Data Analysis of Pooled Dataset

We carried out dataset analysis of pooled dataset as follows.

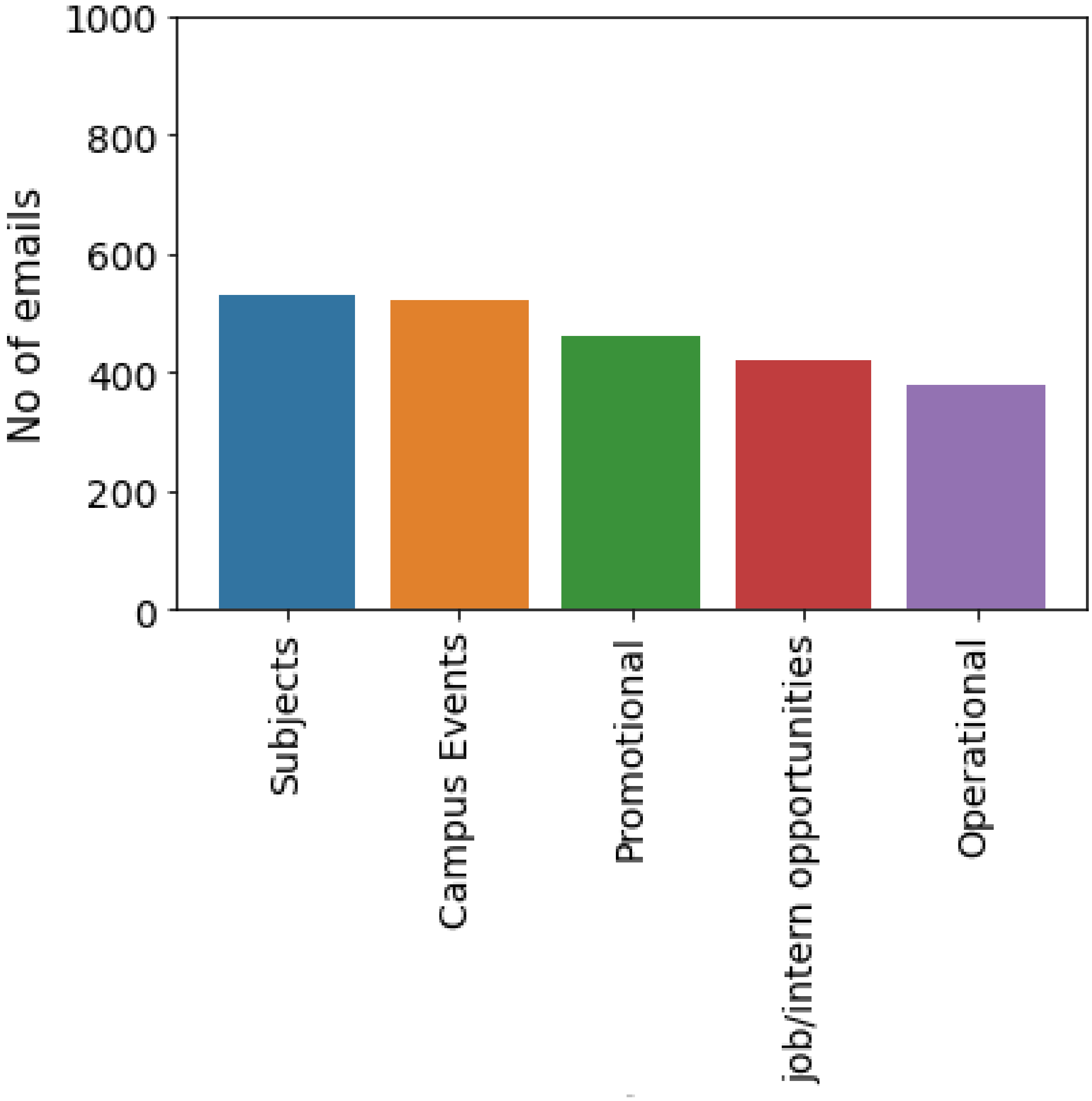


Fig. 1 Number of Emails for each label

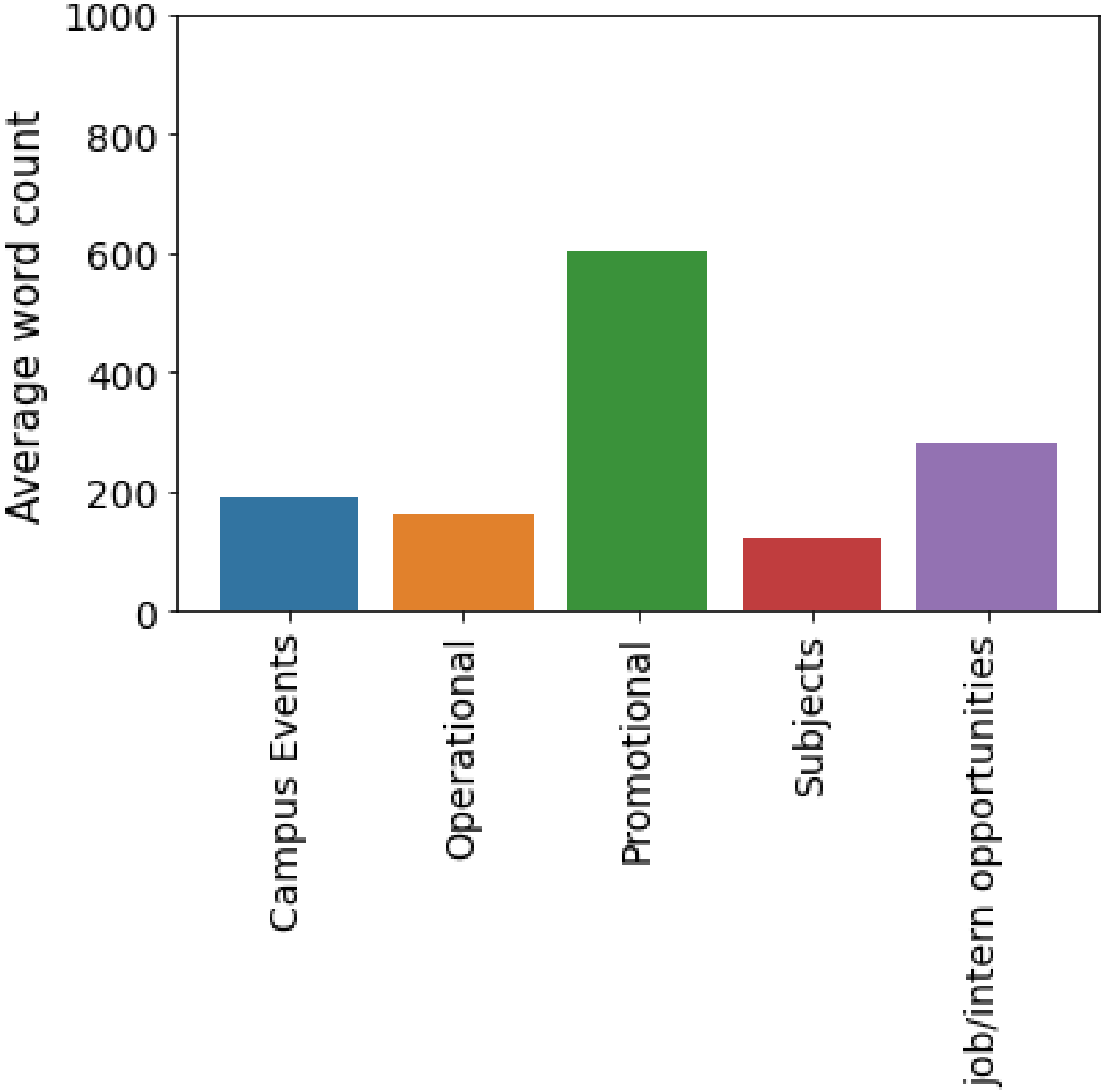


Fig. 2 Average word count for each label

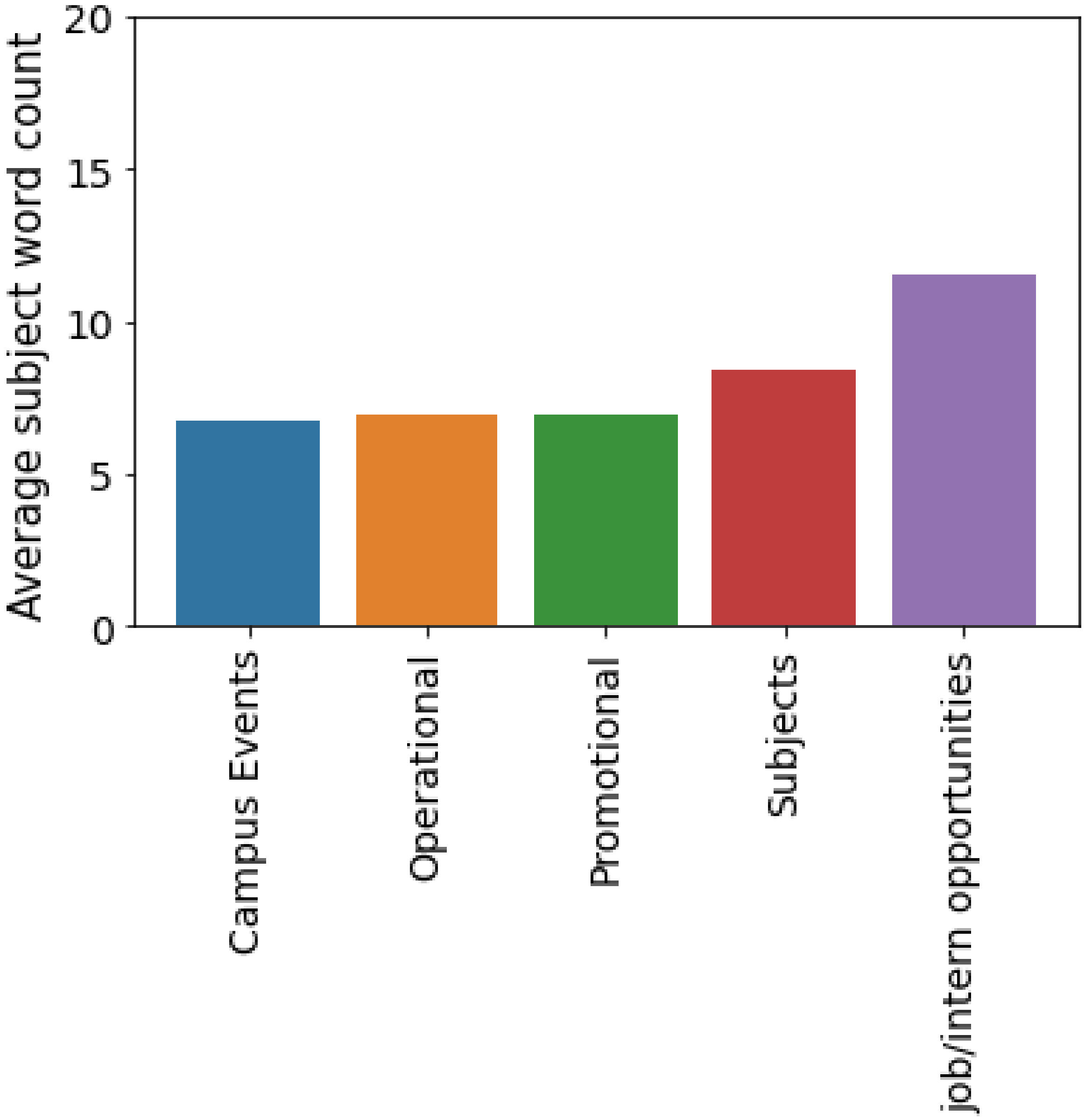


Fig. 3 Average subject word count for each label

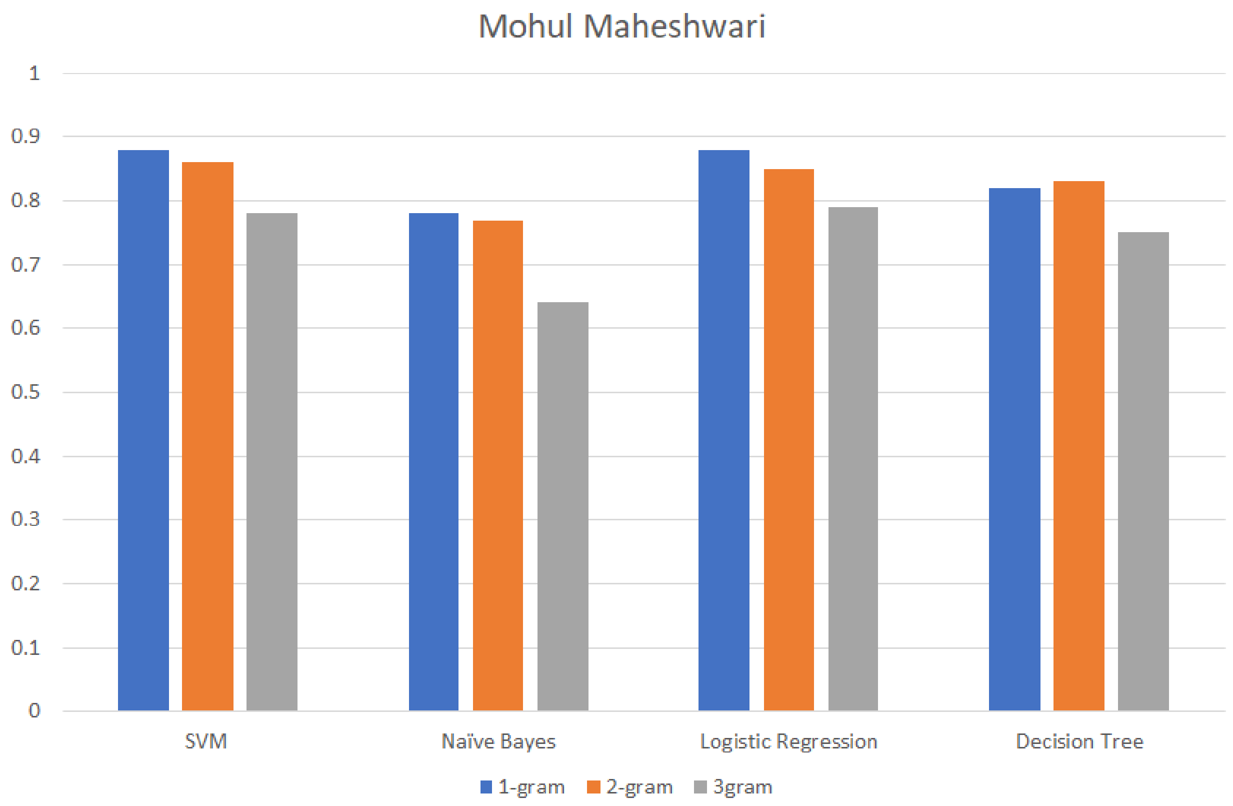
Comparison Study

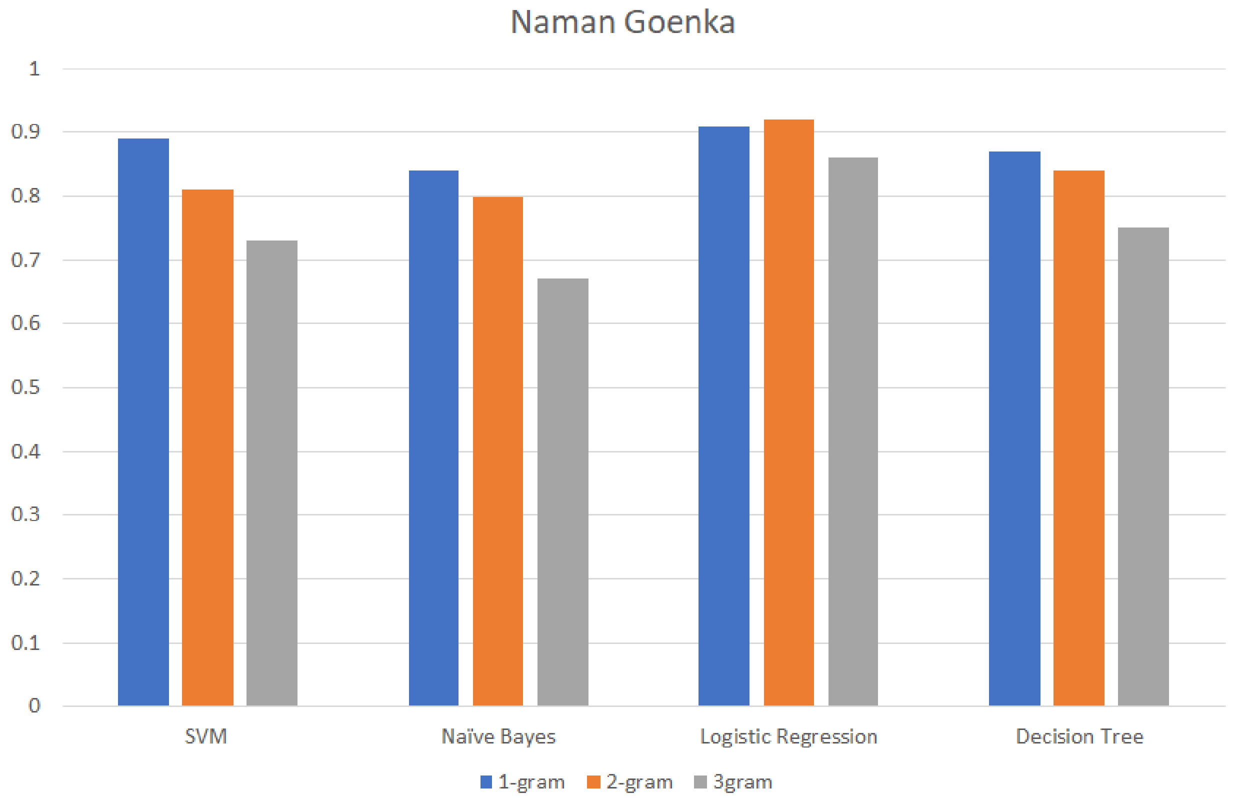
