Spoken and Natural Language Understanding Practical Assignment 2

Submitted by Jiahui Dai

April 26, 2025

Accompanying material: Assignment_2.ipynb

1 SMS Spam Detection

1.1 Data Preprocessing Steps

Steps:

1. Tokenisation: breaking down a text corpus into individual tokens
2. Stop words removal: to remove words that are relatively common and uninformative
3. Stemming and Lemmatization:
   (a) Stemming: the transformation a word into its root form
   (b) Lemmatization: to obtain the grammatically correct forms of words
4. Data Splitting: Random 80-20 data split into training and test data

1.2 Experimental Design and Methods

1. Split the data into training and test data
2. Building of two different vectorisors based on its features:
   (a) **Bag-of-Words** (BoW)
   BoW treats each document as a collection of words, disregarding grammar and word order but keeping track of word frequency.
   (b) **TF-IDF**
   TF-IDF (Term Frequency-Inverse Document Frequency) is used to evaluate how important a word is to a document in a collection or corpus. It helps to weigh terms based on their frequency within a document and their rarity across all documents.

$$TF(t,d) = \frac{\#\text{ terms of } t \text{ appears in document } d}{\text{Total } \#\text{ of terms in document } d} \tag{1}$$

$$IDF(t) = \log\left(\frac{\text{Total } \#\text{ of documents}}{\#\text{ of documents containing } t}\right) \tag{2}$$

$$TF - IDF(t,d) = TF(t,d) \cdot IDF(t) \tag{3}$$

Term Frequency (TF) measures how frequently a term occurs in a document.
Inverse Document Frequency (IDF) measures the importance of a term across all documents in the corpus.
TF-IDF is the product of TF and IDF. It means that

- If a term appears frequently in a document but also in many documents, the TF will be high but the IDF will be low, lowering the overall score.

- If a term appears frequently in one document but is rare across all documents, the score will be high, emphasizing the importance of that term for the document.

3. Trained using Multinomial Naive Bayes form scikit-learn
4. Compare performance on the same train/test split.

1.3 Hyperparameters

Laplace smoothing is used to find tune the model to obtain the best/optimal evaluation metrics. (Default: $\alpha = 1.0$)

1.4 Evaluation Metric [1]

- TP: A spam message is correctly classified as spam
- TN: A ham (non-spam) message is correctly classified as ham
- FP: A ham message is incorrectly classfied as spam
- FN: A spam message is incorrectly classfied as ham

Accuracy is the proportion of all classifications that were correct, whether positive or negative.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Recall or **true positive rate (TPR)** is the proportion of all actual positives that were classified correctly as positives.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

Precision is the proportion of all the model's positive classifications that are actually positive.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

F_β **Score (F1 Score)** provides a balanced measure of a model's precision (P) and recall (R).

$$F_\beta = (1 + \beta^2)\frac{P \cdot R}{\beta^2 \cdot P + R} \tag{7}$$

1.5 Findings

Performance metrics is shown in Figure 1

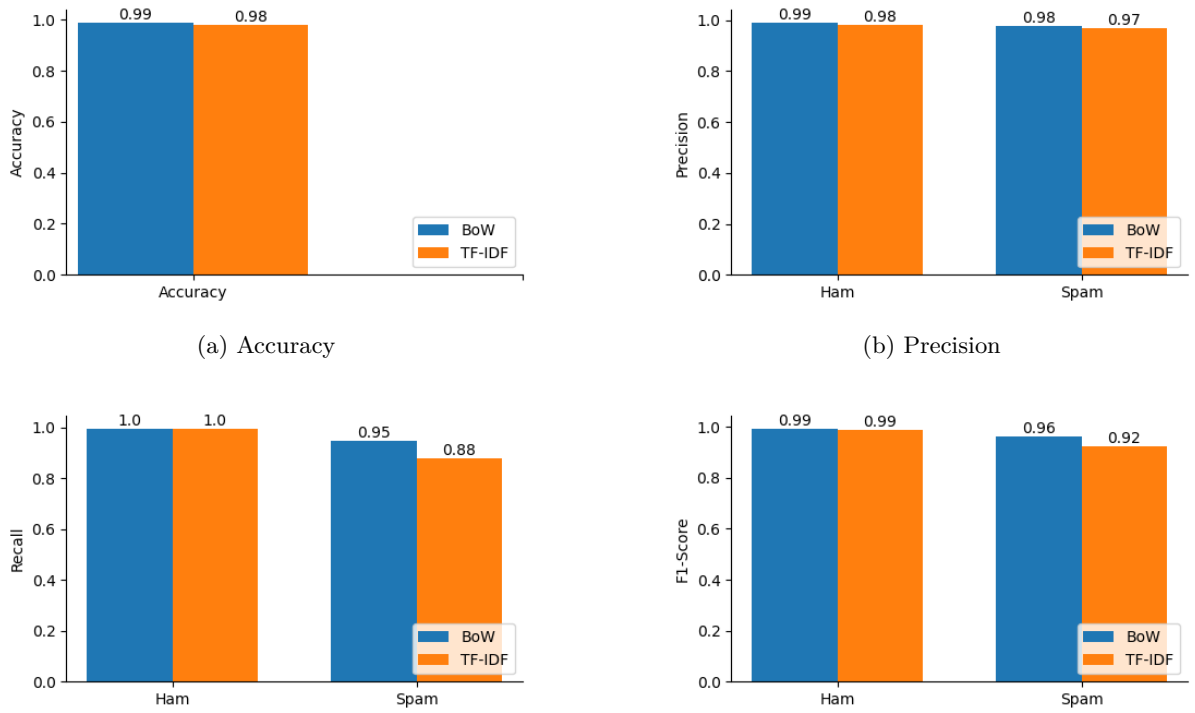
Accuracy In Figure 1a, both models have high accuracies, which suggests that both models are able to correctly predict well and accurately the labels for the test data, whether 'Ham' or 'Spam'.

