COMP472 ASSIGNMENT 3 DESIGN DOC

team: anonymous

Yifan Yang 40038814

1. Analysis of initial dataset

1.1 text processing

"For the average American the best way to tell if you have covid-19 is to cough in a rich person鈥檚 face and wait for their test results"

The initial data has a lot of non-english char, symbols and stopwords, but they do not have significant impact on the decision.

We only keep English letter&&number&&'-' by regular expression and turn them to lower case

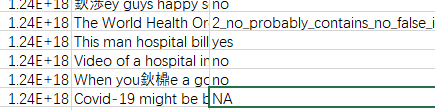


Then in the rest of the words, there are still some stopwords, we remove them by nltk.stopwords

<https://gist.github.com/sebleier/554280>

1.2 q1\_label

There are some q1\_label whose value is not yes or no, this type of data won't be recorded



2 NB-BOW-OV vs NB-BOW-FV

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **NB-BOW-OV** | | **NB-BOW-FV** | |
|  | YES | NO | YES | NO |
| Accuracy | 72.72 % | | 69.09％ | |
| Precision | 72.5% | 73.33% | 70.0% | 66.66% |
| Recall | 87.87% | 50.0% | 84.84% | 45.45% |
| F1\_measure | 79.45% | 59.45% | 76.7% | 54.05% |

Accuracy:

All the measures are better in NB-BOW-OV model because NB-BOW-FV remove some vocabulary so the NB-BOW-OV has larger training data

Time Complexity:

However, there are 5700 words and 408576 cells in NB-BOW-FV,7909 words and 1289967 cells in NB-BOW-OV, word-filtering has greatly reduced the time we spend on processing data with acceptable performance decrease. If dataset is bigger or model is more complex, vocabulary filter will be more useful to save time. We can also assume that words like 'covid-19' that appear more than 2 times, are more important than other words in decision.

Similarity:

In both NB-BOW-OV and NB-BOW-FV, Yes-Recall is significantly better than No-Recall, which means that the model tends to guess 'yes' , one possible reason is that there are more 'yes' in test data and test data is small.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **NB-BOW-OV** | | **NB-BOW-FV** | | **LSTM** | |
|  | YES | NO | YES | NO | YES | NO |
| Accuracy | 72.72 % | | 69.09 % | | 78.18 % | |
| Precision | 72.5% | 73.33% | 70.0% | 66.66% | 75.61% | 85.71% |
| Recall | 87.87% | 50.0% | 84.84% | 45.45% | 93.94% | 54.55% |
| F1\_measure | 79.45% | 59.45% | 76.7% | 54.05% | 83.78% | 66.67% |

Overall, LSTM generates better results than NB-BOW-OV and NB-BOW-FV, and every values is close to NB-BOW-OV and NB-BOW-FV

It also follows the same tendency that YES-recall is far better than NO-recall

Another major difference is the randomness of LSTM. The results of LSTM are not the same after training. While the result of Naïve-Bayes is fixed.

