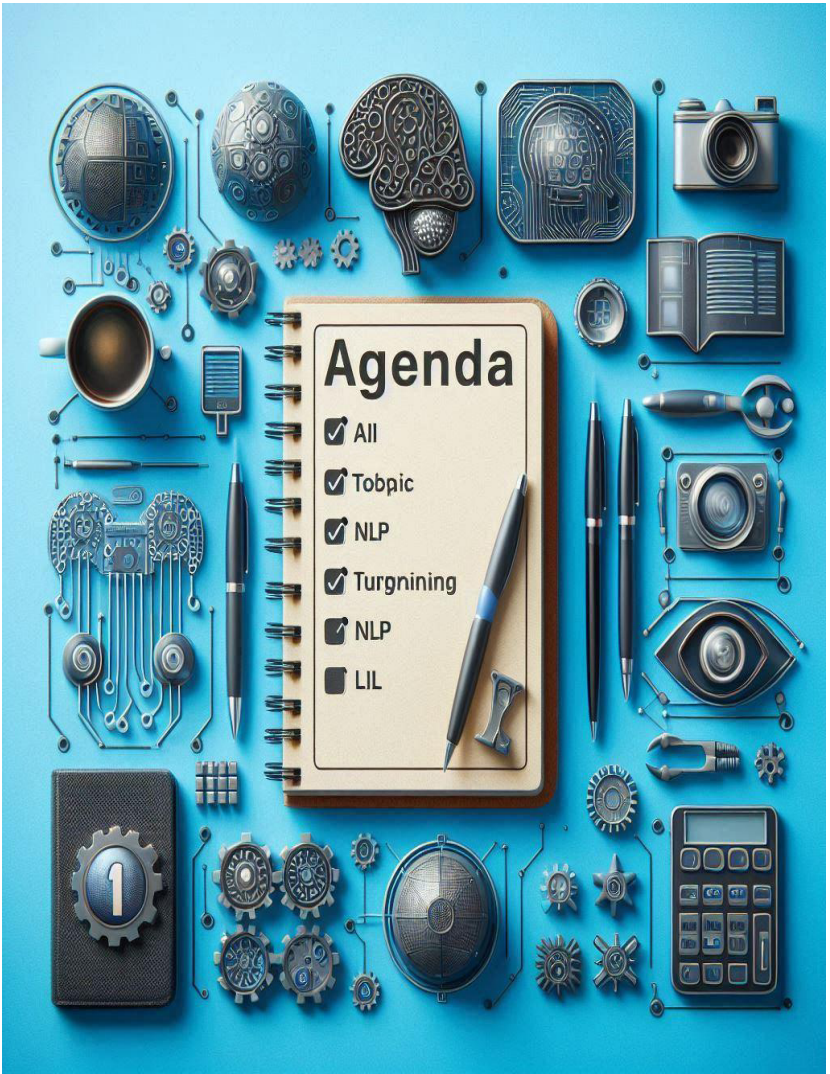
National Telecommunication Institute

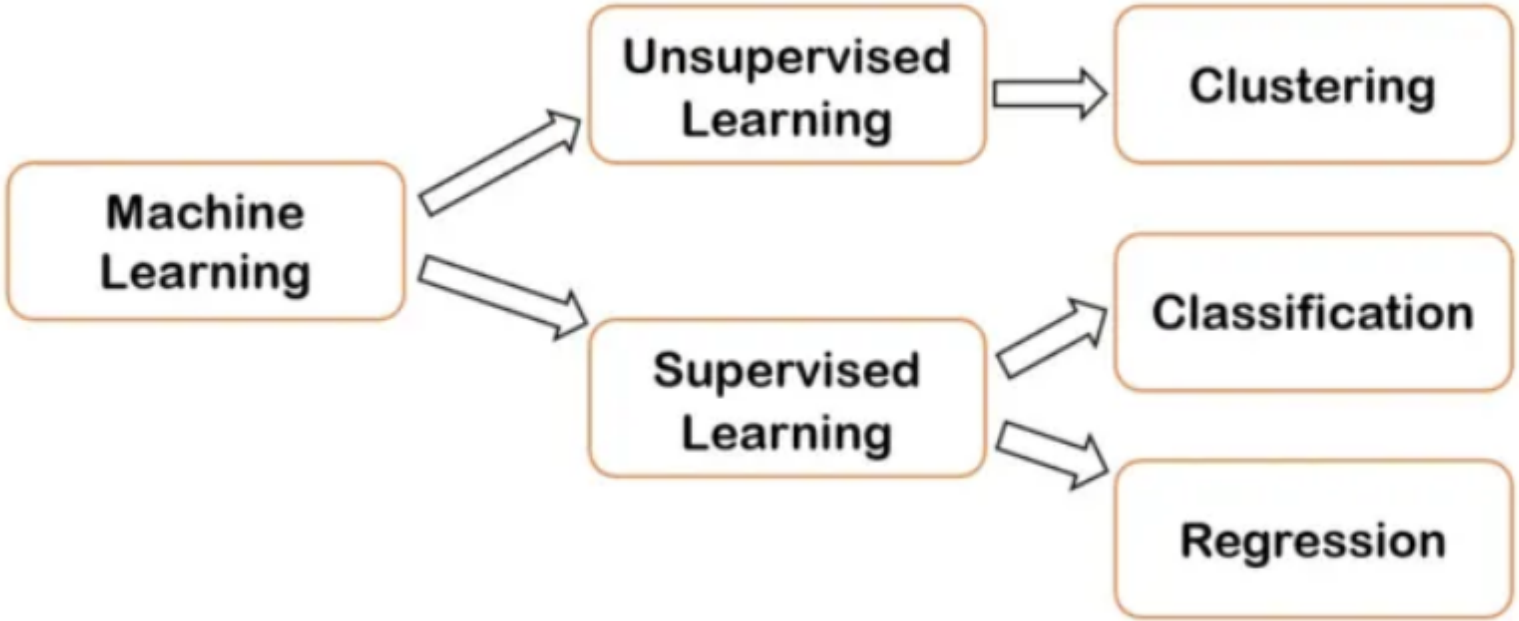
What are your expectations for the course?



Agenda

1. Revision on ML

1. What is Text Classification?
   - Binary Classification
   - Multi class Classification


