IOANNA PAPAGIANNI 2790 – CHRISTOS GOULAS 2677

QUESTION 2

For this analysis we use Jupyter Notebook and the file is in IPython format. We downloaded "training_data.txt" and "test_data.txt" from Kaggle and all files are in the same folder with "spam_sms_detection.ipynb".

Our public score with Multinomial Naive-Bayes & CountVectorizer: 0.96052
Our public score with Multinomial Naive-Bayes & TfidfVectorizer: 0.89130
Our public score with Bernoulli Naive-Bayes & CountVectorizer: 0.91489
Our public score with SVM (linear SVC) & CountVectorizer: 0.95918
Our public score with Logistic Regression & CountVectorizer: 0.94117

The multinomial Naive Bayes classifier is suitable for classification with discrete features and it performs well with text classification. The multinomial distribution normally requires integer feature counts.
However, in practice, fractional counts such as TfidfVectorizer work but do not perform as well as CountVectorizer.

We tried Bernoulli Naive-Bayes, like MultinomialNB, this classifier is suitable for discrete data. The difference is that while MultinomialNB works with occurrence counts, BernoulliNB is designed for binary/boolean features.

Linear Support Vector Classification (linear SVC) has more flexibility in the choice of penalties and loss functions and should scale better to large numbers of samples but does not perform better than MultinomialNB for spam detection. It represents **data** as points in space and tries to create a mapping with a wide as possible gap between the separate categories. Thus **SVM** can efficiently handle **sparse data** but again, not in the level of MultinomialNB.

Performance of Logistic Regression, commonly used to analyze a binary response variable, is questionable in the presence of sparse data.

KAGGLE NAME: GoulasAC, JohannaP
KAGGLE TEAM NAME: Goulas-Papagianni
KAGGLE SCORE: 0.96052