Here we compare different Machine Learning algorithms using 3 different N-gram language models.

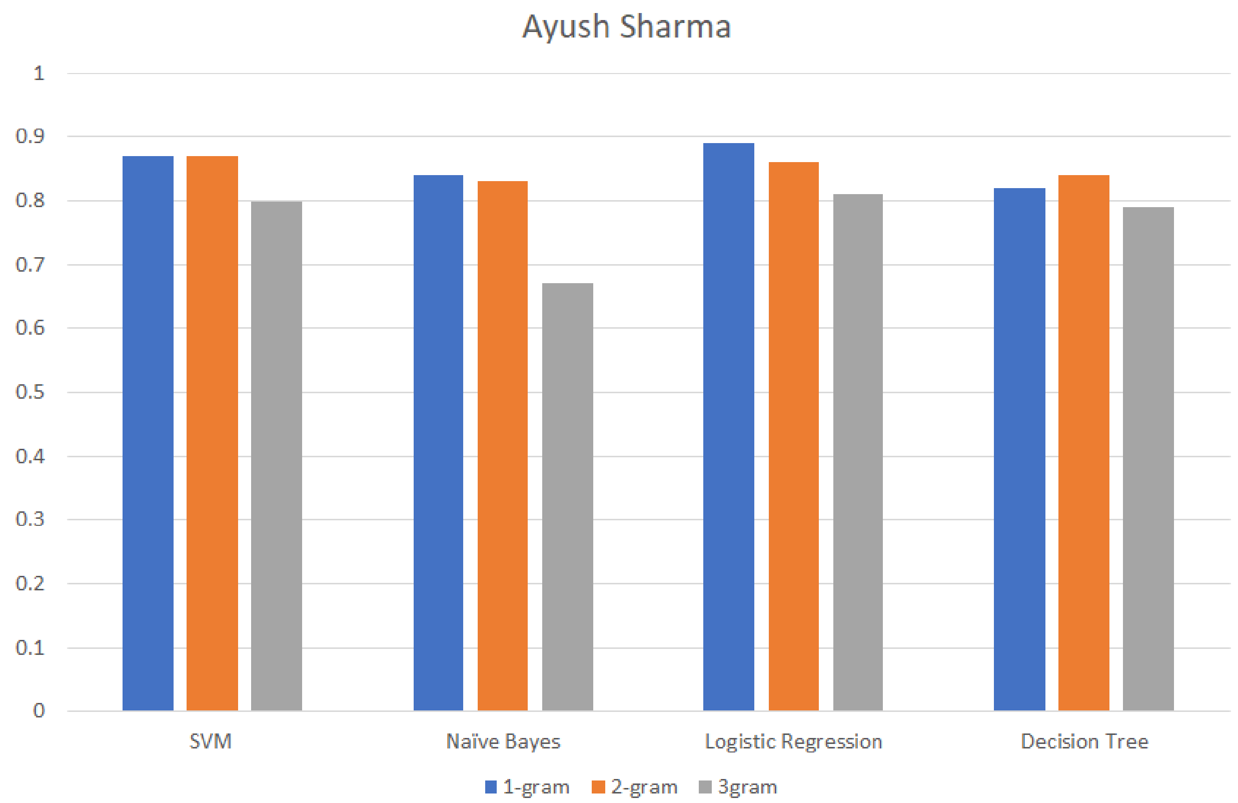
Point to note regarding LDA, we tried it but the system crashed for 2-gram and 3-gram dataset on Google Colab.

The values that follow are weighted average F1 scores.

	1-gram	2-gram	3-gram
Ayush Sharma			
MultinomialNB	0.84	0.83	0.67
SVM	0.87	0.87	0.80
Logistic Regression	0.89	0.86	0.81
Decision Tree	0.82	0.84	0.75
Naman Goenka			
MultinomialNB	0.84	0.80	0.67
SVM	0.89	0.81	0.73
Logistic Regression	0.91	0.92	0.86