Machine Learning
Unsupervised Learning
Supervised Learning
Clustering
Classification
Regression

Text Classification

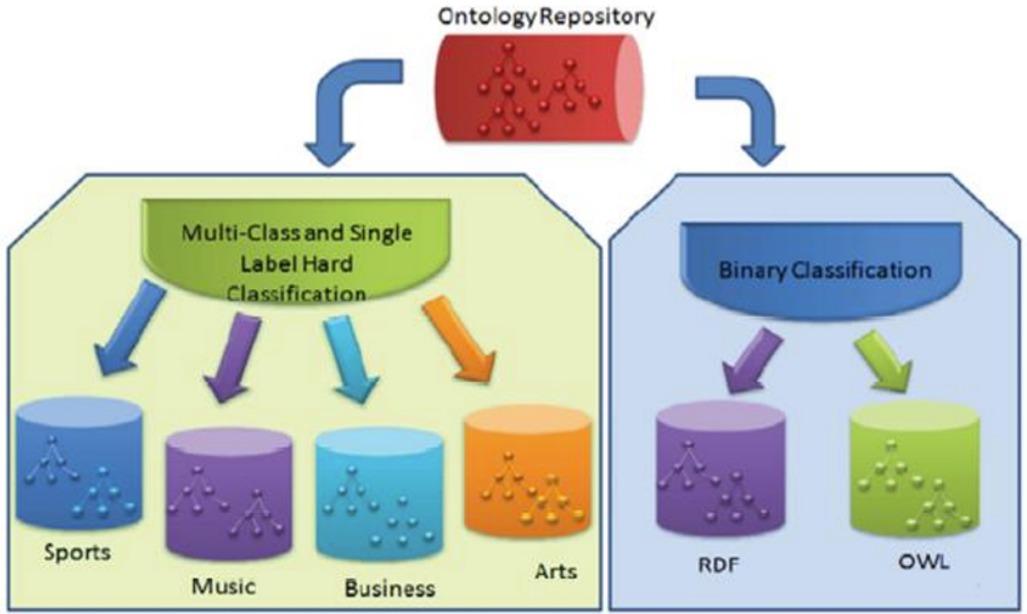
Text classification or Text Categorization is the activity of labeling natural language texts with relevant categories from a predefined set.

