NLP ASSIGNMENT 01 REPORT

Name: Jude Damian Sequeira

Student ID: 220431413

INTRODUCTION

Fake news is the intentional broadcasting of false or misleading claims as news, where the statements are purposely deceitful. It is important to recognize and differentiate between Fake and Real news. One method is to use human labour to check facts for every piece of information, which is time consuming and needs expertise that cannot be shared. The other approach is that we can use machine learning and artificial intelligence tools to automate the identification of fake news.

This report proposes a methodology to use a ML model that will detect if an article is real or fake based on its texts and other features on a tab-separated text file dataset. Then, Pre-processing and feature selection methods are applied to the experiment and the best model evaluated on 10-fold Cross Validation data based on metrics such as accuracy, precision, recall and F1 score. I propose to create the model using different Bag-of-Words (BOW)methods on the SVM classifier. The best model across all the 10-folds for each experiment will test the unseen data, with its confusion matrix plotted for unseen test data.

EXPERIMENTS

The dataset has over 10,000 statements on current affairs which have been annotated with labels for different degrees of fakeness, ranging from 'true' to 'pants on fire'. Each label is given either a tag of Real or Fake to simplify the task of classification.

1) For the first experiment(question1-4), the unigram BOW method is used along with few Pre-processing steps such as converting text tokens to lower case letters and punctuation removal were performed to reduce noisy data. A dictionary of the unigram model was obtained with its keys as unigram tokens and weights as the binary representation(i:e only 1 if token present or else 0). The input to the classifier is the feature vector of BOW and the true labels of the training data. The Train-Test split is in the ratio 80:20 respectively. Before directly testing on the test data, the 10 fold cross validation algorithm is implemented. The average metrics values of Precision, Recall and F1 score is computed for evaluation purposes between other techniques used later. This step gives an idea on how the trained model will behave on any other test data that may be used for classification. The unseen test data is not yet implemented until all other techniques are evaluated based on the average metric score over the 10-fold cross validation.

2) For the second experiment, some extra pre-processing steps were implemented such as stop-word removal and lemmatization along with converting to lowercase and punctuation removal. Stop-words are words like "the", "is", "are" which occur very frequently in texts but not giving much context to predict fake news. Lemmatizing is a technique which converts a word token to its base form. For this experiment the lemmatizing 'pos' parameter was given POS-tagging input as 'v' which means the lemmatizer function will only convert verbs to its base form. Example: "running" will lemmatize to "run". There are also POS-tags that could be used such as 'n' for lemmatizing nouns, 'r' for adverbs and 'j' for adjectives. Secondly, the Bigram model is also used which considers two tokens as keys and weight is given as per the number of times that bigram token is appearing in the text. Hence weights can be 1,2,3,etc which signifies its frequency in the text. After this the 10-fold average performance is computed with the first fold confusion matrix plotted.

3) For the third experiment the same process was used as Bigram Modelling, with only difference in the number of Ngram=3. Similarly from previous experiments, this experiment also computes the mean of the metrics across all 10-folds with its 1st fold confusion matrix displayed. Now the models are ready to be compared to find the best technique for this classification problem. The best model chosen is then used to evaluate the unseen test data with its metrics values and confusion matrix plotted.

RESULTS

Comparison of average metrics scored during 10-Fold Cross Validation

Technique	Precision	Recall	F1 score
a. Unigram-Binary Bow	0.535225	0.561906	0.444666
b. Bigram Bow with further pre-processing	0.538833	0.56458	0.446507
c. Trigram Bow with further pre-processing	0.537757	0.563607	0.440824

The table above shows the values obtained for each experiment over all the 10-folds of the validation data. The scores for each metrics are quite similar although the Bigram Bow model with lemmatization and stop-words removal added in the pre-processing stage scores slightly better for each metric compared to the binary unigram model and the trigram model. Based on these results, the Bigram BOW model is chosen to test the unseen test data which gives an output metrics Precision: 0.535441, Recall: 0.549536, F1 Score:0.425653. The precision values of the test and and 10-fold cross validation are quite similar, although the Recall and F1 score are a little less compared to the Bigram Model.

Precision talks about how precise/accurate your model is out of those predicted positive and how many of them are actual positive. Recall calculates how many of the Actual Positives our model captures through labelling it as Positive (True Positive). F1 Score is a better measure to use if we need to seek a balance between Precision and Recall and if there is an uneven class distribution (large number of Actual Negatives).

From the table and test results, the classifier classifies REAL text data with precision approx. 53% and classified REAL Text correctly with a recall of approx. 56%. But the F1 score is quite low with approx. 44% which means that the number of REAL texts and FAKE texts are imbalanced.

CONCLUSION

After comparing all the approaches in the experiment, the Bigram model shows the best performance compared to unigram binary BOW and Trigram Bow. Although the output metrics values for the best model is low for the Fake review classification problem, many other techniques and approaches can be used to improve model performance such as balancing the number of REAL and FAKE news text and using other Machine Learning techniques for the classification problem.