Precision In Figure 1b, both models have high precision for both 'Ham' and 'Spam' labels, which suggests that both models are able to correctly well and predict precisely the labels for the test data, on both 'Ham' and 'Spam'.

Recall In Figure 1c, both models have 1.0 score for recall on 'Ham' and 'Spam', which suggests that both models were able to correctly identify label 'Ham' and 'Spam' test data. TF-IDF model still have lower recall for 'Spam' compared to BoW model, suggesting that it is not as recallative as BoW model.

F1 Score In Figure 1d, both models have high F1-score on both 'Ham' and 'Spam', which suggests a good balance between precision and recall for 'Ham' and 'Spam'.

Higher score on 'Ham' compared to 'Spam' There is a higher score on 'Ham' than 'Spam' is highly due to the imbalanced data of 'Ham' and 'Spam'. There is a total of 4825 'Ham' labelled documents and 747 'Spam' labelled documents.



(a) Accuracy

(b) Precision

(c) Recall

(d) F1 Score

Figure 1: Performance Metrics between BoW and TF-IDF models

Overall better performance of BoW Model The BoW model has a higher corresponding metric score compared to TF-IDF model, with slightly better overall identification of 'Spam' messages. The TF-IDF model has a slightly lower precision and lower recall for 'Spam' category (which also lowers the F1 score).

1.6 Drawbacks

- Manual vectorization is slower than optimized libraries.
- No deep contextual understanding (can't catch subtle spam phrasing).

1.7 Potential Improvements

- Use higher n-grams for better context (Currently only unigram)

2 Search Engine

2.1 Construct Document Graph

Cosine similarity measures the angle between two vectors (i.e. TF-IDF vectors of SMS messages). It ranges from 0.0 (where messages with no shared words) to 1.0 (where messages are identical in terms of content).

Figure 2 plots the frequency of cosine similarity value between SMS messages. Here we see a histogram that is right skewed towards low similarities (0.0 - 0.2), which suggests that most messages are unrelated, and that the dataset is diverse with many topics.

The similarlity threshold can be set on graph construction which affects which nodes (messages) are connected. In Figure 3, the outer ring shows the nodes of messages that have cosine similarity below the threshold and the core shows the connected nodes with edges. A high threshold (e.g. 0.8) will only connect highly similar messages, shown in Figure 3b, where a low threshold (e.g. 0.2), more messages will be linked and creating a densely connected graph, shown in Figure 3a.