Decision Tree	0.87	0.84	0.79
Mohul Maheshwari			
MultinomialNB	0.78	0.77	0.64
SVM	0.88	0.86	0.78
Logistic Regression	0.88	0.85	0.79
Decision Tree	0.82	0.83	0.75


Mohul Maheshwari
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
SVM
Naïve Bayes
Logistic Regression
Decision Tree
1-gram
2-gram
3gram

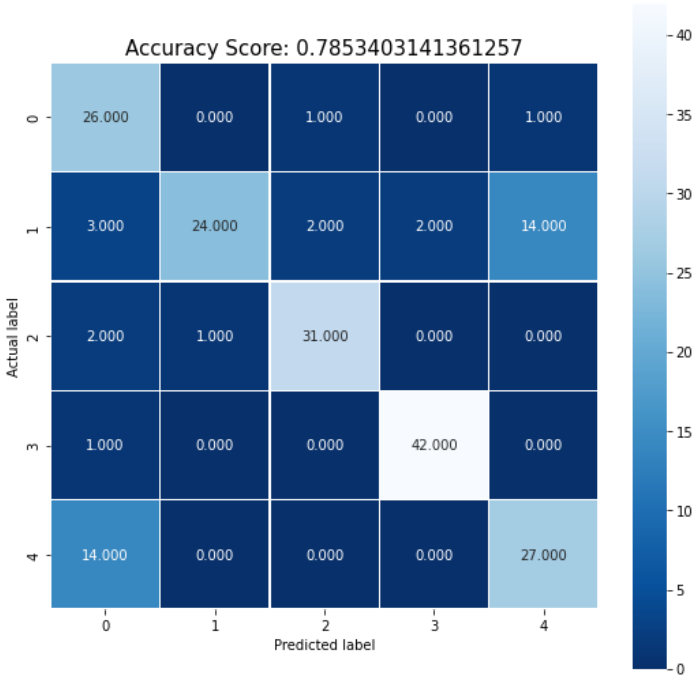

Naman Goenka
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
SVM
Naïve Bayes
Logistic Regression
Decision Tree
1-gram
2-gram
3gram


