COMP30027 Assignment 1

Written Report

Name: Quoc Khang Do

Student ID: 1375531

1. Supervised model training

The prior probabilities for each class (*Table 1*) show that non-malicious messages outnumber scam messages (by around 4 times). This reflects a typical real-world imbalance as receiving a scam text is relatively uncommon.

Class	Prior probabilities
Non-malicious	0.799499
Scam	0.200501

Table 1: Prior probabilities for each class (supervised model)

The most probable words within each class (*Table 2*) are mostly words/tokens that are frequently used in everyday messaging, including punctuation. Notably, the scam class includes words such as *"call"*, *"£"*, and *"free"*, which already begin to suggest a pattern of words with malicious connotations.

Top	Non-malicious		Scam	
	Word	Likelihood $\mathbb{P}(\text{word} \mid \text{non-malicious})$	Word	Likelihood $\mathbb{P}(\text{word} \mid \text{scam})$
1	.	0.079330	.	0.056528
2	,	0.026033	!	0.024350
3	?	0.025585	,	0.023481
4	u	0.018923	call	0.020546
5	…	0.018755	£	0.013915
6	!	0.017187	free	0.010545
7	..	0.014948	/	0.009131
8	;	0.013156	2	0.008805
9	&	0.013100	&	0.008697
10	go	0.011141	?	0.008479

Table 2: Top 10 most probable words in each class (supervised model) and their likelihoods

However, the most predictive words (*Table* 3), which show how much more likely a word is to appear in one class over another, reveal a clearer distinction. Words like *"prize"*, *"claim"*, *"code"*, *"award"*, and *"won"* are strongly skewed toward the scam class, aligning with typical SMS scam tactics involving financial bait or urgency (Australian Competition and Consumer Commission, 2024). Conversely, the non-malicious class is associated with lighthearted, friendly language, such as *“ok”*, *“hope”*, and even emoticons like *“:)”*.

When comparing the top words across classes in *Table 3*, we observe that the highest ratio for the scam class (*“prize”* ~ 99) is significantly larger than the highest ratio for the non-malicious class (*“;”* ~ 61). This suggests that scam words are more class-specific than non-malicious words – meaning when these words appear they are very strong indicators of the scam class.

These results are highly reasonable and suggest that the supervised multinomial Naïve Bayes model is effective at capturing the semantics that differentiates scams from non-malicious.

Top	**Non-malicious**		**Scam**	
	Word	**Probability ratio** $\frac{\mathbb{P}(\text{word} \mid \text{non-malicious})}{\mathbb{P}(\text{word} \mid \text{scam})}$	**Word**	**Probability ratio** $\frac{\mathbb{P}(\text{word} \mid \text{scam})}{\mathbb{P}(\text{word} \mid \text{non-malicious})}$
1	;	60.512960	prize	99.028373
2	…	57.508771	tone	64.077182
3	gt	54.075411	£	49.708360
4	lt	53.560408	select	46.601587
5	:)	47.895364	claim	45.954343
6	ü	31.930243	paytm	36.892923
7	lor	28.840219	code	34.951190
8	hope	24.720188	award	32.038591
9	ok	24.720188	won	31.067725
10	d	21.115161	18	29.125992

Table 3: Top 10 most strongly predictive words in each class (supervised model)

2. Supervised model evaluation

The supervised classifier performed well on the test set (*Table 4*), achieving an overall accuracy of 97.50%. For the scam class, both precision (92.68%) and recall (95.00%) were high, indicating the model effectively identified most scam messages while minimizing false positives. Recall is especially important in scam detection, as being too selective and failing to catch scam messages (false negatives) can have greater consequences than occasionally flagging a non-malicious message. The non-malicious class had slightly higher precision and recall, suggesting it was easier to classify than the scam class – although this is expected due to its greater representation in training (*Table 1*).

Metric	**Non-malicious (-)**	**Scam (+)**
Precision	0.9874	0.9268
Recall	0.9812	0.9500
Overall accuracy	0.9750	

Table 4: Evaluation metrics on test data (supervised model)

Table 5 shows that only 1.632% of words in the test set were out-of-vocabulary. Despite 14.2% of instances including at least one out-of-vocabulary word, the model was still able to classify all of them (no skipped instances), suggesting that most input tokens were already seen in training. This allows the model to compute more meaningful likelihoods for the test instances, instead of discarding them.

