MLRD 2022 Paper 3

7a) i) System A

Prediction \ Actual	T	F
T	5	1
F	9	119,985

System B

Prediction \ Actual	T	F
T	12	20
F	2	119,966

FN : 11
TN : 239951
FP : 21
TP : 17

ii) $A_A = \frac{119,985+5}{120,000} \simeq 0.9999$

$A_B = \frac{119,966+12}{120,000} \simeq 0.9998$

iii) $\text{Precision} = \frac{TP}{TP+FP}$ $\quad$ $\text{Recall} = \frac{TP}{TP+FN}$

$P_A = \frac{5}{5+1} \approx 0.83$ $\quad$ $P_B = \frac{12}{12+20} \simeq 0.38$

$R_A = \frac{5}{5+9} \simeq 0.36$ $\quad$ $R_B = \frac{12}{12+2} \simeq 0.86$

b) i) Null hypothesis: System A is no better than System B.

N is the total number of classified documents. (120,000)
p is the probability of system B beating system A ($p=0.5$).

Count number of times A beats B, B beats A, and they predict same. Use the formula to find if probability is less than some α (usually 0.05 or 0.01). If it is then reject null hypothesis

ii) A beats B (plus) = 1 + 20 = 21
B beats A (minus) = 8 + 0 = 8
Draw (null) = 120,000 - 29 = 119,971

iii) No since the difference in accuracy is very small and the sign test shows A doesn't beat B that much more than B beats A.

iv) Comparing precision and recall:

$P_A = 0.83 > P_B = 0.38$

$R_A = 0.36 < R_B = 0.86$

so A has much higher precision but B has much higher recall.

iv) We can calculate the F-measure for each system as well:

$$F = \frac{2 \cdot P \cdot R}{P + R}, \quad F_A \approx 0.50, \quad F_B \approx 0.53$$

So B slightly better, but not much. The bigger difference depends if we want higher precision or recall specifically.

We can't do a sign test since we can't see for each document which system has higher precision or recall. These can only be calculated over all documents.

8a) Transition probs:

[L→L: 0, L→M: 1, L→H: 0,
M→L: 0, M→M: $\frac{1}{2}$, M→H: $\frac{1}{2}$,
H→L: 0, H→M: $\frac{1}{3}$, H→H: $\frac{2}{3}$]

Emmision probs

[(L, +): 1, (L, ++): 0, (L, +++): 0,
(M, +): 0, (M, ++): 1, (M, +++): 0,
(H, +): 0, (H, ++): $\frac{2}{3}$, (H, +++): $\frac{1}{3}$]

b) 1. For a first order MM the probability of a state given all previous states is equal to the pobability of the state given the previous one state.

2. The observed feature depends only on the hidden state.

Assumption 1 isn't that appropiate as the infection number could depend on multiple previous states. For example if it's been H for 3 or 4 timesteps we might expect it to lower since most people will have developed immunity.
Assumption 2 is appropiate as it's likely the positivity rate depends on the number of currently infected people.

c) $\delta_j(t) = \max_{1 \le i \le 3} [a_{ij} \; \delta_i(t-1) \; b_j(O_t)]$

L: 1, M: 2, H: 3

t = 8 ~~b_j (...)~~

$\delta_L(8) = \max$ ~~[0~~

$\delta_M(8) = \frac{1}{2} \times 1 \times 0 = 0$

$\delta_H(8) = \frac{1}{2} \times 1 \times \frac{1}{3} = \frac{1}{3}$

~~So M at 8~~

$\delta_L(9) = \max \begin{bmatrix} 0 \\ \frac{1}{2} \times 0 \times 0 \\ \frac{1}{2} \times \frac{1}{3} \times 0 \end{bmatrix} = 0$

$\delta_M(9) = \max \begin{bmatrix} 1 \times 0 \times 1 \\ \frac{1}{2} \times 0 \times 1 \\ \frac{1}{3} \times \frac{1}{3} \times 1 \end{bmatrix} = \frac{1}{9}$

$\delta_H(9) = \max \begin{bmatrix} 0 \\ \frac{1}{2} \times 0 \times \frac{2}{3} \\ \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3} \end{bmatrix} = \frac{4}{27}$

c) $\delta_L(10) = 0$

$$\delta_M(10) = \max \begin{bmatrix} 1 \times 0 \times 1 \\ \frac{1}{2} \times \frac{1}{18} \times 1 \\ \frac{1}{3} \times \frac{4}{216} \times 1 \end{bmatrix} = \frac{1}{18}$$

$$\delta_H(10) = \max \begin{bmatrix} 0 \\ \frac{1}{2} \times \frac{1}{9} \times 1 \\ \frac{2}{3} \times \frac{4}{216} \times 1 \end{bmatrix} = \frac{8}{81}$$

So 10 9 8
H H H

d) 1. The first markov assumption from (b) is not realistic for modelling infections
2. It doesn't take into account Cots of other important factors such as changes in behaviour, social distancing rule changes etc.

9a) $P_{NB}(c) = \underset{c \in (a,s)}{\operatorname{argmax}} \left[\log P(c) + \sum_{w_i \in V} \log P(w_i | c) \right]$

$$P(c) = \frac{\text{count}(c)}{\text{count}(\text{all documents})}$$

$$P(w_i | c) = \frac{\text{count}(w_i \text{ in } c)}{\text{count}(\text{all words in } c)}$$

b) 4500 used for training 90 for each brand.
500 for testing, 10 for each brand.

This allows as high amount for training while still leaving enough for testing.

c) i) Attacked: Terrible, avoid, not

Safe: Awesome, Fantastic, Support

Terrible and avoid, awesome, fantastic should generalise well as they are strong words with one main use.
Support and not probably won't generalise well as they aren't strong sentiment indicators and can be used for good or bad e.g. 'it's not great' or 'it's not bad at all'

c) (ii) Extract words following not as a group pair with not to better get the context and sentiment.

(Maybe 3?)

Tag parts of speech as verbs, adjectives etc. to prevent confusion if a word has two meanings.

d) 1. Retrain often with new data to allow it to know any new words that are good indicators

2. Try to use posts in context of ~~surrounding~~ posts to get better understanding.

3. Let the classifier annotate new data to train itself.

e) Adv

- More fine-grained analysis for each post
- Might be quicker to spot individual posts that are starting campaigns against boards
- It's more flexible - can be adapted for different needs.

e) Disadv

- More computational complexity
- Isolating posts may lose contextual data
- Lose broad-level insights such as overall public perception