جامعة بيرزيت

BIRZEIT UNIVERSITY

Faculty of Engineering & Technology
Electrical & Computer Engineering Department

ENCS3340

Project 2

Prepared by :

Tareq Shannak – 1181404

Abd Al-Rahman Mansour - 1182955

Instructor : Dr. Adnan Yahya

Section : 2

Date : 6/6/2021

Specifications

Tool used

We used WEKA 3.8.5 to preprocessing, classifying and simulate the testing results.

Test set used

According to the max ID in our group, we used test set number one which contains the first 10 authors with step size equals to 2.

Algorithms selection

First, we need to test these two algorithms: **Decision Trees** and **Artificial Neural Network (ANN)**. Also, there are two additional algorithms based on our IDs, and according to the IDs it seems that the **Random Forest** is the only algorithm that we need to test, so we chose an additional algorithm randomly which is **Naïve Bayes** algorithm.

Assumptions and Details

For all data files in preprocessing, we chose an unsupervised filter for attribute which is StringToWordVector, this filter converts the texts to a vector of words. Also, we modified this filter by enable IDF and TFT transforms that give a value for each word to represents its importance according to the frequency for this word in the text. We put IteratedlovinsStemmer as a stemmer and MultiStopwords as a stop words handler, where the stemmer algorithm used on the words. The words to keep differ from learning algorithm to another according to Table 1 after some experiments in improving results.

In classifying the data, we chose the class that we want to test which is the names of authors. The name of classifier and the percentage split between the training and test sets differ between learning algorithms according to Table 1 after some experiments in improving results.

	Words To Keep	Percentage Split
Decision Tree	300 Words	85.0%
ANN	15 Words	85.0%
Random Forest	1000 Words	66.0%
Naïve Bayes	500 Words	66.0%

Results

Decision Tree

2 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
1-2	73.6	26.4	0.736	0.233	0.774	0.613	0.721
3-4	87.0	13.0	0.870	0.138	0.872	0.870	0.869
5-6	80.9	19.1	0.810	0.277	0.811	0.810	0.803
7-8	85.7	14.3	0.857	0.134	0.865	0.857	0.857
9-10	85.1	14.9	0.851	0.166	0.850	0.851	0.850
Avg.	**82.5**	**17.5**	**0.825**	**0.19**	**0.834**	**0.8**	**0.82**

4 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
1-4	70.4	29.6	0.704	0.096	0.732	0.704	0.705
5-8	71.9	28.1	0.719	0.101	0.737	0.719	0.720
7-10	76.1	23.9	0.761	0.088	0.765	0.761	0.761
Avg.	**72.8**	**27.2**	**0.728**	**0.095**	**0.745**	**0.728**	**0.729**

6 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
1-6	60.5	39.5	0.605	0.084	0.606	0.605	0.605
5-10	65.7	34.3	0.657	0.072	0.656	0.657	0.655
Avg.	**63.1**	**36.9**	**0.631**	**0.078**	**0.631**	**0.631**	**0.63**

8 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
1-8	60.2	39.8	0.602	0.058	0.613	0.602	0.606
3-10	63.8	36.2	0.638	0.056	0.651	0.638	0.640
Avg.	**62**	**38**	**0.62**	**0.057**	**0.632**	**0.62**	**0.623**

10 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
10	**56.2**	**43.8**	**0.562**	**0.051**	**0.564**	**0.562**	**0.560**

Artificial Neural Network

2 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
1-2	79.2	20.8	0.792	0.246	0.813	0.792	0.783
3-4	75.1	24.9	0.751	0.265	0.756	0.751	0.748
5-6	75.8	24.2	0.758	0.326	0.752	0.758	0.751
7-8	82.1	17.9	0.821	0.169	0.831	0.821	0.821
9-10	78.8	21.2	0.788	0.232	0.787	0.788	0.788
Avg.	**78.2**	**21.8**	**0.782**	**0.248**	**0.788**	**0.782**	**0.778**

4 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
1-4	64.4	35.6	0.644	0.111	0.697	0.644	0.647
5-8	63.5	36.5	0.635	0.129	0.678	0.635	0.639
7-10	67.4	32.6	0.674	0.124	0.677	0.674	0.669
Avg.	**65.1**	**34.9**	**0.651**	**0.121**	**0.684**	**0.651**	**0.652**

6 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
1-6	49.8	50.2	0.498	0.103	0.574	0.498	0.506
5-10	54.0	46.0	0.540	0.105	0.565	0.540	0.535
Avg.	**51.9**	**48.1**	**0.519**	**0.104**	**0.57**	**0.519**	**0.521**

8 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
1-8	51.7	48.3	0.517	0.076	0.584	0.517	0.522
3-10	52.5	47.5	0.525	0.080	0.569	0.525	0.528
Avg.	**52.1**	**47.9**	**0.521**	**0.078**	**0.577**	**0.521**	**0.525**

10 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
10	**46.0**	**54.0**	**0.460**	**0.067**	**0.510**	**0.460**	**0.463**

Random Forest

2 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
1-2	93.7	6.3	0.937	0.062	0.937	0.937	0.937
3-4	93.1	6.9	0.931	0.074	0.933	0.931	0.931
5-6	91.0	9.0	0.910	0.107	0.910	0.910	0.910
7-8	95.3	4.7	0.953	0.060	0.954	0.953	0.953
9-10	91.5	8.5	0.915	0.083	0.916	0.915	0.915
Avg.	**92.9**	**7.1**	**0.929**	**0.077**	**0.93**	**0.929**	**0.929**