Ayush Sharma
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
SVM
Naïve Bayes
Logistic Regression
Decision Tree
1-gram
2-gram
3gram

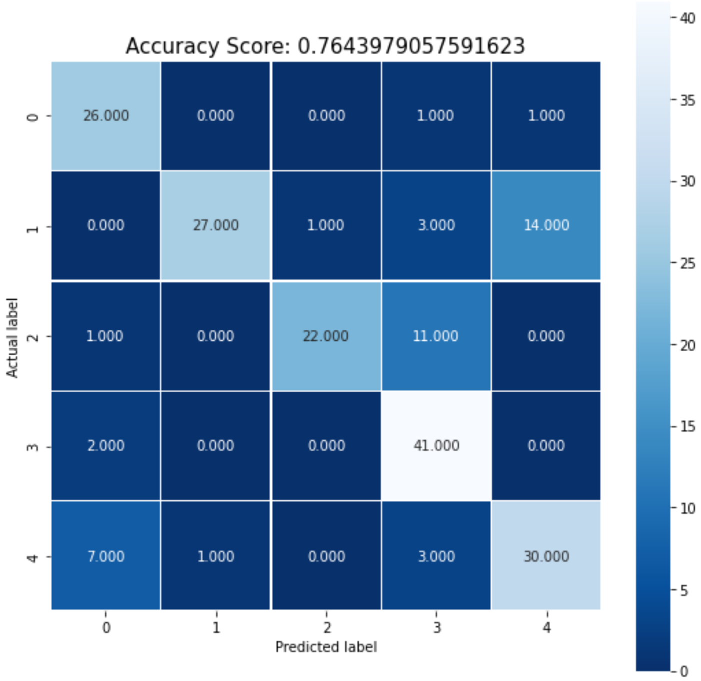
Data_1

MultinomialNB

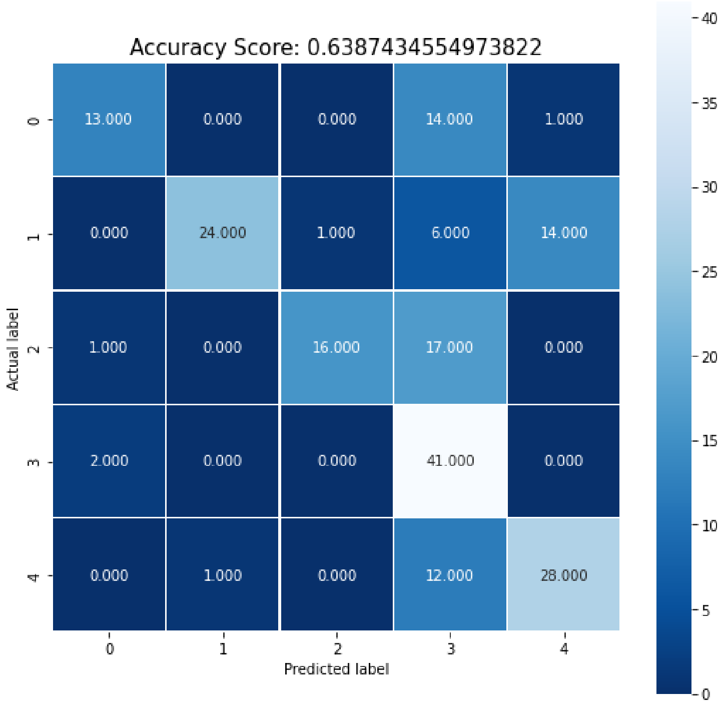
Unigram



Bigram

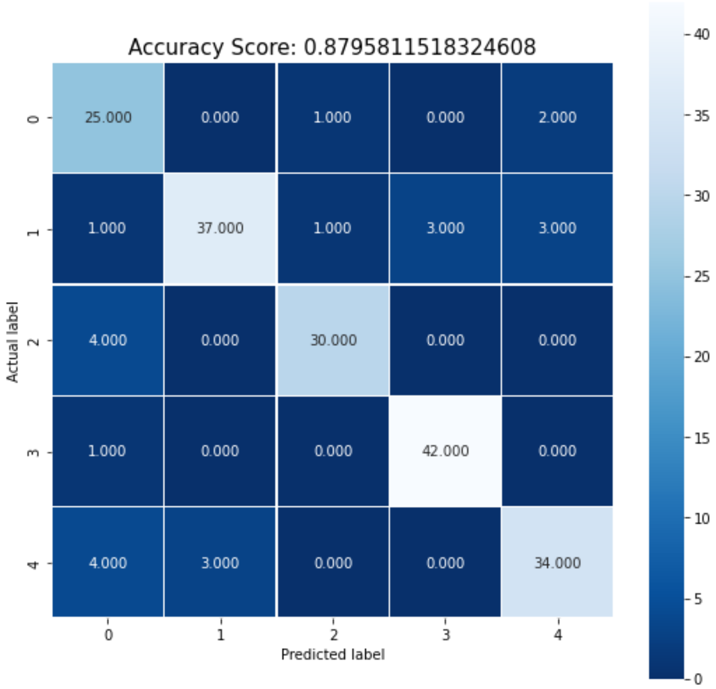


Trigram

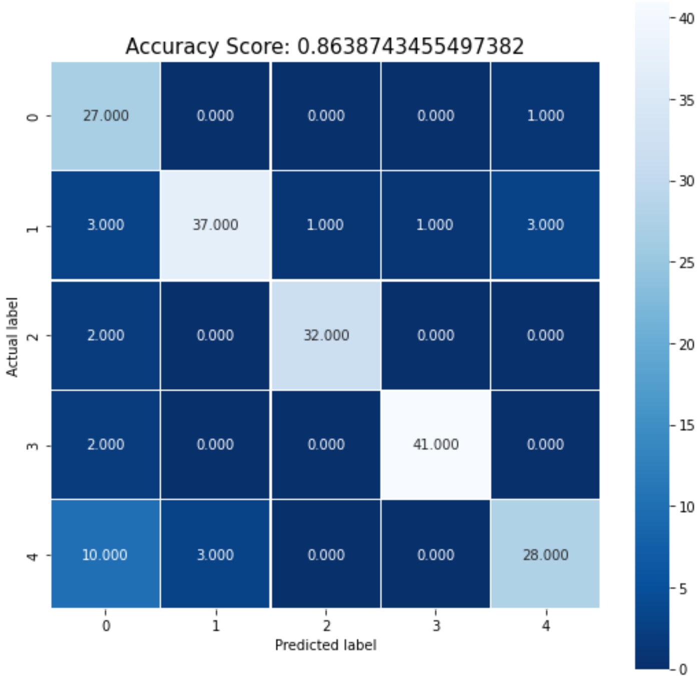


SVM