% of out-of-vocabulary words in test dataset	1.632%
% of test instances containing out-of-vocabulary words	14.2%
Number of skipped (unclassified) instances	0

Table 5: Analysis of out-of-vocabulary words in test dataset

Confidence scores (*Table 6*) provide further insight. High-confidence scam predictions (low ratios) featured typical bait language such as *"£"*, *"holiday"*, *"call"*, and *"prize"*, whereas confident non-malicious instances (high ratios) often contained more informal words, text abbreviations, and occasionally typos – features common in natural conversation. Low-confidence predictions were generally shorter, giving fewer cues for the model to confidently classify them. They also contained ambiguous words like *“call”*, *“reply”*, *“order”* which could easily be in both classes.

Instances classified as…	**Preprocessed Text**	**Confidence Ratio** $\frac{\mathbb{P}(\text{non-malicious} \mid \text{instance})}{\mathbb{P}(\text{scam} \mid \text{instance})}$
Scam with high confidence	. 4 + call £ - * holiday & urgent 18 t landline 150ppm cash cs await collection po box sae complimentary 10,000 ibiza	7.403295e-21
	. 3 4 + ! call : £ offer * holiday & urgent 18 t landline 150ppm cash cs await collection po box sae tenerife 10,000	7.791059e-21
	. . . , please order text call / : customer tone number [[service mobile]] colour colour thanks ringtone reference charge 4.50 arrive = red x49 09065989182	8.736389e-21
Non-malicious with high confidence	time : rs. transaction number & & & & & & & & & & ; ; ; ; ; ; ; ; ; ; lt lt lt lt lt # # # gt gt gt gt gt credit account reference decimal	9.193742e+37
	? ? ? ? .. .. 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 d d d lift bt real dat h girlfrnd girlfrnd moral	2.715691e+29
	. every & & & & & & ; ; ; ; ; ; lt lt lt # # # gt gt gt big hr	3.191048e+25
Low confidence	. call dear	1.014611
	. reply glad	1.041559
	. . tell return re order	0.928124

Table 6: Top 3 confident instance for each class, and top 3 low confidence instances overall (supervised model)

3. Extending the model with semi-supervised training

To extend the Naïve Bayes classifier to a semi-supervised model, a label propagation strategy was adopted (option 1). In this approach, the supervised model from part 1 was used to first classify the unlabelled dataset. These predicted labels were treated as "ground truth" and incorporated into an extended dataset for retraining.

To improve on this baseline method, two key hyperparameters were explored:

- k: only retrain the model on the top-k most confident label-propagated instances. The rationale was that restricting retraining to high-confidence predictions would mitigate noise or incorrect labels.
- $iterations$: the number of steps over which the label propagation is applied. Instead of label-propagating all at once, for each of the n iterations, consecutively train $1/n$ of the unlabelled data – allowing for gradual and more stable learning.

Grid search was used to tune these parameters on a validation set. The hyperparameter values explored were:

- $k \in \{100, 200, 300, 400, baseline\}$ – the $baseline$ value uses all the instances (no filtering).
- $iterations \in \{1, 2, 3, 4, 5\}$.

Validation results are visualized in two heatmaps. *Figure 1* (accuracy) and *Figure 2* (recall for scam class). The recall of the scam class was considered especially important due to the risk associated with false negatives mentioned earlier.

In *Figure 1*, the highest accuracy values (0.9849) were consistently observed when lowering the confidence threshold ($k = baseline$) and using a higher number of iterations ($iterations = 3,4,5$). This suggests that using all the label-propagated data (no filtering for high confidence) is effective when the supervised model is already strong, since its predictions are already reliable. Since most predictions are likely to be correct, adding more data reinforces the trained model rather than introducing harmful noise (Ferreira, Dórea, & Lee, 2023). Additionally, using more iterations likely improved performance by allowing the model to integrate label-propagated data gradually, rather than learning from a large, potentially noisy dataset all at once.

In *Figure 2*, recall for the scam class steadily increased with the number of iterations, regardless of k – at $iteration = 5$, all values reached 0.9583, even for smaller k. Whereas using fewer iterations resulted in lower recall (as low as 0.9167) for most hyperparameter combinations. This indicates that iteration count, rather than the quantity of data alone, plays a more significant role in improving recall.

