# COMP30027 Machine Learning – Assignment 1

Erich Wiguna 1389444

## 1. Supervised model training

Results for Supervised model training

```
Prior probability of scam, P(scam) = 0.2000
Prior probability of non-malicious, P(non-malicious) = 0.8000
```

***Figure 1.1*** *Prior probability values for both scam and non-malicious classes*

```
Most probable words in scam class:
.: 0.0565
!: 0.0243
,: 0.0235
call: 0.0205
£: 0.0139
free: 0.0105
/: 0.0091
2: 0.0088
&: 0.0087
?: 0.0085

Most probable words in non-malicious class:
.: 0.0793
,: 0.0260
?: 0.0256
u: 0.0189
...: 0.0187
!: 0.0172
..: 0.0149
;: 0.0132
&: 0.0131
go: 0.0111
```

***Figure 1.2*** *Top 10 most probable words along with their probability values for both scam and non-malicious classes*

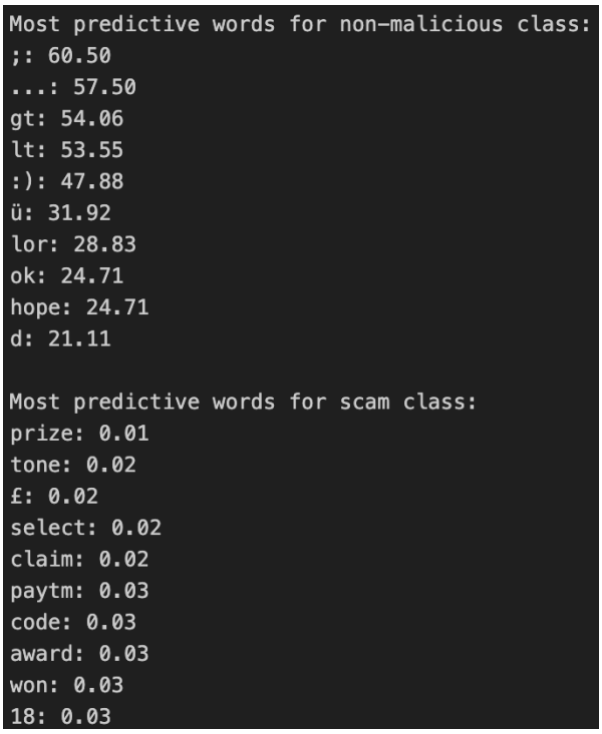


***Figure 1.3*** *Top 10 most predictive words along with their probability ratios for both scam and non-malicious classes*

### Discussion for Supervised model training

As seen in Figure 1.1, there is a clear imbalance in the training dataset where the prior probability of non-malicious instances is 80%. This imbalance in the dataset denotes that the model would highly predict towards non-malicious class unless the message contains a highly probable or predictive words for scam.

The multinomial Naïve Bayes model seems to be quite capable of distinguishing scam and non-malicious messages based on word frequency alone. This can be seen in Figure 1.2, where the most probable words in scam messages are “!”, “call”, “£”, and “free”. These words are usually found in promotional messages that would typically fall into spam messages. In the case of non-malicious messages, common punctuations and words like “u” and “go” are most probable as they are more personal.

This is further supported in Figure 1.3 where words like “prize”, “£”, “claim”, “award”, and “won” are more predictive of scam class. Where words like “:)” (smile emoji), “hope”, and “ok” (informal form of yes/okay) are more likely to appear in non-malicious class.

The clear separation of vocabularies between the two class has allow the multinomial Naïve Bayes model to work quite well in this case. However, as the model assume word independence and there is a loss of word order and hence word context in the text pre-processing, this model is not the optimal solution for this case.

