A Sri Lankan startup company wants to setup a sports information service for sports enthusiasts who want to keep up to date with their favourite sports and teams. In order to support their plan, they want to know how they can access every single relevant piece of sports news as soon as possible.

(a) Using the Twitter API, collect at least 10,000 sports news items in the English language from the past year.

(b) Describe the dataset collected in terms of the number of news items, the total number of words, the number of unique words and the distribution of lengths in tokens of the new items.

(c) Using a suitable mechanism identify possible duplicates (or near duplicates) and remove them from the dataset. Describe the resulting deduplicated dataset as in (b) above.

(d) Check the nature of the dataset and take any action required to clean the dataset. Tokenize the news items and preprocess the data for feature extraction.

(e) Perform feature extraction on the dataset by creating sparse and dense vector representations. Describe the final dataset(s) in terms of number of news items and number of features for each such feature extraction method. Randomly select 300 data points from this final dataset for testing.

(f) Using machine learning1, categorize the rest of the news items in the dataset into an appropriate number of groups using any appropriate technique and justify the categorization scheme.

(g) Manually annotate the 300 randomly selected news items from the dataset into different sports. Check the accuracy of the model learned in (f) using this annotated sample as the ground truth.

(h) Assume that the categorization you arrived at in (f) is close enough to the 'ground truth' and use the categories learnt as their labels. Take necessary action to remedy any imbalance in the data you may have. Use 3 different non-deep learning algorithms from scikit-learn that would help you classify this data, giving reasons for the choice of each such algorithm. Comment on the performance of each of the classifiers.

(i)Use the same assumptions as (h) and explore 2 deep learning architectures2 to try to improve the sports data classification results. Comment on the performance of the deep learning models.

(j) How would you detect if your deep learning model has overfitted the training data and what could you do about it, if so?