Sentimental Analysis using NLP

1. **Finding the Problem** — From the Tweet text: Identifying the Sentiments from the text
2. **Collecting Dataset** — Twitter Sentiment dataset
3. **Load & Summarize dataset**
4. **Segregating Dataset into X & Y**
5. **Removing the Special characters, Symbols with Regular Expressions**
6. **Removing Stop words and Extracting Features**
7. **Splitting Dataset into Train and Test**
8. **Load Machine Learning Algorithm**
9. **Predict with Test Dataset**
10. **Evaluating with Confusion Matrix**

Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.

Example

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

COVID – 19 - Binary prediction (Yes / No)
Total no. of Patient : 165
Real Data:
Covid Yes – 105 Patient
No Covid – 60 Patient

Our ML Predicted Data:
Covid Yes – 110 Patient
No Covid – 55 Patient

Elements

- True positives (TP) — These are cases in which we predicted yes (they have the covid), and they have the covid.
- True negatives (TN) — We predicted no, and they don't have the covid.
- False positives (FP) — We predicted yes, but they don't actually have the covid. (Also known as a "Type I error.")
- False negatives (FN) — We predicted no, but they actually have the covid. (Also known as a "Type II error.")

Calculation

Example

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

- Accuracy & Loss Calculation from Confusion Matrix
  - Accuracy: Overall correct.
  - (TP+TN)/total = (100+50)/165 = 0.91
  - Misclassification Rate: Overall wrong | Error Rate
  - (FP+FN)/total = (10+5)/165 = 0.09
- Other Calculation

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

True Positive Rate: Sensitivity or Recall
TP/actual yes = 100/105 = 0.95

False Positive Rate:
FP/actual no = 10/60 = 0.17

True Negative Rate: Specificity
TN/actual no = 50/60 = 0.83

Precision:
TP/predicted yes = 100/110 = 0.91

Prevalence :
actual yes/total = 105/165 = 0.64