Programme:

MSC IN BUSINESS ANALYTICS

Module:

Data Mining- B9BA103

Assignment Title:

CA 2 – Practical Assignment – Text Mining

Submitted to:	**Submitted By:**
Mr. Kunwar Madan	Jatin Achenkunju……… 10539553

Contents

Introduction (Text Mining) 3
Data Preparation 3
Implementation & Observation (Model Creation) 4
Model Deploy.. 5
Conclusion 5

Introduction

The main purpose of this assessment is to perform sentimental analysis using the machine learning model on the text scrapped data through websites. The classifies models used here are Support Vector Classifier, Naïve Bayes Classifier and AdaBoost Classifier.

Dataset

Two datasets are available. User_reviews.cvs will be used to train the models for the given recommendations and gain the vocabulary. The same model with best performance and gained vocabulary will be used predict recommendations in the new dataset viz. test_reviews.csv. Below are features of user_reviews.

Features	Significance
source	Website from where data is gathered
domain	Website domain
score	Score from 1 -10
score_max	Maximum score which is 10
extract	User reviews regarding the products
product	Product name

test_reviews have exactly same columns with same significance apart from score and score_max columns.

Data Preparation

Data cleaning and normalization of the datasets is carried out during the model creation and deployment. We perform below steps while cleaning the dataset.

- HTML encodings are removed using the BeautifulSoup.
- Removing all the hyperlinks, URLs, #tags, @ID mentions
- Removing all the words having less than or equal to 2 as word length.
- Removing all the Stop words and punctuations.
- Converting all the words to lower case.
- Perform stemming by using Lemmatization to find the exact root word. This will efficiently reduce the number of words.
- Target variable is converted to binary. As we need to provide the recommendations, the scores greater than or equal eight will be considered as recommended and below 8 will be considered as not recommended. So, greater than or equal to 8 is given value 1 and less than 8 is given value 0.

Once the normalization is completed the cleaned data is exported to Cleaned_PhoneReviews.csv.

Model Implementation and Evaluation

The models used here are SVC, Naïve Bayes and AdaBoost. Accuracy will be performance evaluation metric that will be used. Target variable is created with just the score column. The model implementation has to be considered in 3 scenarios using unigrams, bigrams and trigrams as tokens. The identified words are then converted to vocabulary and then converted to csv format. SMOTE is applied to balance the dataset.

10-fold cross validation is applied for both SVC and Naïve Bayes whereas 5-fold cross validation is applied for AdaBoost while deciding the hyperparameter. During each fold, accuracy is calculated in each iteration. The mean of the all the accuracy in cross validation is calculated for both SVC and Naïve Bayes. The model the best accuracy is considered for deployment on the test_reviews dataset.

Below are the three models trained under different scenarios:

1. **Only unigrams are used as tokens, which must occur in at least 10 documents.**

The shape of the dataset is (99990, 4242). After application of unigrams as tokens on the three models, below are the mean accuracy after 10-fold cross validation on SVC / Naïve Bayes and 5-fold cross validation on AdaBoost. The hyperparameter acquired from for AdaBoost grid search is 510.

MODEL	ACCURACY
SVC	0.889
Naïve Bayes	0.841
AdaBoost	0.85

From the above evaluation, after the cross-validation SVC has the best mean accuracy of about 0.88 compared to Naïve Bayes and AdaBoost at 0.84 and 0.85.

2. **Unigrams and Bigrams are used as tokens, which must occur in at least 20 documents.**

The shape of the dataset is (99990, 10034). After application of unigrams and bigrams as tokens on the three models, below are the mean accuracy after 10-fold cross validation on SVC / Naïve Bayes and 5-fold cross validation on AdaBoost. The hyperparameter acquired from for AdaBoost grid search is 520.

MODEL	ACCURACY
SVC	0.892
Naïve Bayes	0.846
AdaBoost	0.851

From the above evaluation, after the cross-validation SVC has the best mean accuracy of about 0.89 compared to Naïve Bayes and AdaBoost at 0.84 and 0.85.

3. **Unigrams, Bigrams and Trigrams are used as tokens, which must occur in at least 30 documents.**

The shape of the dataset is (99990, 7469). After application of unigrams, bigrams and trigrams as tokens on the three models, below are the mean accuracy after 10-fold cross validation on SVC / Naïve Bayes and 5-fold cross validation on AdaBoost. The hyperparameter acquired from for AdaBoost grid search is 560.

MODEL	ACCURACY
SVC	0.879
Naïve Bayes	0.833
AdaBoost	0.848

From the above evaluation, after the cross-validation SVC has the best mean accuracy of about 0.87 compared to Naïve Bayes and AdaBoost at 0.83 and 0.84.

On the basis of the best accuracy, **SVC** has created vocabulary tokens using unigrams and bigrams at accuracy **0.89 or 89%.** Thereafter, creating a .sav file to be deployed on the similar dataset to predict recommendations.

Model Deployment

The best performing model to predict user recommendations is SVC model implemented using unigrams and bigrams model. This model will be deployed on test_reviews.csv. to predict the recommendation using the same vocabulary identified with 1’s and 0’s. 1 means the product is recommended and 0 means it is not recommended.

Conclusion

The deployed model was successfully able to predict the user recommendations on the new dataset, based on the user reviews. This can be used to decide if the new user can should buy the product.