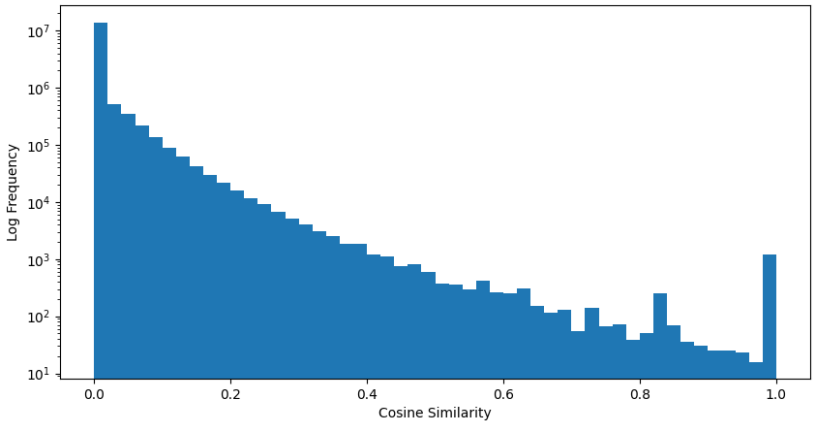


Figure 2: Histogram of Cosine Similarities

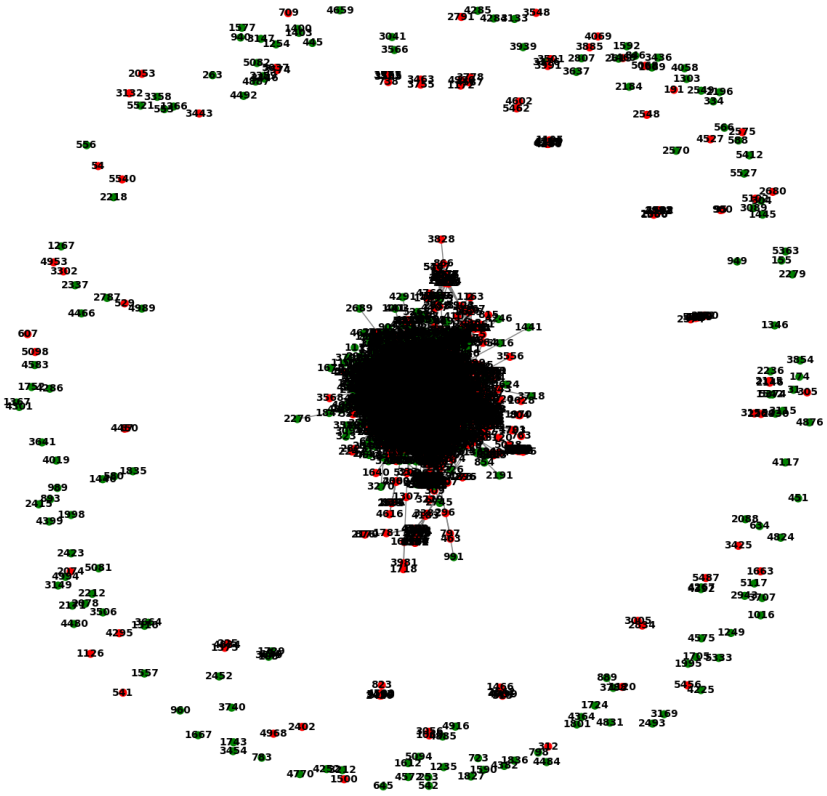
2.2 PageRank Algorithm

PageRank is used to rank pages (nodes) based on their importance. As the edges are connected to the message nodes based on cosine similarity, only with similarity greater than threshold. This means that a message node becomes important if many important or similar nodes point to it.

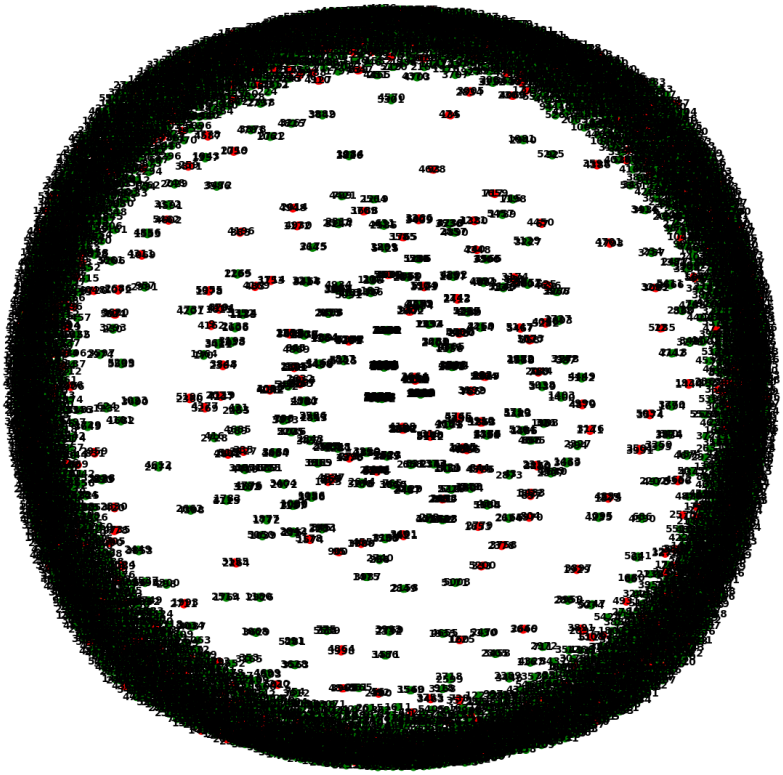
PageRank	Label	Message
0.00017946877243223293	Ham	Beerage?
0.00017946877243223293	Ham	645
0.00017946877243223293	Ham	* Will be september by then!
0.00017946877243223293	Ham	O was not into fps then.
0.00017946877243223293	Ham	I pocked you up there before
0.00017946877243223293	Ham	Or just do that 6times
0.0001794687724319843	Ham	What's the significance?
0.0001794687724319843	Ham	Annoying isn't it.
0.0001794687724319843	Ham	Which channel:-):-):):-).
0.00017946877243180592	Ham	And how's your husband.
0.0001794687724314241	Ham	S:)no competition for him.
0.00017946877243132285	Ham	ALRITE
0.00017946877243132285	Ham	So how's the weather over there?
0.00017946877243132285	Ham	It'll be tough, but I'll do what I have to
0.00017946877243122904	Ham	Then why you not responding
0.00017946877243098068	Ham	I have no idea where you are
0.00017946877243090606	Ham	Then we gotta do it after that
0.00017946877243090606	Ham	What not under standing.
0.0001794687724308693	Ham	hanks lotsly!
0.0001794687724308693	Ham	Those ducking chinchillas