## 2. Supervised model evaluation

### Results for Supervised model evaluation

```
Overall Accuracy: 0.9750
Confusion Matrix:
[[785  15]
 [ 10 190]]
Precision (Non-Malicious, Scam): [0.98742138 0.92682927]
Recall (Non-Malicious, Scam): [0.98125 0.95   ]
```

***Figure 2.1*** *Performance report of the supervised model on test data (Accuracy, Confusion Matrix, Precision, and Recall)*

```
Total OOV words encountered: 1571
Total skipped test instances: 0
```

***Figure 2.2*** *Total out-of-vocabulary words and total skipped instances from test data*

```
Examples of Scam Classified with High Confidence:
R=0.00 | Text: . 4 + call £ - * holiday & urgent 18 t landline 150ppm cash cs await collection po box sae complimentary 10,000 ibiza
R=0.00 | Text: . 3 4 + ! call : £ offer * holiday & urgent 18 t landline 150ppm cash cs await collection po box sae tenerife 10,000
R=0.00 | Text: . . . , please order text call / : customer tone number [ [ service mobile ] ] colour colour thanks ringtone reference charge 4.50 arrive = red x49 09065989182
R=0.00 | Text: . call £ £ guarantee won customer prize prize claim service 1000 yr 2000 representative cash 10am-7pm
R=0.00 | Text: . . 2 free u + ! 1st / / wk wk txt tone gr8 hit 150p 16 poly 8007 8007 nokia nokia nokia tones polys

Examples of Non-Malicious Classified with High Confidence:
R=913499445162972038862535218251358535680.00 | Text: time : rs. transaction number & & & & & & & & & ; ; ; ; ; ; ; ; ; ; lt lt lt lt lt # # # gt gt gt gt gt credit account reference decimal
R=269038185856116170020844732416.00 | Text: ? ? ? ? .. .. u u u u , , ... ... ... ... say person yes ! f : hello hello hello o o wen knw knw girl girl mean @ " " " " t name name g g n d d d
R=318292454118529755303116800.00 | Text: . every & & & & & & ; ; ; ; ; ; lt lt lt # # # gt gt gt big hr
R=6035603737741317898240.00 | Text: , get like second half & & & & ; ; ; ; lt lt # # gt gt run though almost whole gram gram usually
R=38221697184201940992.00 | Text: u , , lor ... ... ... ... food food eat den oso haha well depend mon n la wana okie okie cheap chinese gd ex

Examples on the Decision Boundary (Uncertain Classification):
R=1.02 | Text: . call dear
R=1.04 | Text: . reply glad
R=0.93 | Text: . . tell return re order
R=0.93 | Text: ? ur * just alrite sam
R=0.88 | Text: . . reply send person right ! code confirm sort bank acc
```

***Figure 2.3*** *Example of high, low, and R = 1 confidence ratio instances along with their confidence ratios*

### Discussions for Supervised model evaluation

In Figure 2.1, the multinomial Naïve Bayes classifier achieved an overall accuracy of 97.5% on the test set and can be further observed from the confusion matrix where it identifies 785 non-malicious messages out of 800 and 190 scam messages out of 200. Although, the confusion matrix indicates strong performances on both classes, there are clear distinctions between the prediction of the two classes, where both precision and recall of scam class are lower compared to non-malicious class. This suggests that the model is having a harder time in identifying scam messages compared to non-malicious messages.

As seen in Figure 2.2, there are a total of 1,571 out-of-vocabulary words in the test set with no skipped instances, meaning that there is at least one known word in all of test instances. In terms

of confidence, there is a clear separation between messages with low and high R values that can be observed in Figure 2.3. Messages containing casual and conversational words like “hello”, “u”, and “okie” tend to have higher R, meaning the model is confident in identifying messages containing these words. In contrast, messages that contains promotional and persuasive words like “call”, “£”, “free”, and “claim” are more likely to be classified as scam by the model. Lastly, for messages where the model has low confidence typically contains a short messages like “. call dear” and “. reply glad” as there is not enough information for the model to rely on.