- Assigning subject categories, topics, or genres·
- Spam detection·
- Authorship identification·
- Age/gender identification·
- Language Identification·
- Sentiment analysis.

Text Classification

There are two kinds of classifiers:

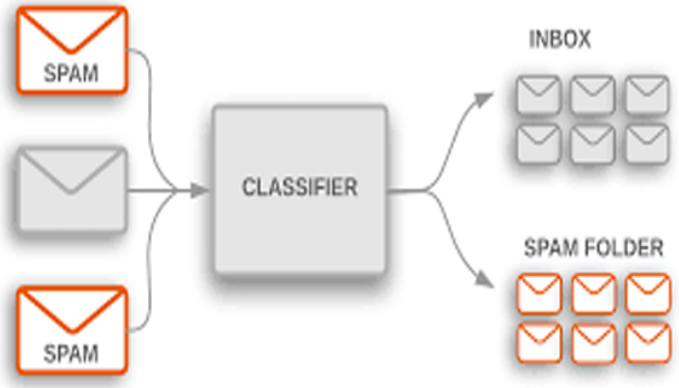
1. Binary Classifiers.
2. Multi-Classifiers



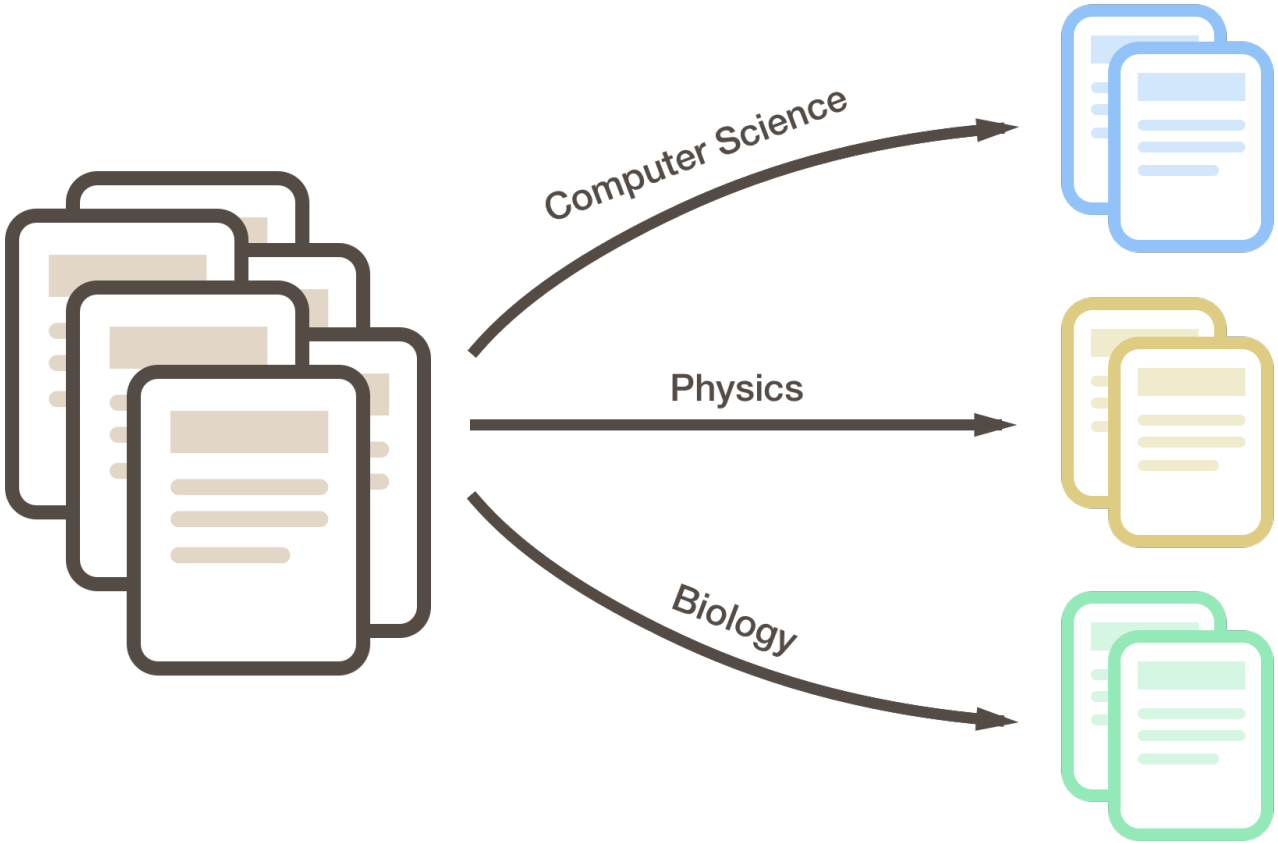
Binary Classifiers

Suppose we have a dataset of SMS messages, each labeled as 'spam' or 'ham'. Our goal is to train a model that can accurately classify new messages.

