NLP-Project Report_Group20

Project Title: Hate Speech Detection

Team Members:

Vishnu Manoj Deepala(vishnumanojdeepala@my.unt.edu)-**UNT ID:11519490**

Seemaparvez Shaik(seemaparvezshaik@my.unt.edu)-**UNT ID:11512343**

Sai Anjali Potula (saianjalipotula@my.unt.edu)-**UNT ID-11519812**

Report:

Project Objective:

In this project, we are going to focus on **three** objectives:

We're going to employ Python-based NLP machine learning techniques to construct a hate-speech detection program. Machine learning is the process of teaching machines to do specific tasks by using data to train them.

- In this scenario, we'll utilize data from Kaggle or extract data from Twitter using a tool called "Twint".
- We'll then extract terms that indicate prominence inside hate speech using an NLP (or Natural Language Processing) approach called Tf-Idf vectorization.
- Finally, using data retrieved from Kaggle or by using "Twint", we'll train the computer to identify hate speech using a machine learning approach called logistic regression, which is commonly used for probability estimates (or any kind of data you wish to utilize for training).

Text Vectorization: Term Frequency — Inverse Document Frequency (TFIDF)

By counting the number of times words appear in a document, Bag of Words (BoW) transforms the text into a feature vector. It disregards the significance of words. TFIDF (Term Frequency — Inverse Text Frequency) is based on the Bag of Terms (BoW) model, which provides information about the less and more important words in a document. In information retrieval, the relevance of a word in the text is extremely important.

For example, if you search for anything on a search engine, TFIDF values can assist us to find the most relevant papers connected to our search. So, We will be using **Text Vectorization: Term Frequency — Inverse Document Frequency (TFIDF)** to extract terms that indicate prominence inside hate speech

Twint:

TWINT is a Python-based sophisticated Twitter scraping application that allows you to scrape Tweets from Twitter accounts without having to use Twitter's API.

Twint takes advantage of Twitter's search operators to allow you to scrape Tweets from specific individuals, scrape Tweets related to specific themes, hashtags, and trends, and sift out sensitive information from tweets like e-mail addresses and phone numbers. It's a fantastic tool for getting information.

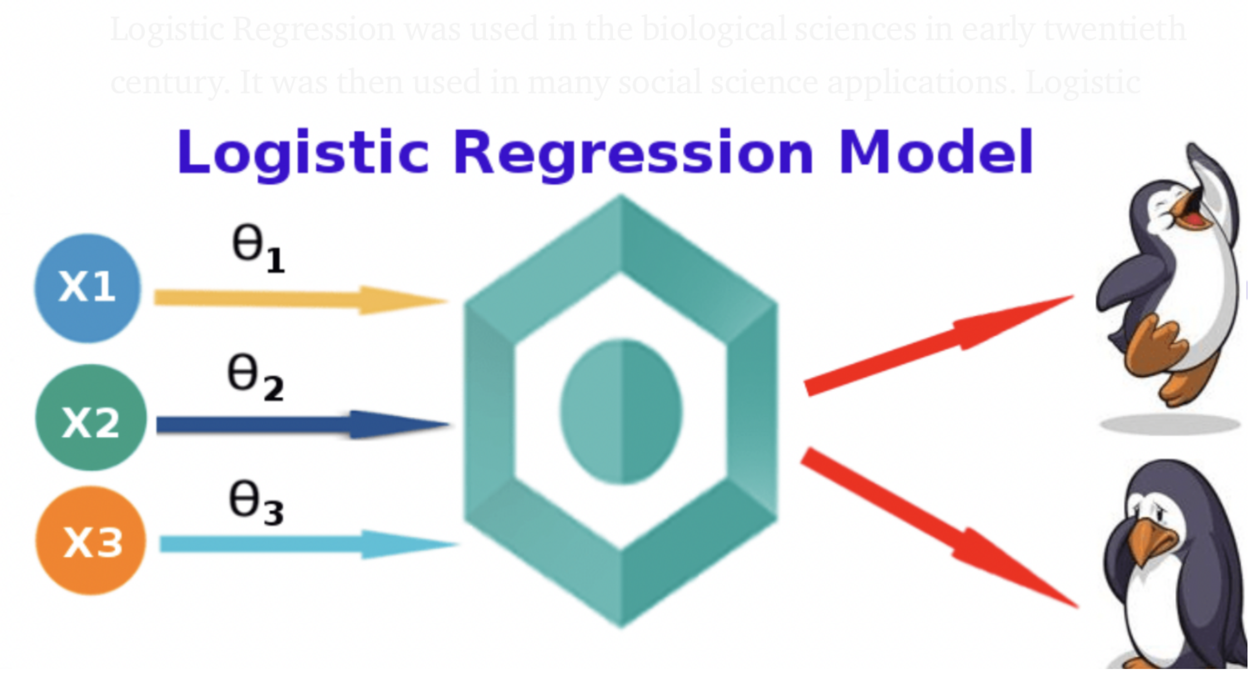
Benefits:

Advantages of Twint vs Twitter API:

Twint can retrieve practically all Tweets, however the Twitter API only allows you to retrieve the last 3200 Tweets.
Twint offers a quick setup time.
It may be used without requiring a Twitter sign-up or log-in.
There are no rate restrictions.

Most probably we'll be using TWINT to extract data.

Logistic Regression:



Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic

Logistic Regression is used when the dependent variable(target) is categorical.

For example,

- To predict whether an email is a spam (1) or (0)
- Whether the tumor is malignant (1) or not (0)

In this case, We will use Logistic regression to predict a tweet is a hate tweet or not