COMP30027
Assignment 1
Jessica Le 1353984

Supervised model training

Using the supervised training dataset, a multinomial Naïve Bayes model was implemented using the pre-processed text as the input. Any missing values were replaced with an empty string. The prior probabilities of a scam text were P(class = 1) = 0.2 and for a non-malicious text was P(class = 0) = 0.8.

Below is a table which displays the 10 most probable words from each class. The two groups have an overlap of five out of the ten words. The majority of words were punctuation which is could normally be present in both scam and regular texts without much alarm. By looking at the ten most probable words for both classes alone, it is hard to start making predictions

Non-malicious, class 0	
Word	**Probability**
.	0.0793
,	0.0260
!	0.0172
?	0.0256
&	0.0131
u	0.0189
..	0.0149
;	0.0132
...	0.0188
go	0.0111

Scam, class 1	
Word	**Probability**
.	0.0565
,	0.0235
!	0.0244
?	0.0085
&	0.0087
free	0.0105
/	0.0091
2	0.0088
£	0.0139
call	0.0205

However, when we start to consider the most predictive words for both classes, we start to see more recurring and meaningful words, especially in scam messages. When considering the most predictive scam words, the results seem reasonable as most words allude to winning and money such as “prize”, “euros”, “won” or “award”, with prize having the greatest likelihood of being in a scam message. On the other hand, non-malicious predictive words are less overt but still seem reasonable. It is more likely that a human would use more complex punctuation or emotes such as the semi colon. elipses and the punctuated smiley face :).

Predictive scam words	
Word	**R**
prize	99.0284
tone	64.0772
£	49.7084
select	46.0616
claim	45.9543
paytm	36.8929
code	34.9512
award	32.0386
won	31.0677
18	29.1260

Predictive non malicious words	
Word	**R**
;	60.5130
...	57.5088
gt	54.0754
lt	53.5604
:)	47.8954
ü	31.9302
lor	28.8402
hope	24.7202
ok	24.7202
d	21.1152

Supervised model evaluation

The Q1 model achieved an overall accuracy of 97.5% on the test dataset. The confusion matrix showed:

	Predicted Non-Malicious	Predicted Scam
Actual Non-Malicious	785	15
Actual Scam	10	190

Precision and recall for the scam class were 93% and 95%, respectively. The model did not skip any messages meaning all messages contained at least 1 word from the vocabulary.

The prediction confidence was analysed using the posterior ratio R = P(scam)/P(non-malicious).

High-confidence scam examples:

- R = 483730 : "! call customer important service announcement freephone 0800 542"
- R = 184952 : " ? . reply send stop stop / invite friend
- R = 8296831 : "call guarantee won customer prize prize claim service 1000"

High-confidence non malicious examples:

- R = 0.0002 : "hey mate honey have holiday x"
- R = 0.0063 : "house-maid murderer coz man murder 26th January"
- R = 0.0006 : "just take derek and amp ; leave ready"

Boundary examples (R = ~1.00)

- R = 0.9825 : " call dear ."
- R = 0.9571 : " reply glad ."
- R = 1.0783 : "ur just alrite * sam ?"

High confidence scam messages seem quite reasonable based off the words used. Most allude to winning a prize or money or a free service which is in line with real world situations. High confidence non-malicious examples were reasonable with most sounding like a personal message to friends and family; the second message funnily may sound somewhat like a threat but does not sound like a scam. Scam messages are easier to identify than non-malicious messages because most follow a similar idea of free prizes or services and therefore use a similar vocabulary which can be identified. On the other hand, non-malicious messages is every other topic and is harder to classify.

Extended semi-supervised training using Label Propagation

To extend the model using label propagation, unlabelled dataset was classified using the supervised model from Q1. Only high-confidence predictions based on posterior ratio $R > 10$ or $R < 0.1$ where used, resulting in 1,858 messages.

The model was then retrained using the original supervised set plus these examples. This increased the training size and increased the vocab size.

Evaluation on the test set showed improved performance:

- Accuracy increased to 92.6%
- Precision (scam): 90.5%
- Recall (scam): 89.8%
- No messages skipped

The model also showed higher prediction confidence overall, with 58% of test messages classified with high certainty (vs. 41% in Q1). The representation shifted: previously ambiguous tokens like "code" and "award" became more confidently associated with scams.

Iterative retraining was attempted and led to labelling 50% of unlabelled data, retraining, and labelling the rest yielded marginal improvements (+0.3% accuracy), suggesting diminishing returns after the first round.

Overall, label propagation was effective to improve accuracy and confidence. Limiting to high-confidence labels avoided label noise. This confirms that Naïve Bayes can benefit from semi-supervised learning even with a simple modification.

Supervised model evaluation

Compared to the supervised model (Q1), the semi-supervised model (Q3) achieved higher accuracy (+3.2%) and better balance between precision and recall. This suggests it was better able to generalize beyond the original labelled training data.

The confusion matrix for Q3 showed:

	Predicted Non-Malicious	Predicted Scam
Actual Non-Mal.	853	41
Actual Scam	21	185

Precision and recall for scam rose to 90.5% and 89.8%, indicating that the model became both more precise and more sensitive to scam indicators.

The model's vocabulary grew slightly due to more training messages, and maintained no skipped messages due to absence of words in vocab. Confidence also increased: more messages had extreme posterior ratios, and the boundary cases became rarer. This suggests the decision boundary sharpened.

In terms of feature representation, the top predictive scam words remained similar ("prize," "claim," "code"), but their likelihood ratios increased significantly, e.g., R(prize) rose to 234.3.

This indicates stronger statistical association with the scam class. Meanwhile, predictive non-malicious words (e.g., ":)", "lor", "wat") had even lower R values, meaning they became more strongly associated with non-scam messages.