Table 1: Top 20 Messages based on PageRank with threshold of 0.8

In Table 1, these messages have high PageRank score, which means that they are well-connected in terms of context. The top 20 messages are labelled 'Ham', which further contribute to contextually relavent content. They have a common topics of casual conversation and personal updates and are usually of shorter length and straight to the point in conversations.



(a) Threshold: 0.2



(b) Threshold: 0.8

Figure 3: Document Similarity Graph (Spam vs Ham)

PageRank	Label	Message
0.00009893988409080199	Spam	URGENT! Your mobile No 077xxx WON a £2,000 Bonus Caller Prize on 02/06/03! This is the 2nd attempt to reach YOU! Call 090663622
0.00009631903682978389	Spam	Congrats! 2 mobile 3G Videophones R yours. call 09063458130 now! videochat wid your mates, play java games, Dload polyPH music, noli
0.00009467170008725136	Spam	URGENT! Your Mobile number has been awarded with a £2000 prize GUARANTEED. Call 09058094455 from land line. Claim 3030. Vali
0.00009467170008725136	Spam	URGENT! Your Mobile number has been awarded with a £2000 prize GUARANTEED. Call 09058094454 from land line. Claim 3030. Vali
0.00009436915435003342	Ham	Yup i'm free...
0.00009019489256296639	Spam	Congratulations ur awarded either £500 of CD gift vouchers
0.00008956599869019722	Spam	important information 4 orange user 0789xxxxxxx. today is your lucky day!2find out why log onto http://www.urawinner.com THERE'S A
0.00008567559136037 02	Spam	8007 FREE for 1st week! No1 Nokia tone 4 ur mob every week just txt NOKIA to 8007 Get txting and tell ur mates www.getzed.co.uk PO
0.00008547593394616635	Ham	Once free call me sir.
0.00006058484083945277	Ham	I'm in a meeting, call me later at
0.00006058484083945277	Ham	I'm in a meeting, call me later at
0.00006058484083945277	Ham	I'm in a meeting, call me later at
0.00004950692122048504	Ham	Sorry, I'll call you later. I am in meeting sir.
0.00002692031586503949	Ham	What you doing?how are you?
0.00002692031586503949	Ham	Where @
0.00002692031586503949	Ham	Can a not?
0.00002692031586503949	Ham	:)
0.00002692031586503949	Ham	What you doing?how are you?
0.00002692031586503949	Ham	:(but your not here....
0.00002692031586503949	Ham	:-) :-)

Table 2: Bottom 20 Messages based on PageRank with threshold of 0.8 (Descending order)

In Table 2, these messages have low PageRank score, which means that they are less important. The bottom 20 messages have some messages labelled as 'Spam', which are typically standalone and contextually irrelavent, and therefore ranks lower and do not contribute to the network relavence. They have a common topics of promotions, unsolicited offers, mobile phone scams and are often longer in length with many keywords to capture attention: URGENT, WON, CALL, etc.

3 Additional Experiments

References

[1] Google Developers. *Accuracy, Precision, and Recall*. Accessed: 2025-04-16. n.d. URL: https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall.

[2] Sebastian Raschka. "Naive Bayes and Text Classification". In: *arXiv preprint arXiv:1410.5329* (2014). URL: https://arxiv.org/pdf/1410.5329.