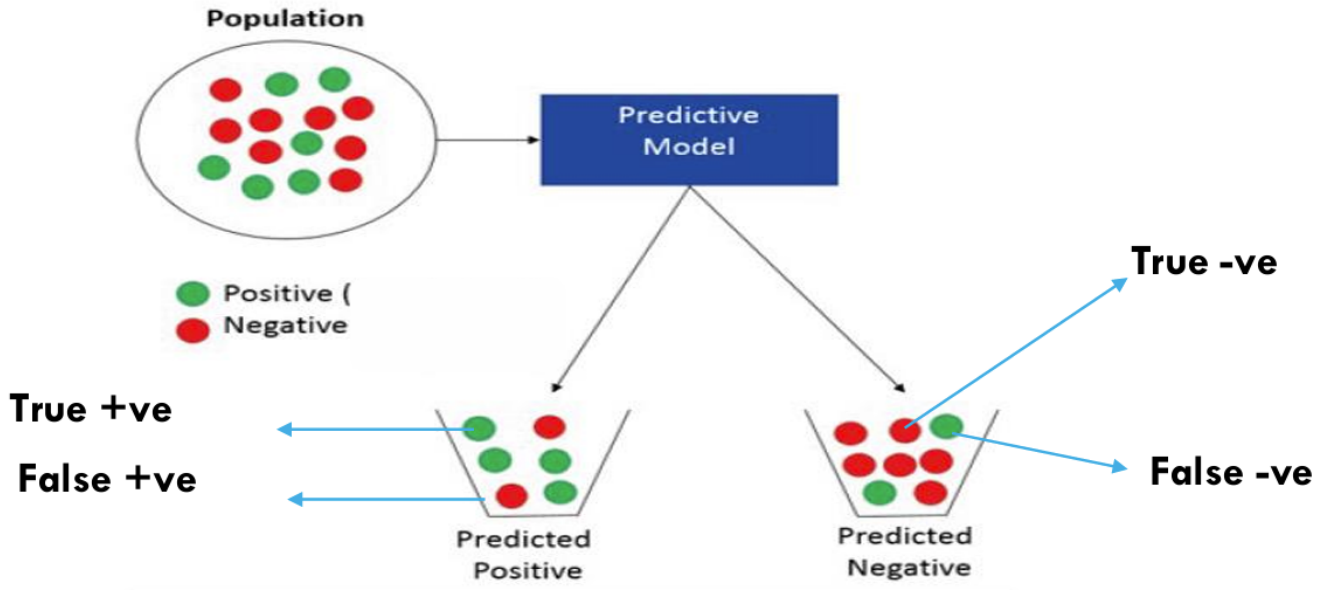
First, we preprocess the messages — removing numbers, special characters, and converting everything to lowercase. Then, we extract features using techniques like Bag-of-Words or TF-IDF ."



