Assignment 3

[Q1 datasets in datasets folder]

Q2.

a) Performance:

- Random Classifier: 20.92%
- Majority Class Classifier: 35.25%

b) [trained and saved parameters in txt file]

c) **Bernoulli Naïve Bayes**:
Range of Hyperparameters: {'alpha': [0.0000001, 0.00001, 0.0001, 0.001, 0.1, 1, 10]}
Best Hyperparameter: 'alpha': 0.1

Decision Trees:
Range of Hyperparameters: {'max_depth': [10,12,14,16,20], 'min_samples_split': [2,4,6,8,10], "max_features": [1,4, 6, 20, 40, 50, 80, 100]}
Best Hyperparameter: {'max_depth': 12, 'max_features': 40, 'min_samples_split': 4}

SVM:
Range of Hyperparameters: {'tol': [0.00001, 0.0001, 0.01, 1, 10], 'C': [0.00001, 0.001, 0.01, 1, 2, 10] }
Best Hyperparameter: {'C': 0.01, 'tol': 1e-05}

d) **Training F1 - Score**
Bernoulli Naïve Bayes: 71.15%
Decision Trees: 40.72%
SVM: 84.78%

Valid F1 – Score
Bernoulli Naïve Bayes: 91.10%
Decision Trees: 45.50 %
SVM: 94.89%

Test F1 – Score:
Bernoulli Naïve Bayes: 86.56%
Decision Trees: 61.31%
SVM: 93.64

e) SVM performed the best amongst all three classifiers. Since the data consisted of binary values he nce SVM was able to correctly classified, maximized the margin and created an optimal hyperpla ne. The parameter 'C' affected the margin maximization and the value of 0.01 maximized the mar gin.

Q3

a) [trained and saved parameters in txt file]

b) **Gaussian Naïve Bayes**: None
Best Hyperparameter: None

Decision Trees: {'max_depth': [40, 60, 80, 90, 100], 'min_samples_split': [10, 20, 30, 40, 50, 60, 10 0], "max_features": [1, 4, 6, 20, 40, 50}
Best Hyperparameter: {'max_depth': 90, 'max_features': 50, 'min_samples_split': 100}

SVM: {'tol': [0.01, 0.01, 0.1, 1, 10], 'C': [0.00001, 0.001, 0.01, 1, 2, 10]}
Best Hyperparameter: {'C': 1, 'tol': 0.01}

c) **Training F1 - Score**
Gaussian Naïve Bayes: 66.42%
Decision Trees: 66.92%
SVM: 81.98%

Valid F1 – Score
Gaussian Naïve Bayes: 93.5%
Decision Trees: 65.60%
SVM: 96.1%

Test F1 – Score:
Gaussian Naïve Bayes: 80.65%
Decision Trees: 86.95%
SVM: 91.83

d) Decision Trees worked best amongst the three classifiers. It was able to pick up best features for classification and hence classified the data accordingly. The hyperparameter max_depth was abl e to pick optimal number of examples which yielded good prediction and max_features picked op timal number of features that classified the data well.

e) The Binary Bag words had better performance as compared to Frequency Bag of words. Since the data consisted of multiple classes hence BBOW classifiers were able to yield better predictions an d classifications. BBOW Naïve Bayes worked better than FBOW Naïve Bayes than in BBOW as the data was binary and was easy to classify. Decision Trees worked better in FBOW than BBOW as t here were varying frequencies for every word in every review. SVM worked better for BBOW tha n in FBOW as it was able to distinguish well due to binary representation of the data making class ification easier.

f) BBOW works best for this type of dataset because the data is imbalanced therefore BBOW repres entation makes it easier to classify the data by converting the data into binary representation and making it a two class problem.

Q4

a) Performance:

- Random Classifier: 49.73%

b) [trained and saved parameters in txt file]

c) **Bernoulli Naïve Bayes**: {'alpha': [0.0000001, 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 20, 30]}
Best Hyperparameter: {'alpha': 1}

Decision Trees: {'max_depth': [10,12,14,16,20], 'min_samples_split': [2,4,6,8,10], "max_features": [1000, 2000, 3000, 4000]}
Best Hyperparameter: {'max_depth': 16, 'max_features': 4000, 'min_samples_split': 8}

SVM: {'tol': [0.00001, 0.0001, 0.01, 1, 10], 'C': [0.00001, 0.001, 0.01, 1, 2, 10]}
Best Hyperparameter: {'C': 0.01, 'tol': 1}

d) **Training F1 - Score**
Bernoulli Naïve Bayes: 86.94%
Decision Trees: 82.69%
SVM: 96.23%

Valid F1 – Score
Bernoulli Naïve Bayes: 88.71%
Decision Trees: 84.46%
SVM: 96.73%

Test F1 – Score:
Bernoulli Naïve Bayes: 87.08%
Decision Trees: 80.72%
SVM: 95.33%

e) SVM performed the best amongst all three classifiers. Since the data consisted of binary values hence SVM was able to correctly classified, maximized the margin and created an optimal hyperplane. The parameter 'C' affected the margin maximization and the value of 0.01 maximized the margin.

Q5

a) [trained and saved parameters in txt file]

b) **Gaussian Naïve Bayes**: None
Best Hyperparameter: None

Decision Trees: {'max_depth': [60, 80, 100, 200, 400, 600, 1000], 'min_samples_split': [10, 20, 40, 60, 80, 90, 100], "max_features": [20, 40, 60, 100, 200, 600, 800, 1000]}
Best Hyperparameter: {'max_depth': 200, 'max_features': 800, 'min_samples_split': 90}

SVM: {'tol': [0.000001, 0.0001, 0.001, 0.1, 0.1, 10], 'C': [0.000001, 0.0001, 0.01, 0.1, 1, 2, 10]}
Best Hyperparameter: {'C': 2, 'tol': 1e-06}

c) **Training F1 – Score**
Gaussian Naïve Bayes: 86.30%
Decision Trees: 62.87%
SVM: 95.10%

Valid F1 – Score
Gaussian Naïve Bayes: 87.39%
Decision Trees: 86.61%
SVM: 95.36%

Test F1 – Score:
Gaussian Naïve Bayes: 95.36%
Decision Trees: 88.02%
SVM: 94.5%

d) Gaussian Naïve Bayes works better amongst all three of the classifiers. Since the data to be classif ied was of two class problem hence GNB was able to make better predictions.

e) FBOW classifiers worked better than the BBOW classifiers. Since the dataset was primarily of tw o class problem but with large amount of data to be classified, FBOW representation was able to make better predictions.

f) FBOW is a better representation for this type of data as IMDB has 2 class problem, balanced data but more data to be classified. Naïve Bayes of BBOW worked slightly better in BBOW because it d id not have to deal with multiple classes and hence could classify better. Decision Trees worked b etter in FBOW than in BBOW due to varying frequencies of different words. SVM worked better i n BBOW due to classification between just two classes.

g) The classifiers yielded better results in IMDB dataset as compared to in Yelp dataset. Since, the da ta was balanced hence it resulted in lower error and better classification.