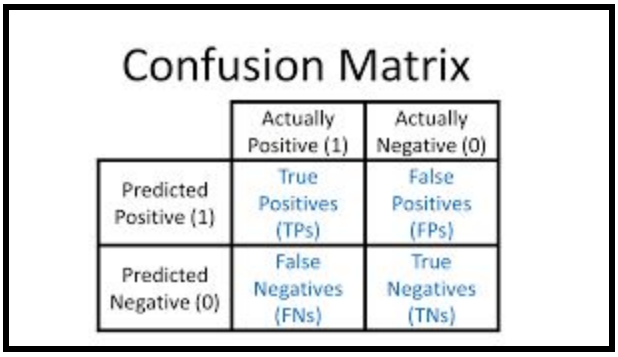
Q1.) When I vary the similarity threshold of the system from 1 – 50 I get different numbers of correct and incorrect answers, that is different numbers of True Positives (TP), False Negative (FN), False Positives (FP) and True Negatives (TN) tweets. For example, when my system correctly identifies the tweets as being about the election and it was indeed about the election, it's a True Positive. When my system says that the tweet is about the election and it is not, then I have got a False Positive. Taking this data, can you compute the Precision and Recall for the system at each threshold and identify the threshold values at which it does best, according to the F1 measure?

Confusion Matrix:

The given data consists of True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) values which represent the confusion matrix. The confusion matrix is used for performance measurement for classification problems. It is useful to find precision, recall and F1 score of the model. Following is an example of confusion matrix :-



Precision : *"how many of the selected objects were correct"*
Recall : *"how many of the objects that should have been selected were actually selected"*

Precision : True positives / (True positives + False positives)
Recall: True positives / (True positives + False negatives)

Recall expresses the ability to find all relevant instances in a dataset. Precision expresses the proportion of the data points our model says was relevant actually were relevant.

F1 measure : harmonic mean of precision and recall. It uses harmonic mean instead of arithmetic mean because it punishes extreme values. Both precision and recall have same numerators but different denominators. To get the right mean we need to use their reciprocals, thus harmonic mean.

F1 : 2 * [(precision * recall) / (precision + recall)]

Below is the precision, recall and F1 measure snippet for each threshold value -

	Threshold	TP	FN	FP	TN	Correct	Incorrect	Test set	precision	recall	F1_Measure
0	1	20	80	2	98	100	100	200	0.909091	0.20	0.327869
1	5	50	50	5	95	100	100	200	0.909091	0.50	0.645161
2	10	60	40	10	90	100	100	200	0.857143	0.60	0.705882
3	15	80	20	20	80	100	100	200	0.800000	0.80	0.800000
4	20	88	12	30	70	100	100	200	0.745763	0.88	0.807339
5	25	90	10	40	60	100	100	200	0.692308	0.90	0.782609
6	30	95	5	50	50	100	100	200	0.655172	0.95	0.775510
7	35	96	4	60	40	100	100	200	0.615385	0.96	0.750000
8	40	97	3	70	30	100	100	200	0.580838	0.97	0.726592
9	50	98	2	80	20	100	100	200	0.550562	0.98	0.705036

From the above results, F1 measure for the **threshold 20 does the best as it is highest(0.807).**
The results at threshold values 15 and 20 do better than all other scenarios because of their high F1 scores. At threshold 15, the same precision and recall is obtained. At threshold 20, the precision is less than the recall but the F1 score is the highest amongst all.

Q2.) Now, can you plot the ROC for this data?

Receiver operating characteristic curve (ROC) —> Illustrates the diagnostic ability of a binary classification system as its discrimination threshold is varied. More generally, the ROC curve is a performance measurement for binary classification at different threshold settings. The ROC is a probability curve and (Area under the curve) AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better is the model is at predicting 0s as 0s and 1s as 1s.

The ROC curve is plotted with True positive rate(TPR) against the False positive rate(FPR) where TPR is on y-axis and FPR is on the x-axis.

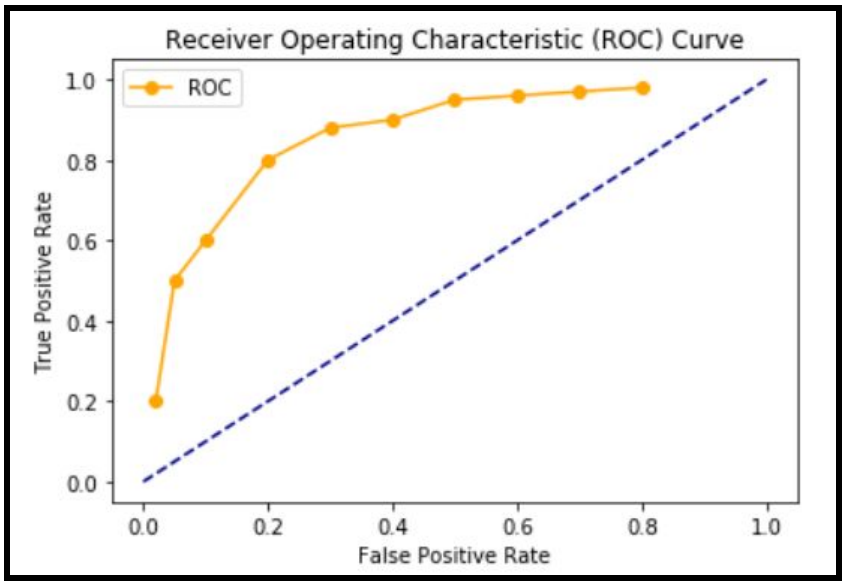
TPR : True positives / (True positives + False negatives)

FPR : False positives / (True negatives + False positives)

Following are the TPR, FPR and FNR values for each threshold -

	Threshold	TPR	FPR	FNR
0	1	0.20	0.02	0.80
1	5	0.50	0.05	0.50
2	10	0.60	0.10	0.40
3	15	0.80	0.20	0.20
4	20	0.88	0.30	0.12
5	25	0.90	0.40	0.10
6	30	0.95	0.50	0.05
7	35	0.96	0.60	0.04
8	40	0.97	0.70	0.03
9	50	0.98	0.80	0.02

Following is the ROC curve obtained -



ROC curves that are closer to the top-left corner indicate a better performance. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). More closer is the curve to the diagonal, less accurate are the results. From the curve we can infer that the ROC curve lies far from the diagonal but not very close to the top-left corner of the ROC space.

Q3.) Now, can you plot the DET curve on the same data?

Detection error tradeoff is the plot of error rates for binary classification systems, plotting the false rejection rate vs. false acceptance rate. They show the range of operating points of systems performing detection tasks as a threshold is varied to alter the miss and false alarm rates and plotted using a normal deviate scale for each axis.

DET curves have the property that if the underlying score distributions for the two types of trials are normal, the curve becomes a straight line. DET curves focus on errors more and “zoom in” to key parts of the ROC curve by using log axes. Since DET curves plot error rates, they give uniform treatment to both types of errors and they use scaling for both the axes which spreads out the plot and produces almost linear plots. They use (FPR) and False negative Rates(FNR) where FNR is given by -

FNR : False negatives / (False negatives + True positives)

Following is the DET curve obtained -