Multi class Classification



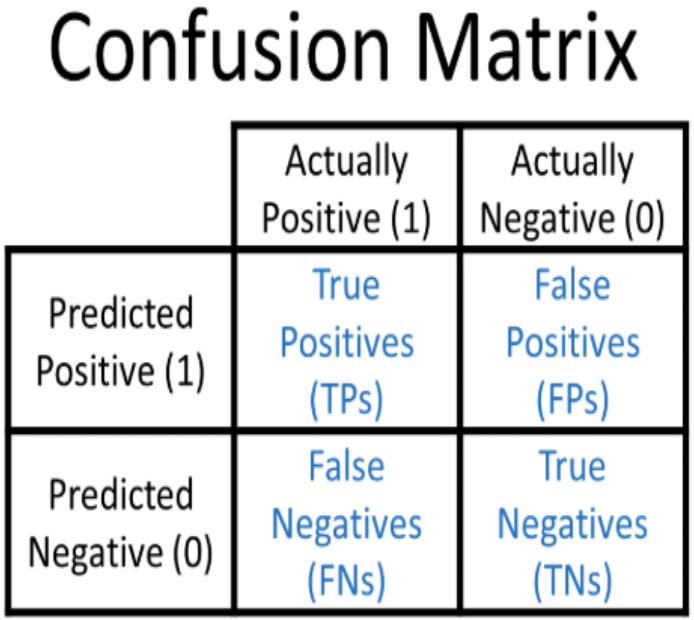
Text Classification Model



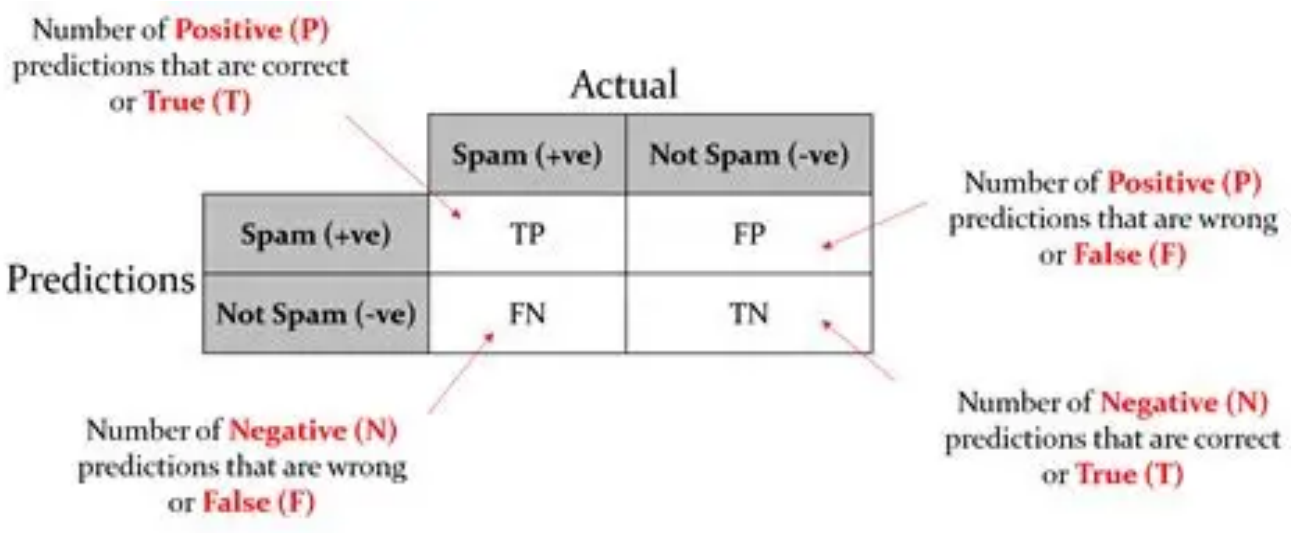
Text Classification Model

Key Components of a Confusion Matrix

- ***True Positive (TP)***: The model correctly predicts a positive outcome.
- ***True Negative (TN)***: The model correctly predicts a negative outcome.
- ***False Positive (FP)***: The model incorrectly predicts a positive outcome (Type I error).
- ***False Negative (FN)***: The model incorrectly predicts a negative outcome (Type II error)