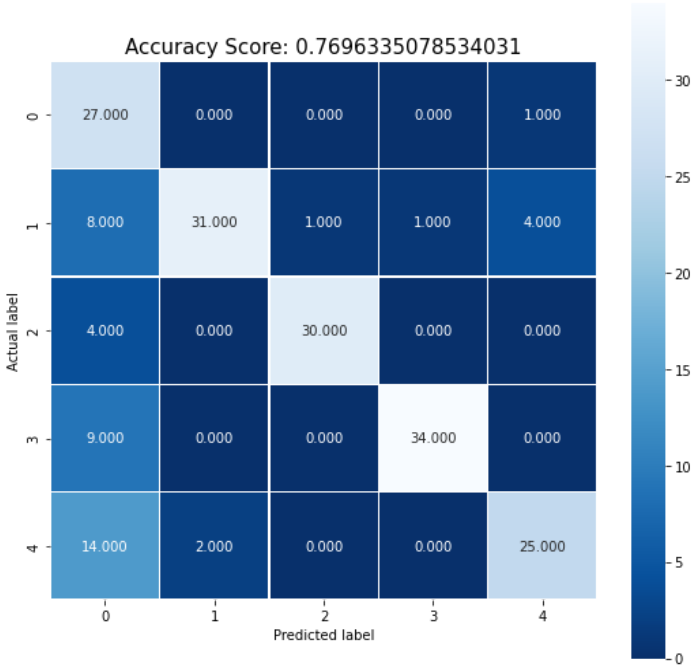
Unigram



Bigram

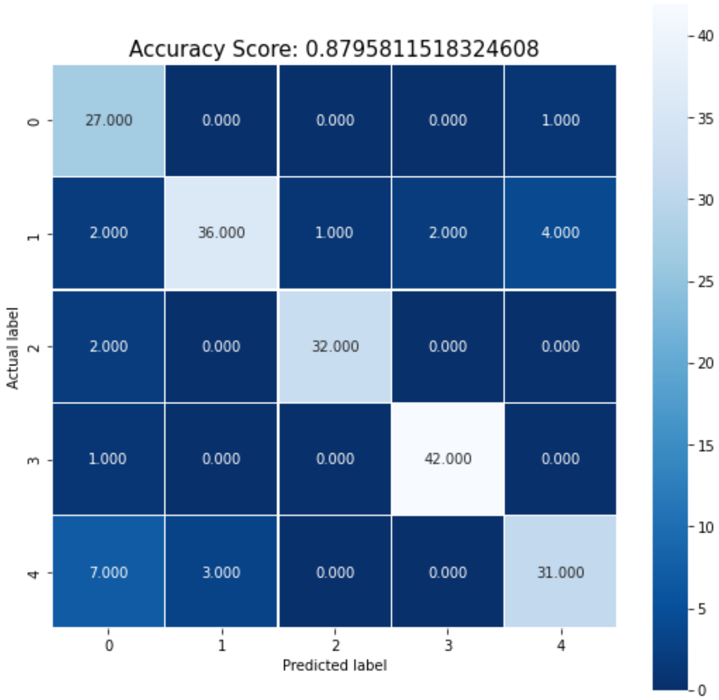


Trigram

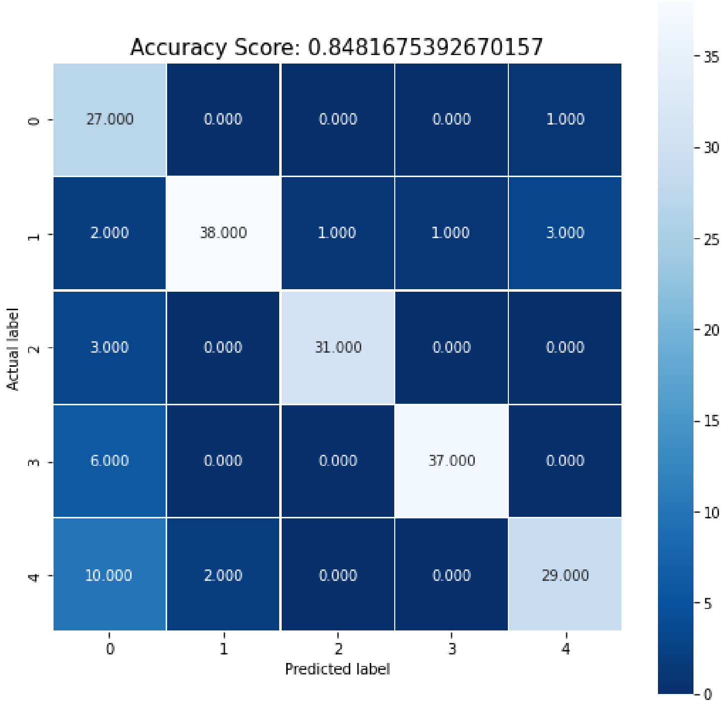


Logistic Regression

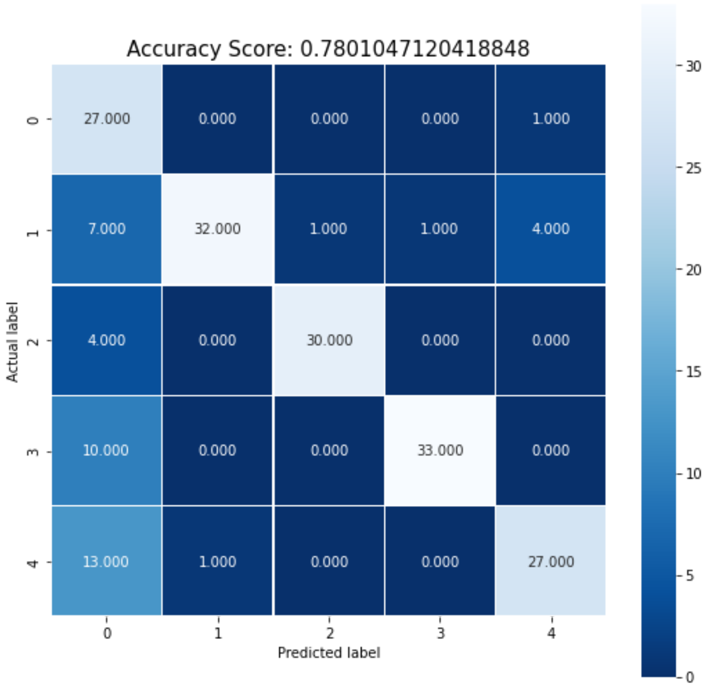
Unigram



Bigram

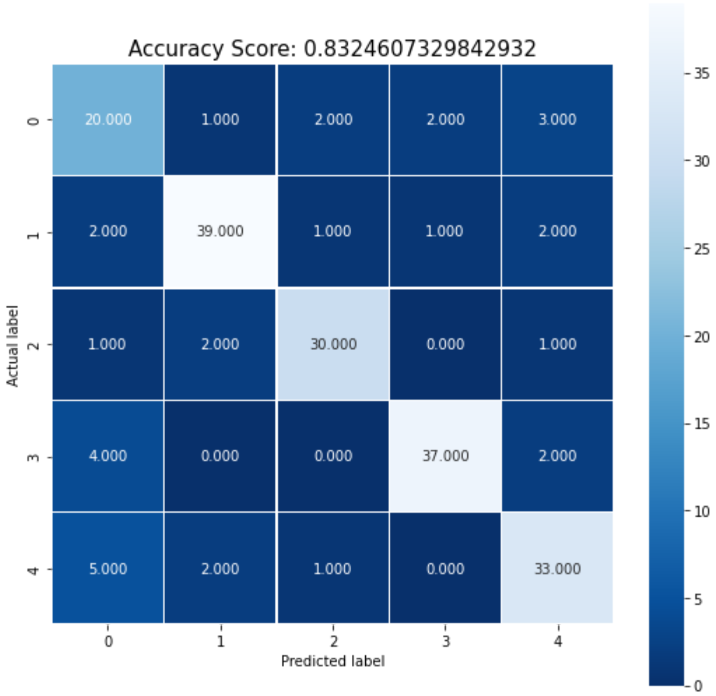


Trigram

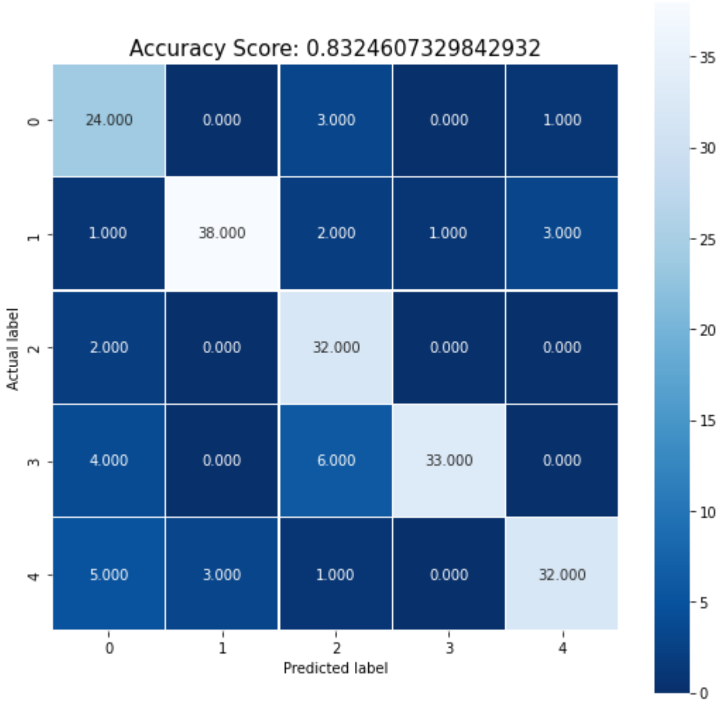


Decision Tree

