Name - Vijay Tadikamalla
Roll. no- CS17BTECH11040

Assignment 1 - Classification tasks in NLP

Foundation of Machine Learning

Choices and Design decisions

- **Dataset:** Thomas Davidson's Dataset (Davidson et al., 2017).
- **Smoothing kernel choice:** The kernels used in KDE based Bayes classifier are
    a. Gaussian
    b. Epanechnikov
    c. Cosine

Feature map and the principles

- **Bag-of-words (bow):** We consider each attribute/feature as the count of the number of times a particular word appears in the sentence/para.
- **TF-IDF** is used to select important terms from Unigram, Bigram and Trigrams occurring in the tweets.
- **Pos-Tagging** is used to tag features (part of sentences) from tweets.
- Some other features like **no. of hashtags, URLs,** and **mentions** are used as a part of the feature vector.

Non-probabilistic k-NN Classifier

Hyperparameter: nearest neighbors : 5, p: 1

k-NN	Precision	Recall	F1-score
Hate Speech	0.36	0.05	0.09
Offensive Language	0.83	0.92	0.87
Neither	0.51	0.4	0.45
Avg	0.57	0.46	0.47
Accuracy	0.78		

KDE based Bayes classifier

Hyperparameter: Kernal: Gaussian, h: 10

KDE	Precision	Recall	F1-score
Hate Speech	0	0	0
Offensive Language	0.77	0.99	0.87
Neither	0.59	0.07	0.13
Avg	0.45	0.35	0.33
Accuracy	0.77		

Kernelized SVM

Kernelized SVM	Precision	Recall	F1-score
Hate Speech	0.47	0.04	0.08
Offensive Language	0.79	0.99	0.88
Neither	0.77	0.12	0.21
Avg	0.68	0.38	0.39
Accuracy	0.78		

Hyperparameter: C: 10, gamma: 0.01

Regularized Logistic regression

Hyperparameter: C: 0.1

Logistic regression	Precision	Recall	F1-score
Hate Speech	0.53	0.28	0.36
Offensive Language	0.93	0.95	0.94
Neither	0.83	0.88	0.85
Avg	0.76	0.70	0.72
Accuracy	0.90		

Gaussian based Bayes classifier

Gaussian	Precision	Recall	F1-score
Hate Speech	0.09	0.44	0.15
Offensive Language	0.86	0.60	0.71
Neither	0.44	0.39	0.42
Avg	0.46	0.48	0.43
Accuracy	0.56		