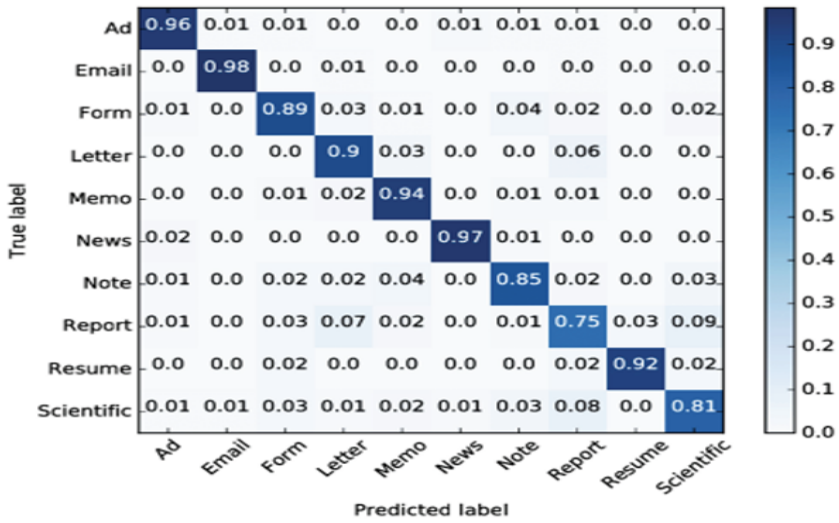
Text Classification Model

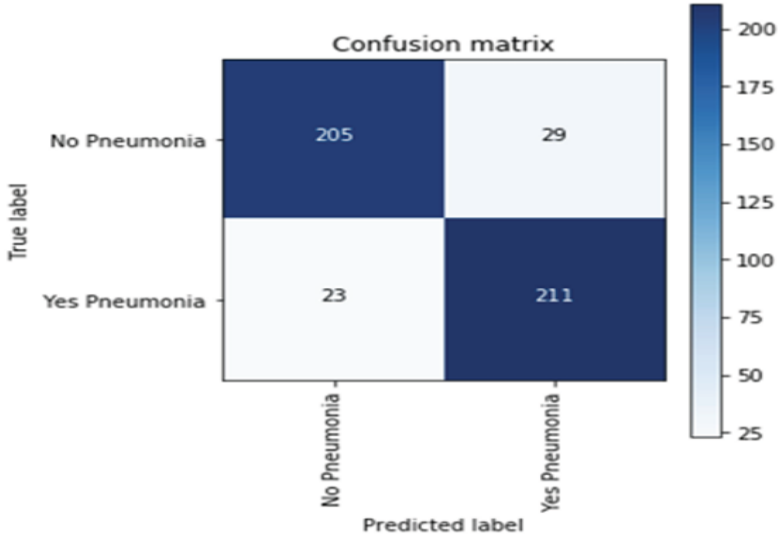


T or F : if prediction = actual or not

P or N : based on the prediction only

Text Classification Model





Whenever the diagonal of the propagation matrix has higher values= better accuracy

TEXT CLASSIFICATION Steps:

- 1. Data Collection: (Labelled Data)
- 2. Text Pre-processing (Stop Words removal, normalization,…etc)
- 3. Feature Extraction : BOW, TF-IDF…etc(specify which features are important for your classification, for example chemical labels refer to scientific articles….etc)
- 4. ML Algorithm Selection (Logistic Regression, Naïve Bayes, decision tree…etc)
- 5. Classification Evaluation (Precision, recall, F1).

Challenges in Text Classification

- Imbalanced Data : When one class (e.g., ham) dominates the other (spam), the model might become biased.
- Ambiguity in Language : Words can have multiple meanings depending on context.
- Preprocessing Decisions : Choosing what to clean and what to keep is critical.

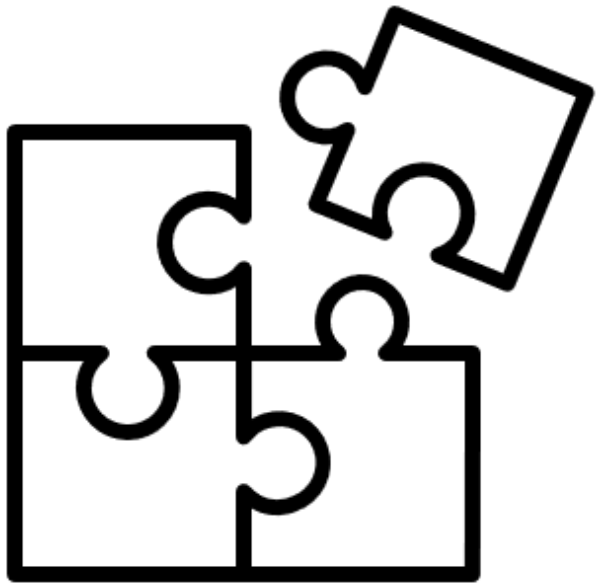
Hands-On

Other Useful Datasets for Text Classification:

Use Case	Dataset	Link
Sentiment Analysis	IMDB Movie Reviews	Kaggle IMDB Dataset
Topic Classification	BBC News Dataset	Kaggle BBC News Dataset
General Text Classification	AG News Dataset	FastText Dataset Page

National Telecommunication Institute

Quiz




Thanks