tradit | mr |
| 5 | growth | mr | tradit |

**Table 6. Words with highest coefficients for Business**

|  |  |  |  |
| --- | --- | --- | --- |
| Entertainment & Arts | | | |
| No. | **Naïve Bayes** | **SVM** | **Softmax** |
| 1 | film | art | film |
| 2 | star | film | star |
| 3 | award | star | art |
| 4 | actor | theatr | artist |
| 5 | play | bbc | bbc |

**Table 7. Word with highest coefficients for Entertainment & Arts**

|  |  |  |  |
| --- | --- | --- | --- |
| Health | | | |
| No. | **Naïve Bayes** | **SVM** | **Softmax** |
| 1 | patient | nh | nh |
| 2 | nh | care | patient |
| 3 | health | health | care |
| 4 | hospit | cancer | health |
| 5 | cancer | patient | hospit |

**Table 8. Word with highest coefficients for Health**

|  |  |  |  |
| --- | --- | --- | --- |
| Science & Environment | | | |
| No. | **Naïve Bayes** | **SVM** | **Softmax** |
| 1 | space | scienc | climat |
| 2 | scientist | rocket | scienc |
| 3 | earth | climat | rocket |
| 4 | climat | speci | engin |
| 5 | scienc | engin | scientist |

**Table 9. Word with highest coefficients for Science & Enrionment**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Technology | | |
| No. | **Naïve Bayes** | | **SVM** | **Softmax** |
| 1 | user | | googl | firm |
| 2 | firm | | site | googl |
| 3 | googl | | game | site |
| 4 | compani | | devic | devic |
| 5 | devic | | firm | game |

**Table 10. Word with highest coefficients for Technology**

**Discussion**

**Effectiveness**

Naïve Bayes performs fine if consider only micro averages. Its macro average for recall, however, is substantially low. This might due to Naïve Bayes largely favors categories with larger proportions. A low macro average disables the possibility to use prediction to predict, since eventually it will predict only categories with larger proportion. This is described more in detailed in *Prediction with Limited Data* section in this report.

Both SVM and softmax regression have similar effectiveness. Not only do they perform well on precision and recall, but also on micro average and macro average. These are positive signs for a good model in text classification.

**Coefficients**

Since we are using linear models without dimension reduction for Naïve Bayes, SVM and softmax regression, it is easy to interpret the fitted models by observing the words with highest coefficients for each category.

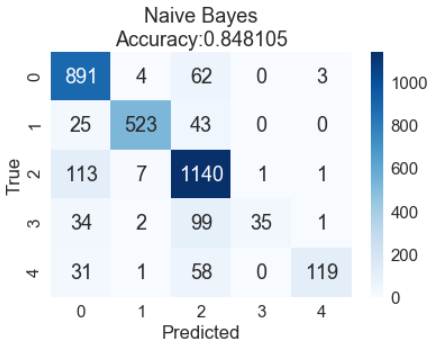
Words with highest scores in SVM and softmax regression are highly identical. They have almost the same words with only slightly differences in order. We may conclude that SVM and softmax regression had resulted in very similar models. This might also explain why they have similar precision and recall.

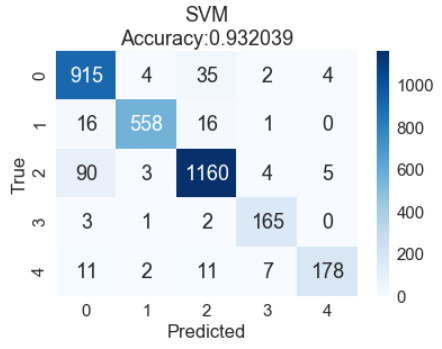
Although there is similarity between the words of highest coefficient between Naïve Bayes and the other two models, there is still substantial differences.

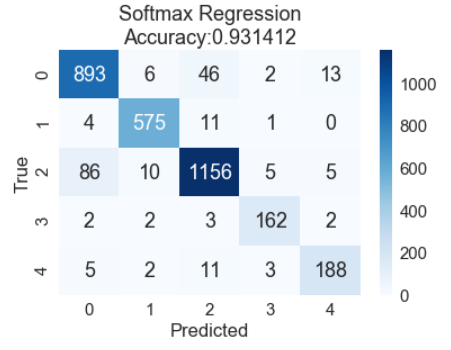
**Conclusion**

The effectiveness of Naïve Bayes is lower than SVM and softmax regression. We should choose the best model between SVM and softmax regression. These two models, however, not only have similar effectiveness, but also have similar models while analyzing words with higher coefficient. There is, however, slight differences between their macro average recall in this validation set. Hence, we shall choose softmax regression as the best model for this dataset.

Appendix:







|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1-score |
| 0 | 0.814 | 0.928 | 0.868 |
| 1 | 0.974 | 0.885 | 0.927 |
| 2 | 0.813 | 0.903 | 0.856 |
| 3 | 0.972 | 0.205 | 0.338 |
| 4 | 0.960 | 0.569 | 0.715 |
| MicroAvg | 0.861 | 0.848 | 0.836 |
| MacroAvg | 0.907 | 0.698 | 0.741 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1-score |
| 0 | 0.884 | 0.953 | 0.917 |
| 1 | 0.982 | 0.944 | 0.963 |
| 2 | 0.948 | 0.919 | 0.933 |
| 3 | 0.922 | 0.965 | 0.943 |
| 4 | 0.952 | 0.852 | 0.899 |
| MicroAvg | 0.934 | 0.932 | 0.932 |
| MacroAvg | 0.930 | 0.927 | 0.931 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1-score |
| 0 | 0.902 | 0.930 | 0.916 |
| 1 | 0.966 | 0.973 | 0.970 |
| 2 | 0.942 | 0.916 | 0.929 |
| 3 | 0.936 | 0.947 | 0.942 |
| 4 | 0.904 | 0.900 | 0.902 |
| MicroAvg | 0.932 | 0.931 | 0.931 |
| MacroAvg | 0.930 | 0.933 | 0.932 |