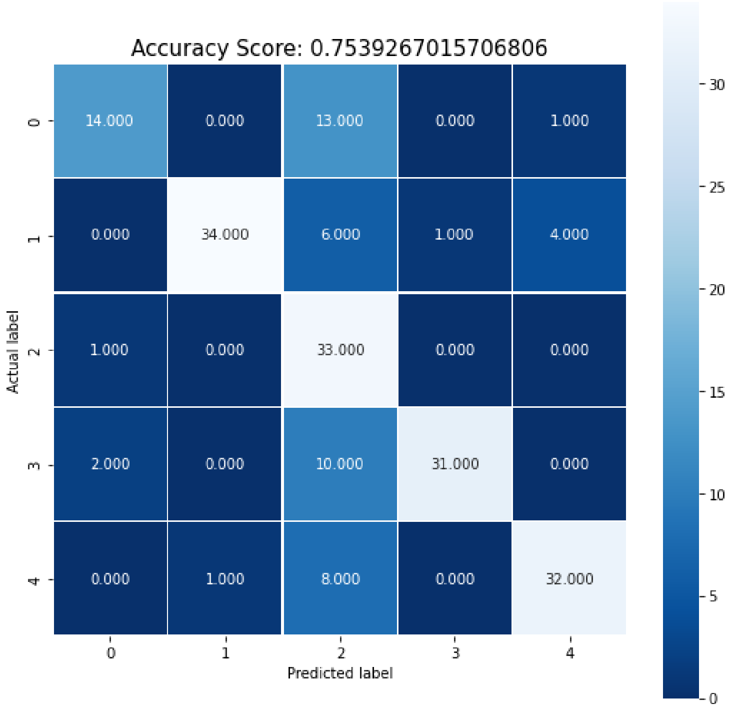
Unigram



Bigram



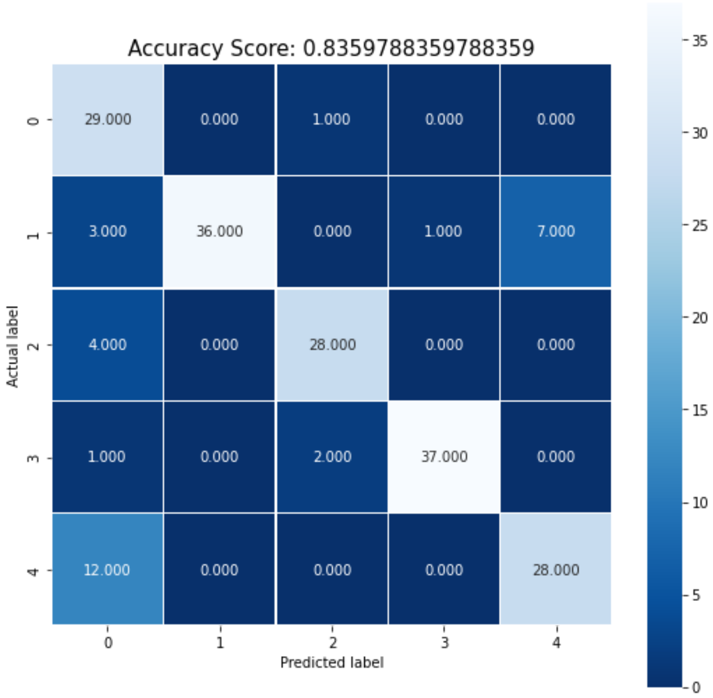
Trigram



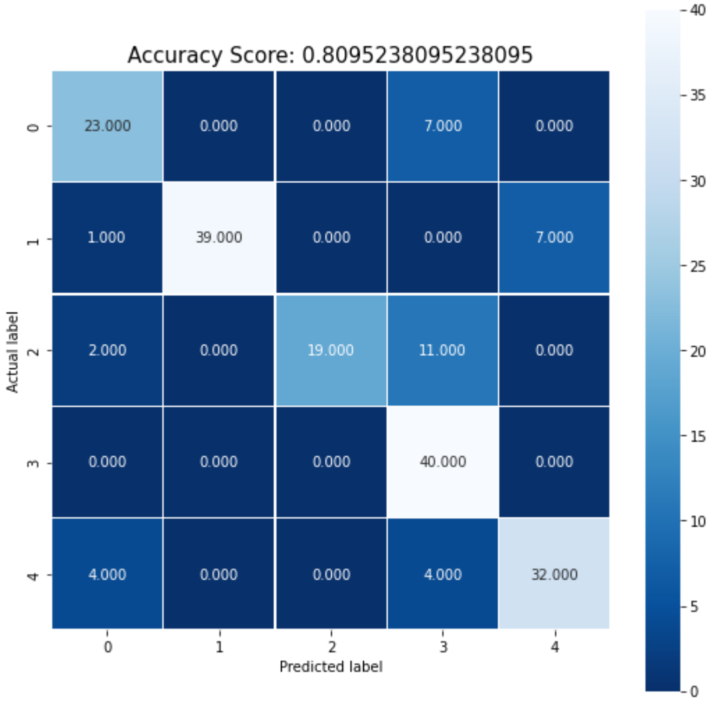
Data_2

MultinomialNB

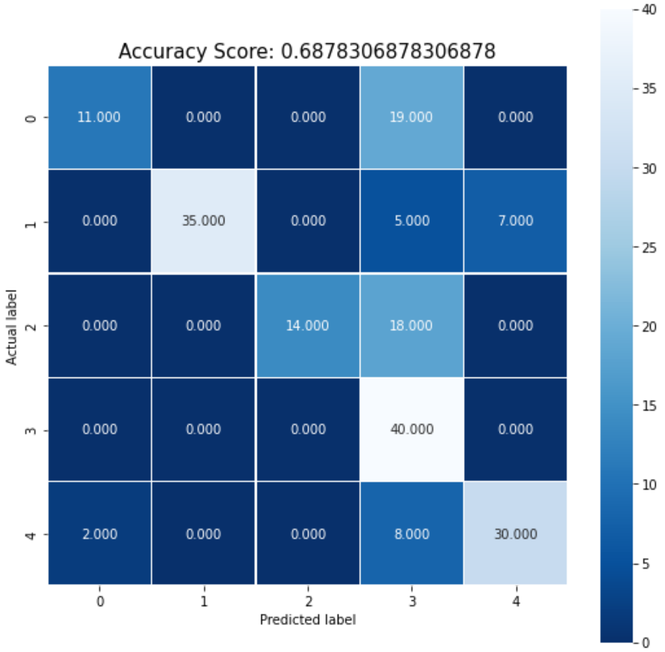
Unigram



Bigram



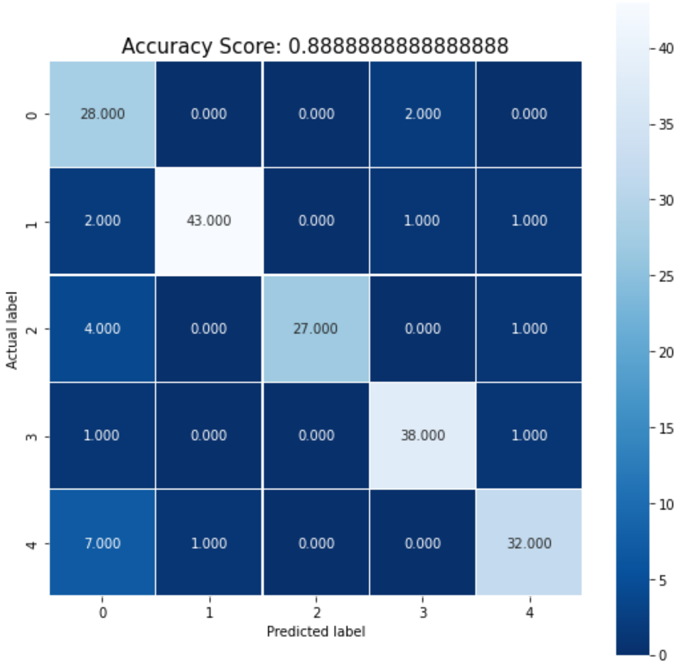
Trigram



	0	1	2	3	4
0	11.000	0.000	0.000	19.000	0.000
1	0.000	35.000	0.000	5.000	7.000
