**Analysis**

We choose the following classification: J48 pruned tree the following classifications are depended on the features and they are faster than the classification on the function and meta sector in WEKA.

The results for the first 10 files of Google syntactic n-grams are:

**Test mode: 10-fold cross-validation.**

Cross-Validation is a technique used in model selection to better estimate the test error of a predictive model. The idea behind cross-validation is to create a number of partitions of sample observations, known as the validation sets, from the training data set. After fitting a model on to the training data, its performance is measured against each validation set and then averaged, gaining a better assessment of how the model will perform when asked to predict for new observations. The number of partitions to construct depends on the number of observations in the sample data set as well as the decision made regarding the bias-variance trade-off, with more partitions leading to a smaller bias but a higher variance.

**J48 classification:**

J48 algorithm which is used to create Univariate Decision Trees. The research study also discuss about the idea of multivariate decision tree with process of classify instance by using more than one attribute at each internal node. The core concept behind the topic is to get depth knowledge with new areas of research by explore more about data, information, knowledge, data mining techniques, and tools. All the results with experiment on Weka are finally examined

**The Precision, Recall and F1 measure of each run.**

Scheme: weka.classifiers.meta.FilteredClassifier -F "weka.filters.unsupervised.attribute.Remove -R first" -W weka.classifiers.trees.J48 -- -C 0.25 -M 2

Relation: similar

Instances: 14047

Attributes: 22

Test mode: evaluate on training data

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 42 77.7778 %

Incorrectly Classified Instances 12 22.2222 %

Kappa statistic 0.2703

Mean absolute error 0.2383

Root mean squared error 0.4559

Relative absolute error 67.6083 %

Root relative squared error 109.3273 %

Total Number of Instances 54

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.333 0.095 0.500 0.333 0.400 0.279 0.663 0.368 True

0.905 0.667 0.826 0.905 0.864 0.279 0.663 0.842 False

Weighted Avg. 0.778 0.540 0.754 0.778 0.761 0.279 0.663 0.737

=== Confusion Matrix ===

a b<-- classified as

4 8 | a = True

4 38 | b = False

**false-negative**

1 2:False 1:True + 0.833 (battleship-arrow)

1 2:False 1:True + 1 (battleship-air)

4 2:False 1:True + 1 (ant-antenna)

1 2:False 1:True + 1 (battleship-anchor)

**true-negative**

6 1:True 2:False + 1 (ant-animal)

5 1:True 2:False + 1 (bag-accessory)

6 1:True 2:False + 1 (battleship-artifact)

5 1:True 2:False + 0.9 (ant-arthropod)

5 1:True 2:False + 0.938 (axe-artefact)

5 1:True 2:False + 0.944 (ambulance-artifact)

5 1:True 2:False + 1 (axe-artifact)

5 1:True 2:False + 0.9 (ambulance-artefact)

**false-positive:**

1 2:False2:False 1 (ambulance-barge)

2 2:False2:False 0.944 (ambulance-air)

3 2:False2:False 0.944 (battleship-abandonment)

4 2:False2:False 0.944 (alligator-addition)

1 2:False2:False 1 (apple-banana)

2 2:False2:False 1 (bag-wallet)

3 2:False2:False 1 (ambulance-accelerator)

4 2:False2:False 1 (alligator-arrears)

2 2:False2:False 0.938 (apple-acne)

3 2:False2:False 0.938 (ambulance-aspect)

**true-positive:**

6 1:True1:True 0.833 (bag-artefact)

5 1:True1:True 1 (battleship-artefact)

6 1:True1:True 1 (bag-artifact)

5 1:True 1:True 1 (alligator-animal)

Cross-validation is a good technique to test a model on its predictive performance. While a model may minimize the Mean Squared Error on the training data, it can be optimistic in its predictive error. The partitions used in cross-validation help to simulate an independent data set and get a better assessment of a model’s predictive performance.

SUMMERY OF THE RESULTS:

After looking on the results between 10 first and 100 we can see that in the 10 first we had a small dataset,

The correctly and incorrectly classified instances show the percentage of test instances that were correctly and incorrectly classified. The raw numbers are shown in the confusion matrix, with a and b representing the class labels.

In the 100 we had 14047 instances, so the percentages and raw numbers add up, aa + bb = 403 + 12679 =13082, ab + ba = 31 + 934 = 965.