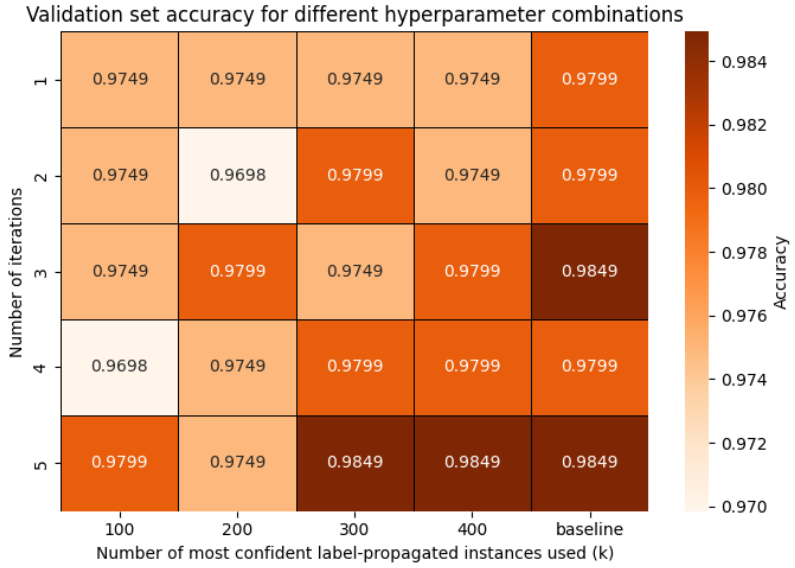

Figure 1: Heatmap showing accuracy of validation set for different hyper parameter combinations

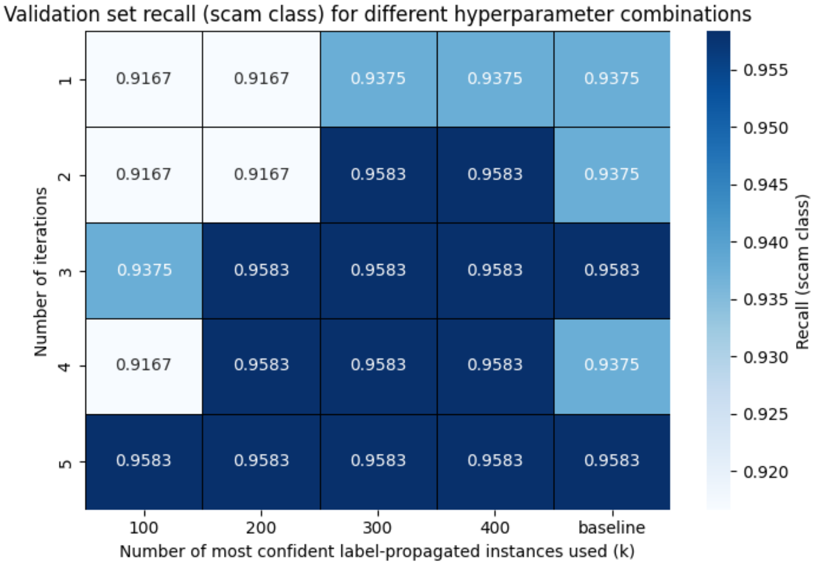

Figure 2: Heatmap showing recall (scam class) of validation set for different hyper parameter combinations

Based on these results, the optimal configuration selected was $k = baseline$, and $iterations = 5$, as an increasing number of iterations benefited both accuracy and recall, and not filtering with top-k confident instances benefited accuracy without affecting recall.

It's important to note that hyperparameter tuning and model selection were performed using a 90-10 test-validation single random holdout split. While efficient and easy to implement, this method can be sensitive to the data split and/or the random seed used to split the data. A more robust alternative would be to perform k-fold cross-validation, however this was not used due to it being more computationally expensive.

4. Semi-supervised model evaluation

The semi-supervised model outperformed the original supervised model across all key metrics (*Table 7*). Accuracy increased from 97.50% to 97.80%. More notably, recall and precision for the scam class rose from 92.68% to 93.63%, and 95.00% to 95.50%, respectively. These gains reflect an improvement in correctly identifying scams while reducing false positives. However, the improvement was relatively insignificant, likely due the supervised already being very strong, leaving little room for improvement.

Metric	Non-malicious (-)	Scam (+)
Precision	0.9887	0.9363
Recall	0.9838	0.9550
Overall accuracy	0.9780	

Table 7: Evaluation metrics on test data (semi-supervised model)

In terms of representation, the distribution of confidence ratios has shifted after semi-supervised training (*Figure 3* and *Figure 4*). The median log-confidence ratio increased from 6.7668 to 7.7430, and variation increased (IQR from 9.9249 to 11.4823). Since the confidence ratio is defined as the posterior of the non-malicious class over the posterior of the scam class, so the higher values imply the model was more decisive when classifying instances in the non-malicious class compared to the scam class. This is again likely due to exposure of more non-malicious instances within the label-propagated data.