2	0.000	0.000	14.000	18.000	0.000
3	0.000	0.000	0.000	40.000	0.000
4	2.000	0.000	0.000	8.000	30.000

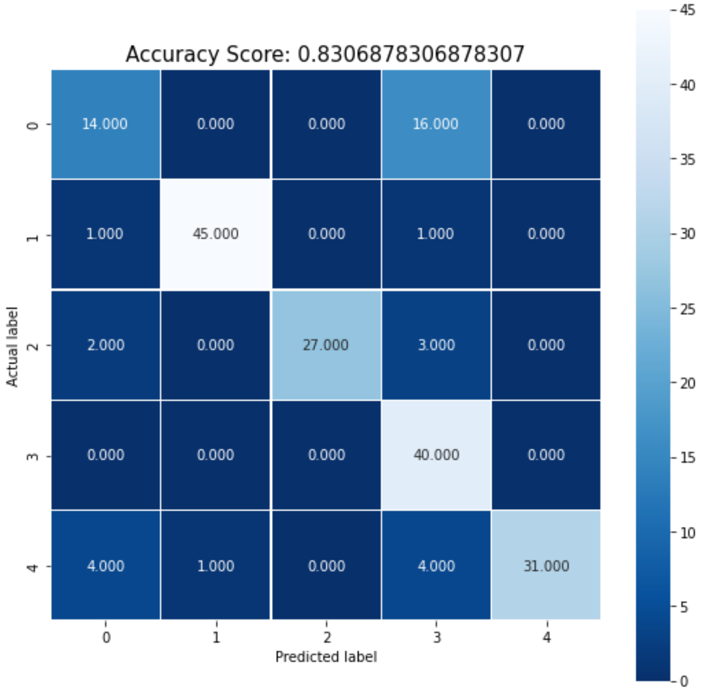
SVM

Unigram

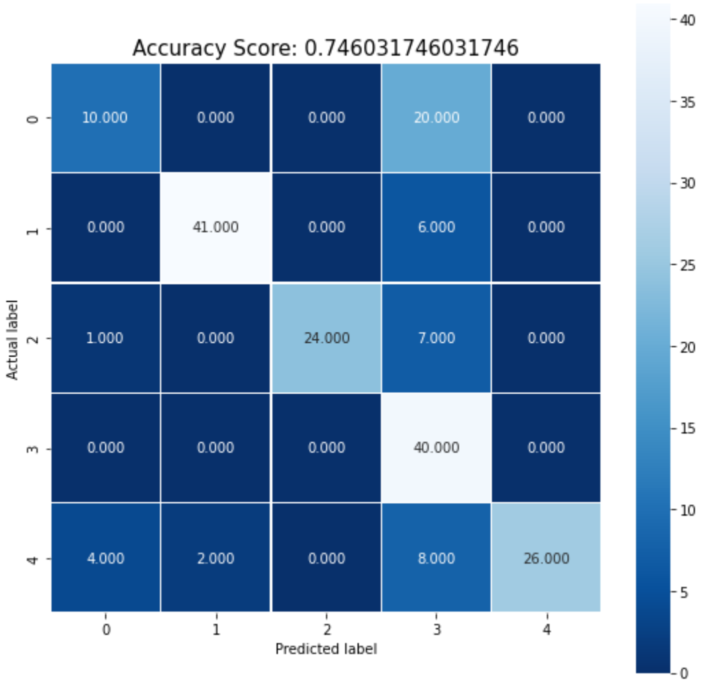


	0	1	2	3	4
0	28.000	0.000	0.000	2.000	0.000
1	2.000	43.000	0.000	1.000	1.000
2	4.000	0.000	27.000	0.000	1.000
3	1.000	0.000	0.000	38.000	1.000
4	7.000	1.000	0.000	0.000	32.000

