

PGData Science Sem 2

Case Study:– NLP

Marks :- 60

Date of Submission:- 30/04/2023

- 1.Find any dataset having text and labels(binary or multiclass)
- 2.Preprocess the text
- 3.Perform Descriptive stats on processed text including but not limited to Count of words per class, word cloud etc.
- 4.Create word embeddings using TF-IDF/bag of words
- 5.Implement different supervised ML techniques to build a classification model. Calculate accuracy metrics.
- 6.Implement the K means algorithm to build a cluster analysis model. Find out the right number of clusters based on elbow method.
- 7.Compare the Cluster analysis output with classification output(supervised) to see if you can find any patterns.
- 8.Implement the topic modelling LDA algorithm to find the topics from the text. Compare the output with the Cluster analysis model.