The study of gender in the twiter

Abstract:

The goal of this project was to use classification models to predict With the credibility of the tweeter's gender, whether that is true or not most popular words.

DESIGN

This project helps us to know the characteristics of the tweeters in terms of their types and the most used words for them.

Data

The dataset contains 20050 datapoint with 26 features for each, The features I am using, such as gender ,gender confidence , profile , profile confidence , fav number , text , tweet count , The features are integer , float , string and bool.

ALGORITHMS

Feature Engineering

Converting categorical features into numerical features.

Models

1. Logistic Regression
2. Decision Tree Classifier
3. Gaussian Naive-Bayes
4. Random Forest Classifier
5. K-Nearest Neighbours

Model Evaluation and Selection

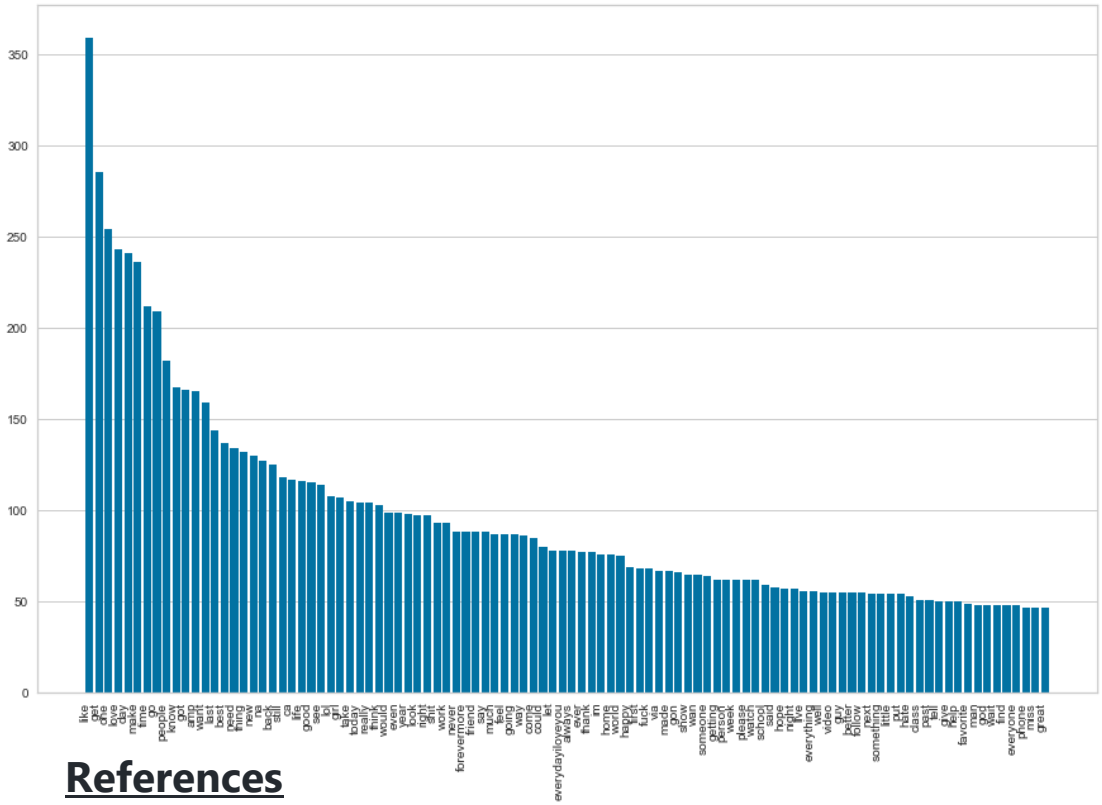
The entire dataset of records was split into 80/20 train vs. test. Below is the evaluation of each model.

TOOLS

- Numpy and Pandas for data processing
- Scikit-learn for modeling
- Matplotlib and Seaborn for visualization

COMMUNICATION

Presentation that includes visuals for communicating the objectives and findings.



References

Dataset

https://www.kaggle.com/siddheshshelke/twitter-sample-dataset?select=twitter+dataset.csv

***https://www.scikit-yb.org/en/latest/api/text/index.html**