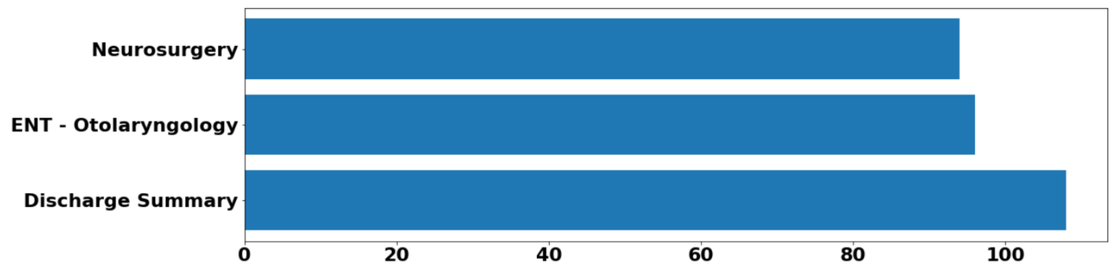
Text Classification of Medical Transcript

01 Problem

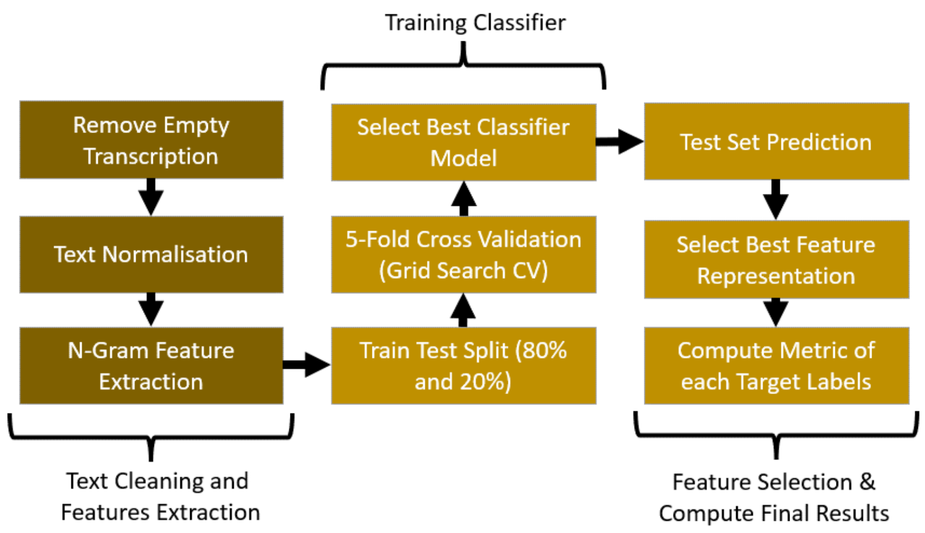
1. Healthcare workers spend **a lot of time** to identify key issues of each transcript
2. Transcripts with complex cases that lead to information **overload and delays**

02 Dataset

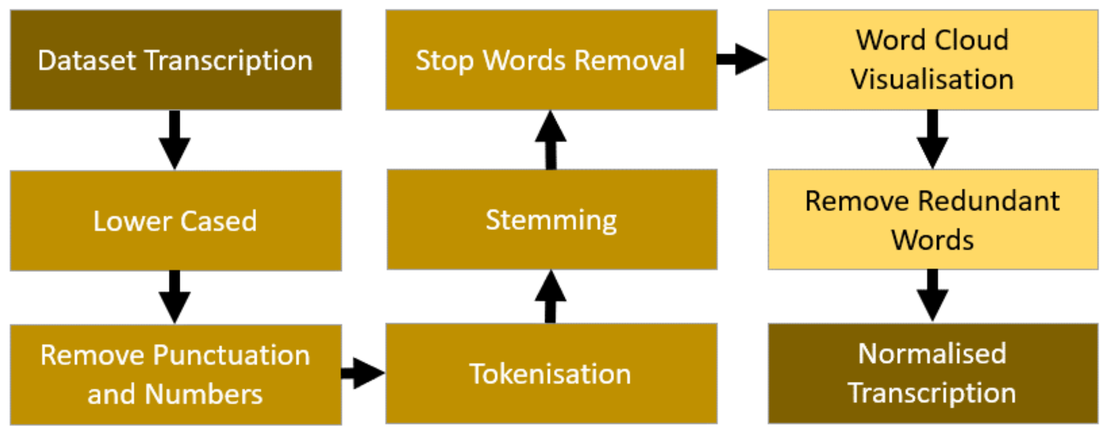
Medical transcript dataset is used, which is sourced from Kaggle and MT samples. It has 300 observations with 3 label classes



