Bahria University

Lahore Campus
Department of Computer Sciences

Natural Language Processing

ASSIGNMENT # 02

DUE DATE: 10, Dec 2023

Instructor Name: Tayyab Mir	**Course Code:**
Program: BSCS	**Max marks: 10**

Instructions:

- The assignment should be submitted before the deadline.
- Late submission is not allowed.
- Plagiarism will be considered as serious academic offense and may result in F grade.
-

PROBLEM/TASK DESCRIPTION

Download "Twitter US Airline Sentiment" dataset from the following link
https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

Problem definition:
Social media sentimental analysis is interesting field with the aim to analyze social conservation and determine deeper context as they apply to a topic or theme. However, it is challenging as tweets are unstructured, informal, and noisy in nature. Also, it involves natural language complexities like words with same meanings (Polysemy). Most of the existing approaches mainly rely on clean textual data, however Twitter data is quite noisy in real life.

Dataset:

The dataset was scraped from February of 2015 and contributors were asked to first classify positive, negative and neutral tweets, followed by categorizing the negative reasons (i.e. "late flight" or "Rude services"). A labelled dataset is uploaded on **Github.**

In this assignment, you are required to explore the data using python script. (perform exploratory data analyses). Display the input data characteristics without making any changes. Now apply all the natural language processing techniques i.e. normalization, removing stop words, stemming/lemmatization, filtered columns, etc. After data pre-processing, show the difference between the data before pre-processing and after pre-processing using any visualization techniques.