ASSIGNMENT 2
Report on Naïve Bayes Classifier

Submitted By : Anima Prasad (21CS60R66)
Arkapravo Ghosh (21CS60R64)

1. Dataset Analysis and Preprocessing
    - The dataset used here is "Twitter Sentiment Analysis, https://www.kaggle.com/arkhoshghalb/twitter- sentiment-analysis-hatred-speech.". The task is to classify racist or sexist tweets from other tweets. Given a training sample of tweets and labels, where label '1' denotes the tweet is racist/sexist and label '0' denotes the tweet is not racist/sexist.
    - The data is read from "train.csv" file.
    - The text in column "Tweet" is tokenized and all uninformative words are removed.
    - Function used for tokenization is "re.findall("[a-z0-9]+", text.tolower())".
    - Moreover, list of stopwords from https://gist.github.com/sebleier/554280 is used.
    - Using this set of tokens, M_{ij} i.e, feature matrix is created. $M_{ij} = 1$, if j-th token is present in i-th tweet/ example.
    - Feature matrix will have 31962 rows = No of examples / tweets.
    - Feature matrix will have 38961 columns = size of vocabulary.
    - Numpy array is used to represent the feature matrix and datatype used is bool.
    - Since storing integer will increase size of feature matrix by 8 times as size of integer is 8 bytes and that of bool is 1 byte.
    - When $M_{ij} = 1$, the value stored is True and for $M_{ij} = 0$, the value stored is False. There is no loss of any information using this representation.
    - Since the size of feature matrix is very large so sparse representation of matrix is used. "scipy.sparse.csr" matrix is used to represent the feature matrix as sparse matrix.

2. Data Split
    - "Train.csv" is randomly split into train and test sets with ratio 70:30.
    - Since the tweets with label "0" is much more than the tweets with label "1", hence stratification is done on the label while splitting to get a good split.

3. Naïve Bayes Classifier
    - Column "label" from dataset is used as a classifier column.
    - A naïve Bayes classifier is an algorithm that uses Bayes' theorem to classify object.
    - $$P(Y = y_k | X_1 \dots \dots X_n) = \frac{P(Y=y_k) \prod_i P(X_i|Y=y_k)}{\sum_j P(Y=y_j) \prod_i P(X_i|Y=y_j)}$$

- In above equation, y_k = each label in data i.e, '0' and '1'.
- X_i represents each token in data.
- Assumption is there is conditional independence among X_i.
- For each new tweet, the conditional probability of each token over each label is computed.
- The label with the highest probability is assigned as the predicted label of the new tweet.
- The train accuracy is 99.32%.
- The test accuracy is 95%.

4. Naïve Bayes Classifier using Laplace Correction
    - A small unit alpha (here its 1) is added in the numerator for avoid the cases where numerator becomes zero.
    - For denominator alpha*no of attributes(size of vocab) is added for normalization.
    - $P(Y = y_k|X_1 \dots \dots X_n) = \frac{P(Y=y_k)\prod_i P(X_i|Y = y_k)+l}{\sum_j P(Y=y_j)\prod_i P(X_i|Y=y_j)+lM}$
    - where l is alpha and M is no of attributes.
    - The train accuracy is 95.17%.
    - The test accuracy is 94.55%.

5. 95% confidence interval of the accuracy, precision, f-score, sensitivity and specificity.

    - Specificity is defined as the proportion of actual negatives, which got predicted as the negative.
    - Sensitivity is a measure of the proportion of actual positive cases that got predicted as positive.
    - Precision = true positive / true positive + false positive
    - F-score provides a way to combine both precision and recall into a single measure that captures both properties, giving each the same weighting.
    - 95% confidence interval, the value of constant is 1.96 based on statistics.

    - Test 95% Confidence Interval of Naive Bayes Classifier: [0.9457, 0.9544]
    - Test Precision of Naive Bayes Classifier: 0.85
    - Test F-score of Naive Bayes Classifier: 0.5
    - Test Sensitivity of Naive Bayes Classifier: 0.35
    - Test Specificity of Naïve Bayes Classifier: 1.0

    - Test 95% Confidence Interval of Naïve Bayes Classifier using Laplace Correction: [0.9409, 0.95]
    - Test Precision of Naive Bayes Classifier using Laplace Correction: 0.89
    - Test F-score of Naive Bayes Classifier using Laplace Correction: 0.4
    - Test Sensitivity of Naive Bayes Classifier using Laplace Correction: 0.25
    - Test Specificity of Naive Bayes Classifier using Laplace Correction: 1.0

- Confusion matrix : It’s a summary of prediction results on a classification problem.