Week 8: Deliverables
Group Name: Awwab Ahmed
Name: Awwab Ahmed
Email: awwab.k.ahmed@gmail.com
Country: USA
College/Company: University of South Florida
Specialization: NLP

Problem description:

The term hate speech is understood as any type of verbal, written or behavioral communication that attacks or uses derogatory or discriminatory language against a person or group based on what they are, in other words, based on their religion, ethnicity, nationality, race, color, ancestry, sex or another identity factor. In this problem, We will take you through a hate speech detection model with Machine Learning and Python.

Hate Speech Detection is generally a task of sentiment classification. So for training, a model that can classify hate speech from a certain piece of text can be achieved by training it on data that is generally used to classify sentiments. So for the task of hate speech detection model, We will use the Twitter tweets to identify tweets containing hate speech

Data Understanding:

| | |
|---|---|
| **Total number of observations** | 29530 |
| **Total number of files** | 1 |
| **Total number of features** | 3 |
| **Base format of the file** | csv |
| **Size of the data** | 3.1 MB |

What type of data you have got for analysis:

The data consists of one column with IDs and the other is a tweet. The data is merely entries of language and sentences and there are not a variety and columns for correlations to be generated between. This is a natural language processing task.

What are the problems in the data ( number of NA values, outliers , skewed etc):

The problem is that the data is not cleaned. It consists of URLs as well as @ symbols that are not correlated with the tweets or will include hate-speech.

What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?:

I will need to remove all URLs and @ symbols from the data in order for the model to train in the most correct way possible. This will be done in my data preprocessing procedure and will consist of using the apply() function to all rows in the dataset.