03 Solution Overview



04 Text Normalisation



-42.9% Less words after text cleaning

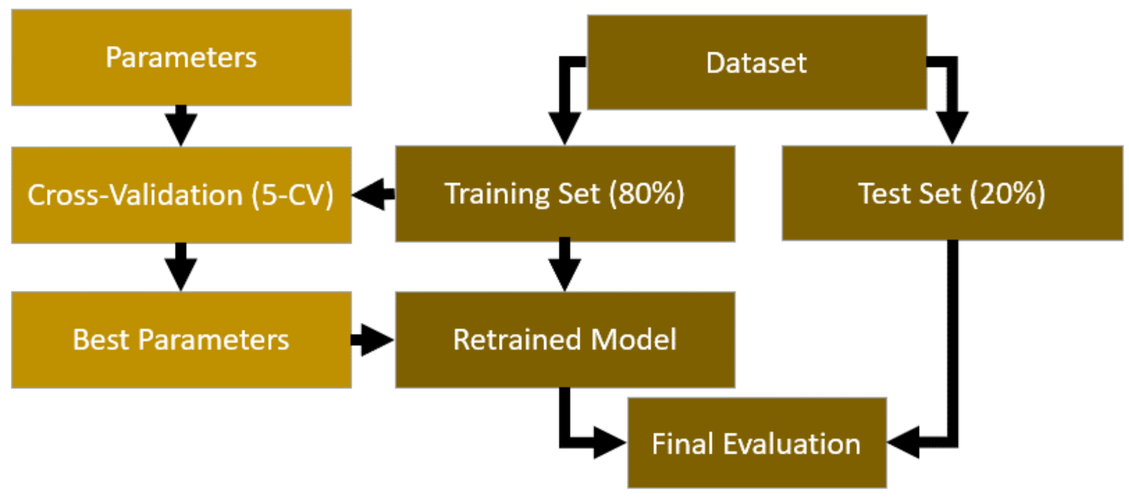
Deployed text normalisation tools in website
http://www.morris-lee.com/nlp

05 N-Gram Feature Extraction

5 N-gram Feature vectors are extracted

Feature Vectors	Lengths of array dimension	
	Rows	*Columns*
Unigram	298	7038
Unigram + Bigram	298	56270
Bigram	298	49232
Bigram + Trigram	298	114215
Trigram	298	64983

06 Experiment Setup



07 Evaluation Criteria

1. **Metric Score:** Macro F1
2. **Train-test split** 80% and 20%.
3. **5-fold cross validation** on training set to find optimised hyperparameters**.**
4. Using trained model with best parameter to make a unseen prediction.

08 Text Classifiers

1. K-Nearest Neighbour (n_neighbors = 7)
2. Decision Tree Classifier
   (max_depth = None, min_samples_split = 2)
3. Random Forest Classifier (RFC)
   (max_depth = 35, n_estimators = 16)

09 Results and Findings

Test Prediction Metric Score using RFC

Features	macro F1
Unigram	0.9336
Unigram + Bigram	0.8526
Bigram	0.8372
Bigram + Trigram	0.8499
Trigram	0.5773

10 Key Findings

1. Unigram is the best feature vector in classifying the medical transcript, obtaining best result, **0.93 macro F1.**
2. This may indicates that the adjacent word ordering less important in the classification task **in this context**.