IT550: Information Retrieval

Assignment 8: Fake News Classification

Instructor: Prasenjit Majumder

Learning Outcome: At the end of this assignment you will learn text classification using ensemble approach. And perform evaluation using accuracy

1 Problem description

Text classification is the task of assigning a text to its proper category. Each classifier that we use for classification has its own limitations and disadvantages. In ensemble learning we can combine all the classifiers. And use the predictions of all the classifiers and give the final prediction. So reduces the disadvantages of a single classifier. Ensemble based approach consists of three techniques: Voting classifier, Stacking classifier and Bagging classifier.

- Voting Classifier: In this approach, we get the predictions of all the classifiers and based on majority vote the prediction is selected
- Stacking Classifier: In this approach we get the predictions from all the classifiers and feed as input to a meta classifier. Meta classifier then predicts the output.

2 Implementation

2.1 Dataset

- For this assignment we will use Fake news dataset from Kaggle.
  https://drive.google.com/file/d/1q0ZimHCtMlhftljfVuy0i-wcgPnDUYjl/view?usp=sharing
- train.csv: A full training dataset with the following attributes:
  1. id: unique id for a news article
  2. title: the title of a news article
  3. author: author of the news article
  4. text: the text of the article; could be incomplete
  5. label: a label that marks the article as potentially unreliable 1: unreliable 0: reliable
- test.csv: A testing training dataset with all the same attributes at train.csv without the label.
- submit.csv: A sample submission

2.2 Exercise

- Use the train.csv file for this experiment.
- Use TF-IDF to represent document.
- Classify the documents using Voting Classifier and Stacking classifier
- The classifiers that you have to make use of for this ensemble approach are:
  - Mulitnomial Naive Bayes
  - RandomForest Classifier
  - Logistic Regression
- Perform 5 fold cross validation.

- Use accuracy for evaluation of the classifier
- First use only the titles of the documents to perform the task. Then use the text of the documents to perform the task.
- Report the accuracy for all the approaches

3 References

- https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_files.html
- https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_validate.html
- https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
- https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-
- https://towardsdatascience.com/applying-machine-learning-to-classify-an-unsupervised-text-document

4 Submission

- You have to submit your assignment in Jupyter notebook with proper comments and explanation of your approach.
- For each of your approach you have to show the accuracy wrt each approach
- The submission deadline for this assignment is **1st April 2020 11 pm**