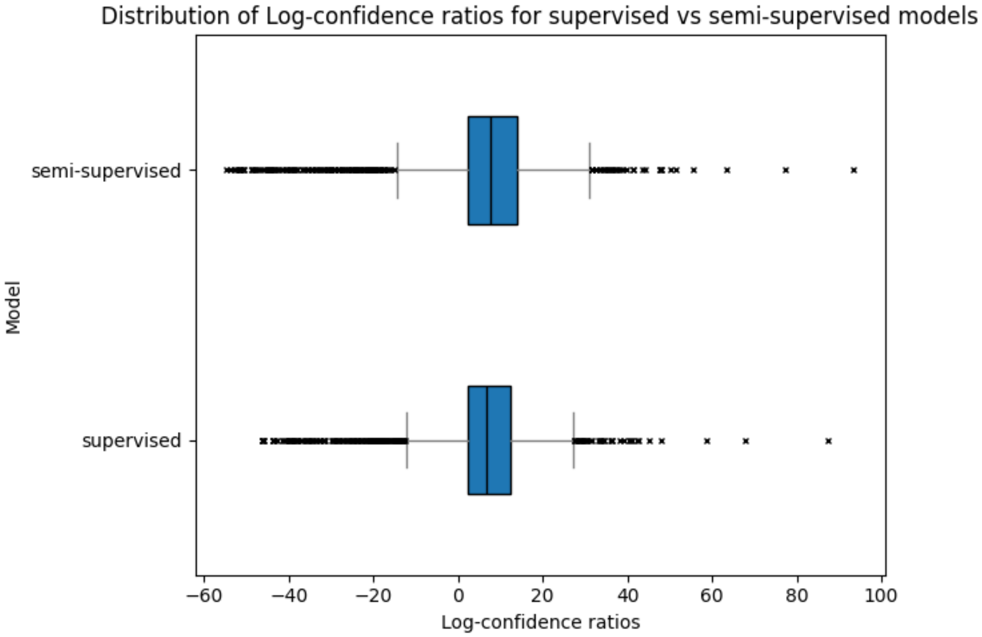

Figure 3: Boxplots showing the distributions of log-confidence ratios for supervised vs. semi-supervised model

	Supervised	**Semi-supervised**
Median	6.7668	7.7430
IQR	9.9249	11.4823

Table 8: Summary statistics of the distributions of log-confidence ratios for supervised vs. semi-supervised model

Despite the change in distribution of confidence ratios mentioned above, the most predictive words in each class remained largely unchanged between the supervised vs. semi-supervised models (*Table 9*), suggesting the model's core representation of each class has remained consistent. However, the associated probability ratios for each class increased significantly, indicating the model has become more confident that these words are highly distinctive to their respective classes.

Top	Non-malicious		Scam	
	Word	**Probability ratio** $\frac{\mathbb{P}(\text{word} \mid \text{non-malicious})}{\mathbb{P}(\text{word} \mid \text{scam})}$	**Word**	**Probability ratio** $\frac{\mathbb{P}(\text{word} \mid \text{scam})}{\mathbb{P}(\text{word} \mid \text{non-malicious})}$
1	gt	92.459497	prize	233.356235
2	lt	92.459497	tone	132.771651
3	:)	83.511803	claim	95.890637
4	;	71.747244	award	78.455975
5	ü	59.651288	code	76.444284
6	…	50.774608	guaranteed	74.432592
7	lor	48.715219	£	72.420900
8	da	46.229748	paytm	66.385825
9	:-)	35.790773	>	60.350750
10	wat	32.311114	ringtone	60.350750

Table 9: Top 10 most strongly predictive words in each class (semi-supervised model)

References

Australian Competition and Consumer Commission. (2024, August 15). *Text or SMS scams*. Retrieved April, 2025 from Scamwatch: https://www.scamwatch.gov.au/types-of-scams/text-or-sms-scams

Ferreira, R. E., Dórea, J. R., & Lee, Y. (2023). Using pseudo-labeling to improve performance of deep neural networks for animal identification. *Sci. Rep. 13, 13875*.