4 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
1-4	83.1	16.9	0.831	0.065	0.849	0.831	0.834
5-8	84.9	15.1	0.849	0.056	0.850	0.849	0.846
7-10	88.0	12.0	0.880	0.045	0.881	0.880	0.879
Avg.	**85.3**	**14.7**	**0.853**	**0.055**	**0.86**	**0.853**	**0.853**

6 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
1-6	74.2	25.8	0.742	0.060	0.748	0.742	0.742
5-10	77.7	22.3	0.777	0.050	0.782	0.777	0.772
Avg.	**76**	**24.1**	**0.76**	**0.055**	**0.765**	**0.76**	**0.757**

8 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
1-8	72.6	27.4	0.726	0.042	0.727	0.726	0.725
3-10	74.7	25.3	0.747	0.040	0.752	0.747	0.745
Avg.	**73.7**	**26.3**	**0.737**	**0.041**	**0.74**	**0.737**	**0.735**

10 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
10	**68.8**	**31.2**	**0.688**	**0.038**	**0.696**	**0.688**	**0.684**

Naïve Bayes

2 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
1-2	89.0	11.0	0.890	0.113	0.890	0.890	0.889
3-4	87.8	12.2	0.878	0.123	0.878	0.878	0.878
5-6	86.4	13.6	0.864	0.162	0.863	0.864	0.864
7-8	90.2	9.8	0.902	0.091	0.906	0.902	0.903
9-10	89.4	10.6	0.894	0.102	0.897	0.894	0.784
Avg.	**88.6**	**11.4**	**0.886**	**0.118**	**0.887**	**0.886**	**0.864**

4 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
1-4	77.9	22.1	0.779	0.072	0.784	0.779	0.779
5-8	75.9	24.1	0.759	0.080	0.766	0.759	0.761
7-10	82.8	17.2	0.828	0.061	0.843	0.828	0.830
Avg.	**78.9**	**21.1**	**0.789**	**0.071**	**0.798**	**0.789**	**0.79**

6 Authors

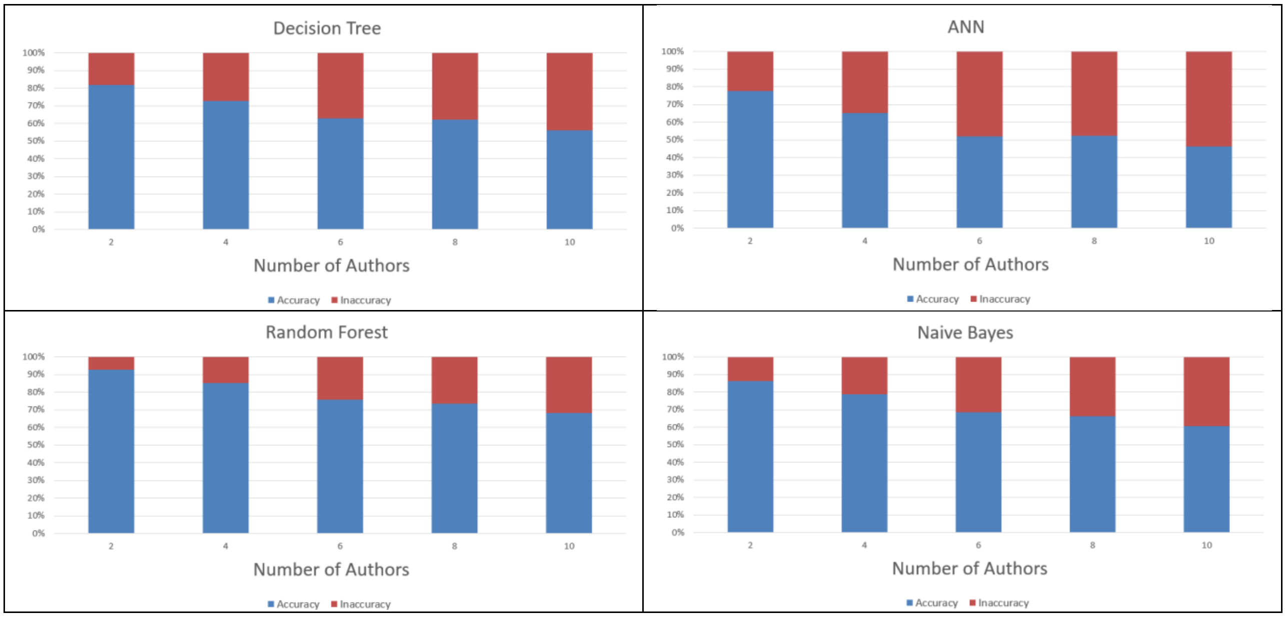
	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
1-6	67.0	33.0	0.670	0.068	0.674	0.670	0.670
5-10	69.9	30.1	0.699	0.060	0.718	0.699	0.701
Avg.	**68.5**	**31.5**	**0.685**	**0.064**	**0.696**	**0.685**	**0.686**

8 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
1-8	65.4	34.6	0.654	0.049	0.662	0.654	0.653
3-10	67.2	32.8	0.672	0.049	0.686	0.672	0.674
Avg.	**66.3**	**33.7**	**0.663**	**0.049**	**0.674**	**0.663**	**0.664**

10 Authors

	Accuracy (%)	Inaccuracy (%)	TP Rate	FP Rate	Precision	Recall	F Measure
10	**60.8**	**39.2**	**0.608**	**0.045**	**0.623**	**0.608**	**0.608**



Conclusion

We can notice that when the training set increases, the testing set give us much more positive results. Also when we increases the words, the results become better but the space will increase. Hence in ANN learning algorithm, the space will increase horribly, so the number of words that we chose is very low (15 words) and in all cases it give us good results even if the number of words are low.

Random Forest is the best learning algorithm based on our data in selecting the author for each document. Where the accuracy didn't be less than 68% in all test cases.