# 3. Extending the model with semi-supervised training: Active Learning

## Results for Extending the model with semi-supervised training: Active learning

```
Selecting 200 random instances

New expanded prior probabilies {0: 0.8, 1: 0.2}
Accuracy: 0.9600
Confusion Matrix:
[[309  11]
 [  5  75]]
Precision (Non-Malicious, Scam): [0.98407643 0.87209302]
Recall (Non-Malicious, Scam): [0.965625 0.9375  ]
```

***Figure 3.1*** *Performance report of the random selection model on validation dataset (Accuracy, Confusion Matrix, Precision, and Recall)*

```
Selecting 200 most uncertain instances

New expanded prior probabilies {0: 0.8045454545454546, 1: 0.19545454545454546}
Accuracy: 0.9675
Confusion Matrix:
[[311   9]
 [  4  76]]
Precision (Non-Malicious, Scam): [0.98730159 0.89411765]
Recall (Non-Malicious, Scam): [0.971875 0.95    ]
```

***Figure 3.2*** *Performance report of the uncertain selection model on validation dataset (Accuracy, Confusion Matrix, Precision, and Recall)*

## Discussions for Extending the model with semi-supervised training: Active learning

To further enhance the performance of the supervised Naïve Bayes classifier, an active learning approach is implemented by selecting 200 additional labelled instances from the sms_unlabelled.csv. Before selecting the additional instances, a stratified sampling is used to sample 20% of the sms_unlabelled.csv dataset as validation set. For the baseline of the active learning approach, random sampling method is used where the 200 additional instances are selected at random. After adding those additional instances to the previous supervised training set, a test on the validation set is run that produce a performance result like in Figure 3.1. In contrast to the random sampling, the uncertainty-based sampling is used where the 200 additional instances are selected based on their ratio (R-value). Using the model from Q1, the posterior probability ratio for each unlabeled instance is calculated. The 200 instances with the lowest R-value are then selected as lower R-value means that the previous model has low confidence in identifying these 200 instances. After selecting, the same method is undergone as before where the previous model is retrained with the expanded training set. In addition, the same testing is undergone using the validation set where result like Figure 3.2 is produced.

Although, the random selection might be faster and easier to implement, the uncertain selection provides better insights about instances where the model has low confidence in, hence improving the model's confidence for those instances. This is supported by both figures where the uncertainty selection outperformed random selection slightly across all metrics.

## 4. Supervised model evaluation

Results for Supervised model evaluation

```
Evaluation on test set
New expanded prior probabilies {0: 0.8045454545454546, 1: 0.19545454545454546}
Accuracy: 0.9770
Confusion Matrix:
[[786  14]
 [  9 191]]
Precision (Non-Malicious, Scam): [0.98867925 0.93170732]
Recall (Non-Malicious, Scam): [0.9825 0.955 ]
```

***Figure 4.1*** *Performance report of the expanded semi-supervised model on test data (Accuracy, Confusion Matrix, Precision, and Recall)*

```
Most probable words in scam class:
.: 0.0647
!: 0.0274
,: 0.0267
call: 0.0228
£: 0.0151
free: 0.0117
/: 0.0101
&: 0.0101
2: 0.0099
?: 0.0097

Most probable words in non-malicious class:
.: 0.0863
,: 0.0283
?: 0.0273
u: 0.0194
...: 0.0189
!: 0.0186
..: 0.0153
&: 0.0133
;: 0.0133
go: 0.0113
```

***Figure 4.2*** *Top 10 most probable words along with their probability values for both scam and non-malicious classes in the semi-supervised model*

```
Most predictive words for non-malicious class:
gt: 101.24
lt: 100.28
:): 89.62
ü: 59.58
lor: 53.77
;: 45.44
d: 39.24
da: 37.30
...: 36.01
let: 33.43

Most predictive words for scam class:
prize: 0.00
tone: 0.01
select: 0.01
paytm: 0.01
code: 0.01
18: 0.02
won: 0.02
ringtone: 0.02
claim: 0.02
>: 0.02
```

