Final Project/Case-study Description

Students may opt to do simple **project** by **applying ML algorithms** on text data. This data is to be **extracted** from **twitter**.

OR

Students may opt for **research-based case-study**. Here the students will have to choose an **NLP specific research topic** such as NER for Medical Domain, Stemmers for Indian Languages, Word Embeddings for Indian Languages; and **explore** on the **different aspects** of research. Students will have to **present** a **survey report** on these and a **demo** to either **existing techniques** or **innovate** a **new approach**.

Project: ML Algorithms on Text Data

Each student must extract tweets from twitter. Perform pre-processing and text representation. Apply ML algorithms for classification/clustering.

1. Creating Datasets
    a. Extract 5000 tweets with any 5 search labels of your choice. (1000 each). Eg(#cricket, #football, #basketball, #tennis, #hockey).
    b. Create one dataset for all the tweets extracted along with labels as second column. Shuffle the dataset.
2. Pre-processing
    a. Clean the data by removing tags, user handles, numbers, and other characters.
    b. Stem tokens for basic vectorization
    c. Lemma tokens for embeddings
3. Text representation
    a. Vectorise each document in the dataset with tf-idf vectorization with n-grams (use stemmed data).
    b. Create document embeddings by summation of word vectors taken from any two pre-trained models. The tokens must be lemmas.
4. Apply machine learning techniques (any two algorithms) for classification/clustering on
    a. 3.a data
    b. 3.b data
5. Evaluate the results (4.a and 4.b) which outperforms.
    a. For clustering compare at least 10 records' label with the clusters created.
    b. Present a chart as for classification:

	knn	svm
Basic Vectorization	80%	90%
Embeddings	85%	92%

Research based Case-study in NLP

1. Choose a topic of your interest in NLP domain such as NER, coreference resolution, POS tagging for Indian Languages, Machine Translation, Question Answering, Text Summarization, Chat-bots, Grammar checkers, Sentiment Analysis, etc.
2. Identify min 3 papers in domain. Summarize them as a report with charts and other visuals.
3. Try implementing any one method or innovate new approach.

Final Presentation

Each student has to do a presentation on the project/case-study (5 min) and demo(5 min).