In the 10 first we had 54 instances, so the percentages and raw numbers add up, aa + bb = 3 + 3 =6, ab + ba = 9 + 39 = 48.

The percentage of correctly classified instances is often called accuracy or sample accuracy. It has some disadvantages as a performance estimate (not chance corrected, not sensitive to class distribution)

We can see that in the 10 first we had 77% and in 100 we had 93% which indicates that in bigger dataset we have more accuracy then in 10 first.

* **TP Rate**: rate of true positives (instances correctly classified as a given class)

**First 10**: 0.778 ,100: 0.931

* **FP Rate**: rate of false positives (instances falsely classified as a given class)

**First 10**: 0.599 , 100: 0.632

* **Precision**: proportion of instances that are truly of a class divided by the total instances classified as that class

**First 10**:0.743 , 100: 0.931

* **Recall**: proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate)

**First 10**: 0.778 , 100: 0.931

* **F-Measure**: A combined measure for precision and recall calculated as 2 \* Precision \* Recall / (Precision + Recall)

**First 10**: 0.748 , 100: 0.915

**ROC:** First 10: 0.822 , 100: 0.897 (0.5 indicates on random guessing and closely to 1 indicates on  "optimal" classifier )

We didn’t applied the *stemmer* which can cause the results be less good that expected and increase the error rates

Results for all files:

=== Summary ===

Correctly Classified Instances 12990 92.4753 %

Incorrectly Classified Instances 1057 7.5247 %

Kappa statistic 0.4405

Mean absolute error 0.1127

Root mean squared error 0.2644

Relative absolute error 65.3832 %

Root relative squared error 90.1052 %

Total Number of Instances 14047

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.360 0.016 0.705 0.360 0.476 0.470 0.721 0.400 True

0.984 0.640 0.936 0.984 0.959 0.470 0.721 0.939 False

Weighted Avg. 0.925 0.581 0.914 0.925 0.913 0.470 0.721 0.888

=== Confusion Matrix ===

a b <-- classified as

481 856 | a = True

201 12509 | b = False

True-Negative:

1402 1:True 2:False + 0.974 (vest-commodity)

1403 1:True 2:False + 0.952 (wasp-insect)

1399 1:True 2:False + 0.976 (fork-tool)

1400 1:True 2:False + 1 (glider-artifact)

1397 1:True 2:False + 0.976 (salmon-vertebrate)

1393 1:True 2:False + 1 (blouse-clothes)

1391 1:True 2:False + 0.719 (cow-beast)

1387 1:True 2:False + 0.931 (elm-tree)

1388 1:True 2:False + 0.979 (piano-furniture)

1385 1:True 2:False + 1 (cannon-arm)

False-Negative:

609 2:False 1:True + 1 (desk-table)

637 2:False 1:True + 1 (glove-socks)

719 2:False 1:True + 1 (hotel-show)

787 2:False 1:True + 0.938 (cello-bridge)

795 2:False 1:True + 1 (missile-engine)

844 2:False 1:True + 1 (library-outside)

845 2:False 1:True + 1 (gun-barrel)

907 2:False 1:True + 1 (dishwasher-pan)

982 2:False 1:True + 0.75 (trumpet-cello)

980 2:False 1:True + 1 (train-effort)

True-Positive:

1376 1:True1:True 0.933 (beaver-vertebrate)

1377 1:True1:True 1 (rat-chordate)

1378 1:True1:True 0.929 (villa-building)

1379 1:True1:True 0.923 (jacket-covering)

1380 1:True1:True 1 (vest-good)

1381 1:True1:True 1 (donkey-creature)

1356 1:True1:True 1 (moth-animal)

1357 1:True1:True 1 (corn-produce)

1358 1:True1:True 1 (celery-veggie)

1359 1:True1:True 0.923 (pistol-artefact)

False-Positive:

4 2:False2:False 0.8 (corkscrew-outsourcing)

5 2:False2:False 0.972 (dolphin-coyote)

6 2:False2:False 0.972 (stereo-television)

7 2:False2:False 0.983 (hornet-body)

8 2:False2:False 0.972 (tanker-bus)

9 2:False2:False 0.962 (sparrow-pheasant)

10 2:False2:False 0.962 (fox-cow)

11 2:False2:False 0.919 (falcon-communion)

12 2:False2:False 0.972 (poplar-cypress)

13 2:False2:False 0.919 (willow-totality)