***Figure 4.3*** *Top 10 most predictive words along with their probability ratios for both scam and non-malicious classes in the semi-supervised model*

```
Examples of Scam Classified with High Confidence:
R=0.00 | Text: . 4 + call £ − * holiday & urgent 18 t landline 150ppm cash cs await collection po box sae complimentary 10,000 ibiza
R=0.00 | Text: . 3 4 + ! call : £ offer * holiday & urgent 18 t landline 150ppm cash cs await collection po box sae tenerife 10,000
R=0.00 | Text: . . . , please order text call / : customer tone number [ [ service mobile ] ] colour colour thanks ringtone reference char
R=0.00 | Text: . . 2 free u + ! 1st / / wk wk txt tone gr8 hit 150p 16 poly 8007 8007 nokia nokia nokia tones polys
R=0.00 | Text: . call £ £ guarantee won customer prize prize claim service 1000 yr 2000 representative cash 10am-7pm

Examples of Non-Malicious Classified with High Confidence:
R=14766753787954904928609148912159455969 28.00 | Text: time : rs. transaction number & & & & & & & & & ; ; ; ; ; ; ; ; ; ; lt lt lt lt lt
R=665550098570132317756084505804 8.00 | Text: ? ? ? ? .. .. u u u u , , ... ... ... ... say person yes ! f : hello hello hello o o wen knw
R=13254466384147295354277068 8.00 | Text: . every & & & & & & ; ; ; ; ; ; lt lt lt # # # gt gt gt big hr
R=264243125767869657251 84.00 | Text: u , , lor ... ... ... ... food food eat den oso haha well depend mon n la wana okie okie cheap chines
R=229508563455706262405 12.00 | Text: , get like second half & & & & ; ; ; ; lt lt # # gt gt run though almost whole gram gram usually

Examples on the Decision Boundary (Uncertain Classification):
R=1.00 | Text: . send link picture
R=1.02 | Text: . reply glad
R=1.07 | Text: . free / give otherwise
R=0.89 | Text: . university
R=1.16 | Text: . call dear
```

***Figure 4.4*** *Example of high, low, and R ≈ 1 confidence ratio instances along with their confidence ratios for the expanded model*

```
Amount of test instance with R-value between 0.8 and 1.2 (using Q1 model): 8
Amount of test instance with R-value between 0.8 and 1.2 (using Q3 model): 5
```

***Figure 4.5*** *Comparison of total instances with R-values around 1 between Q1 model and Q3 model*

## Discussions for Supervised model evaluation

In comparison to the previous supervised model from Q1, the semi-supervised model seems to perform better on the test set given the performance result in Figure 4.1. There is noticeable performance increase in accuracy, precision, and recall which can be better observed in the confusion matrix where there is an increase from 785 true positive to 786 and from 190 true negative to 191. However, this model also suffers from lower precision and recall in classifying scam messages compared to non-malicious meaning that the model performs better in identifying non-malicious than scam.

Additionally, the semi-supervised training has changed the model's representation in the most probable words, most predictive words, and in the model's confidence. As for the most probable words in Figure 4.2, there are some increases in the probability of some of the already frequent words from the first model, suggesting that the additional 200 instances have similar frequent words as the previous dataset. As for the most predictive words in Figure 4.3, there are also some increases in the confidence ratio in some words denoting that the model becomes more confident in identifying non-malicious message if it contains those words. In contrast, we cannot really tell if the model becomes more confident in identifying scam message as the confidence ratio of scam words cannot be lower than zero. A deeper observation can be made regarding the model's confidence as seen in Figure 4.4 where there is almost no change in both scam and non-malicious class where the model's confidence is at its lowest and highest respectively. However, there are some changes in the model's confidence in the uncertain instances that can be further observed in Figure 4.5, where there are only 5 instances that have an R-value between 0.8 and 1.2 compared to the previous model with 8 instances, suggesting a decline in the model's uncertainty.