Week 7: Deliverables

Group Details

- Name: William Harvey
- Email: will.harvey@att.net
- Country: United States of America
- College: University of California, Berkeley
- Specialization: NLP Analyst

Problem Description

- I am doing the advance NLP project on detecting hate speech on Twitter. Hate speech is a very important problem, as it attacks a person for their identity by using defamatory language. It is commonly used in today's world because Twitter makes it so you can connect with millions of people from across the globe, however hate speech should be monitored and shut down. I will be using sentiment Twitter data to train a classifier to identify hate speech on the platform.

Business Understanding

- Twitter is one of the biggest companies in the world, and it should always try to focus on expanding its market share and influence on the world. In order to do this, it is important to create a safe environment for people of all races and backgrounds to be able to share their voice. By being able to immediately identify hate speech, Twitter can remove these tweets from their platform and continue to foster the community they wish.

Project Lifecycle (with deadlines)

- July 26, 2022: All the data understanding and exploration will be complete.
- August 2, 2022: Data will be clean and balanced, and ready for visualizations and models to be made
- August 9, 2022: EDA and visualizations will be made, including word clouds, topic models, and comparison graphs
- August 16, 2022: EDA will be presented to business, along with my proposed modeling technique
- August 23, 2022: Model will be perfected, and code will be vectorized.
- August 30, 2022: Final Report and Model code will be presented to the business

Data Intake Report

Name: <Detecting Hate Speech>
Report date: <7/19/22>
Internship Batch:<LISUM10>
Version:<1.0>
Data intake by:<William Harvey>
Data intake reviewer:< >
Data storage location: <https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech?select=train_E6oV3lV.csv>

Tabular data details:

Total number of observations	31962
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	1635543 bytes

Total number of observations	17197
Total number of files	1
Total number of features	2
Base format of the file	.csv
Size of the data	3103165 bytes

Note: Replicate same table with file name if you have more than one file.

In the first dataset, the data is imbalanced, so I will use the resample method from sklearn to create a balanced set that won't favor the majority.