Bigram

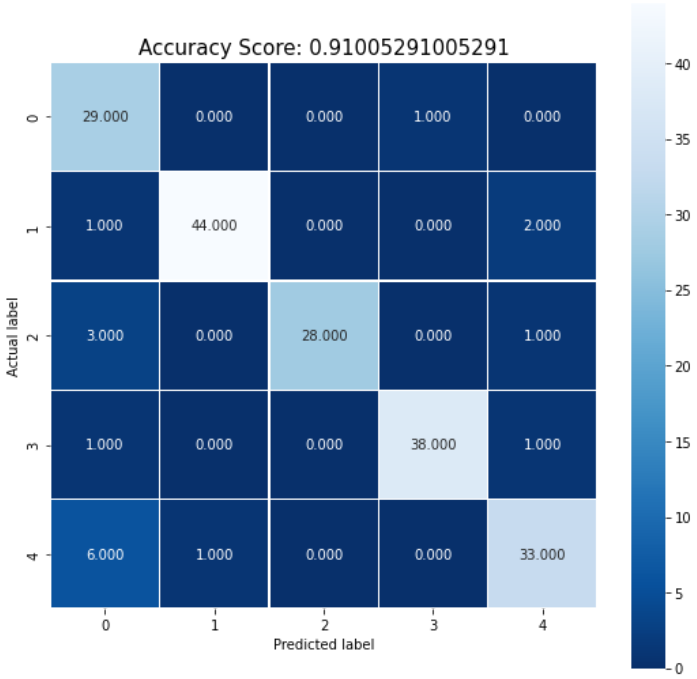


Trigram

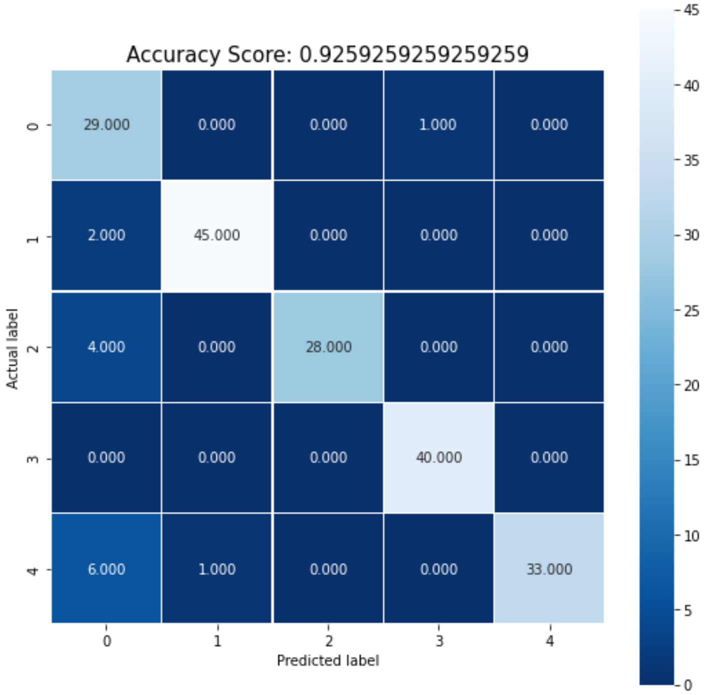


Logistic Regression

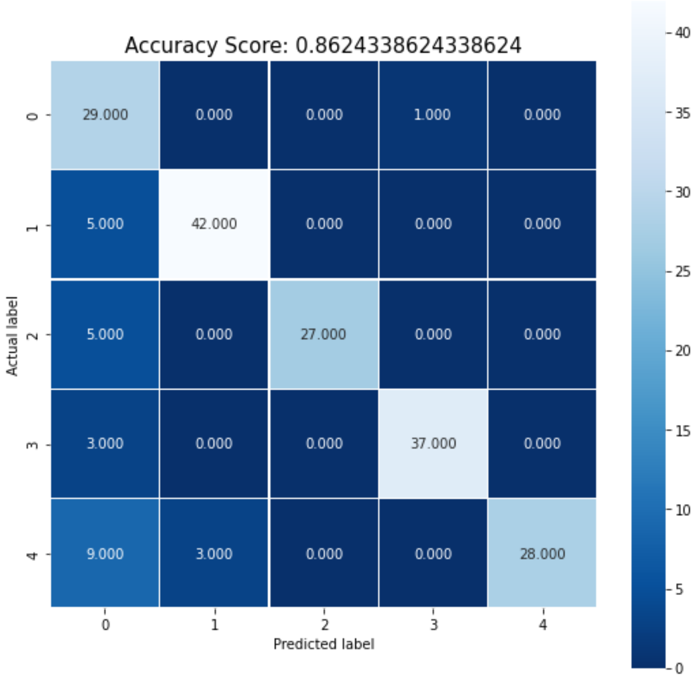
Unigram



Bigram

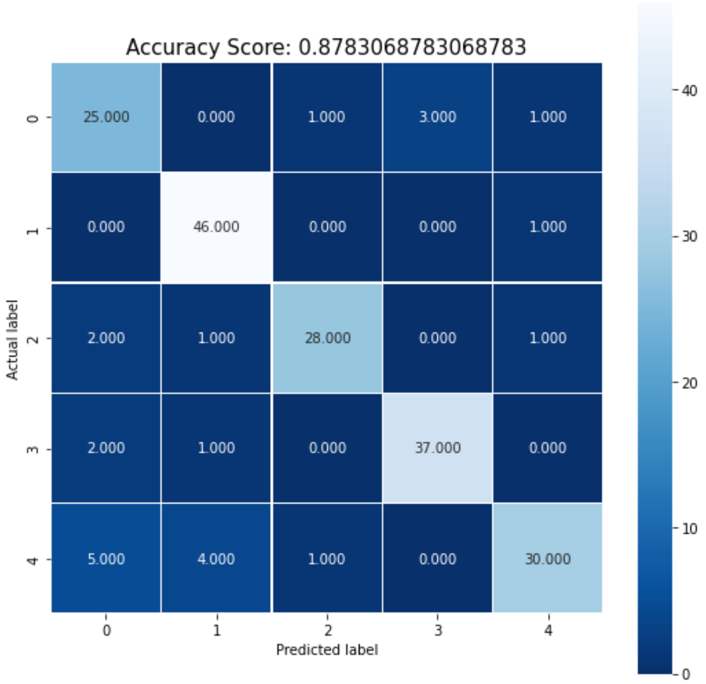


Trigram

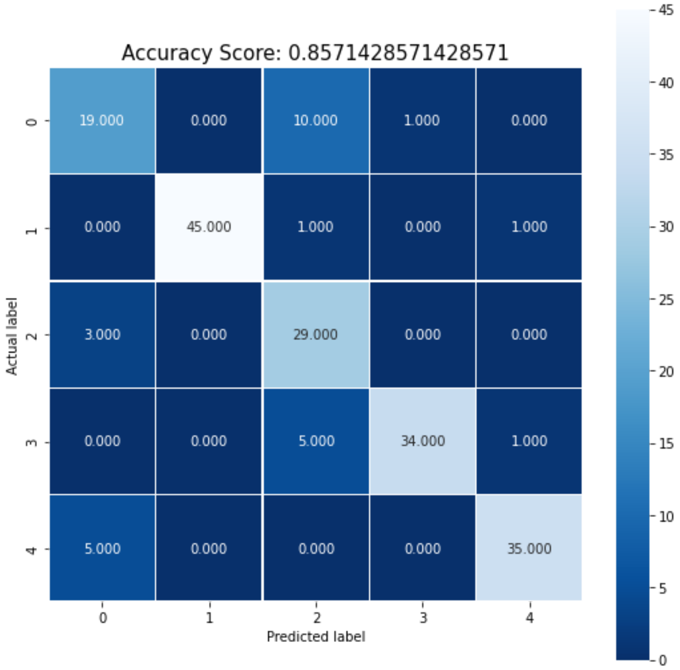


Decision Tree

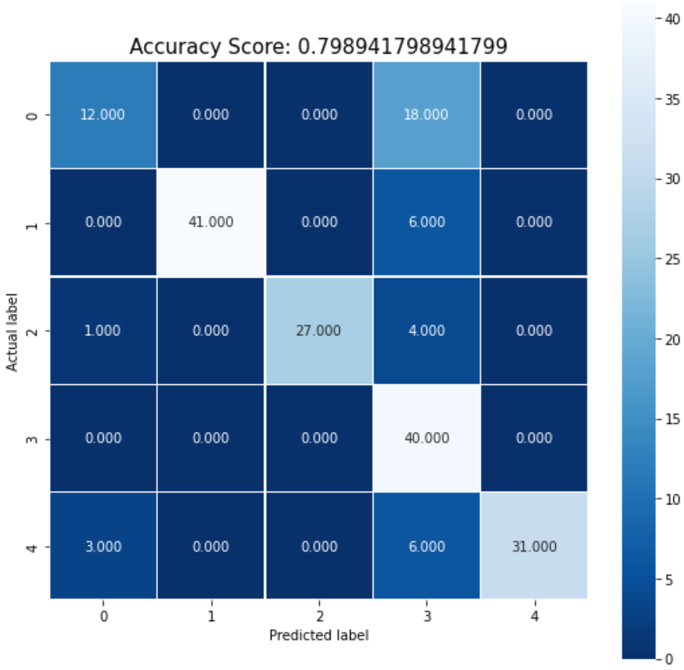
Unigram



Bigram



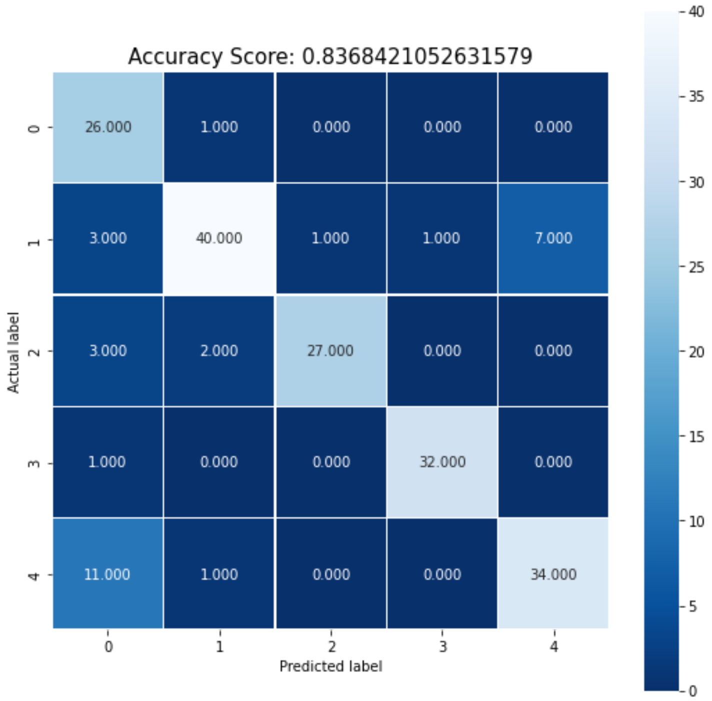
Trigram



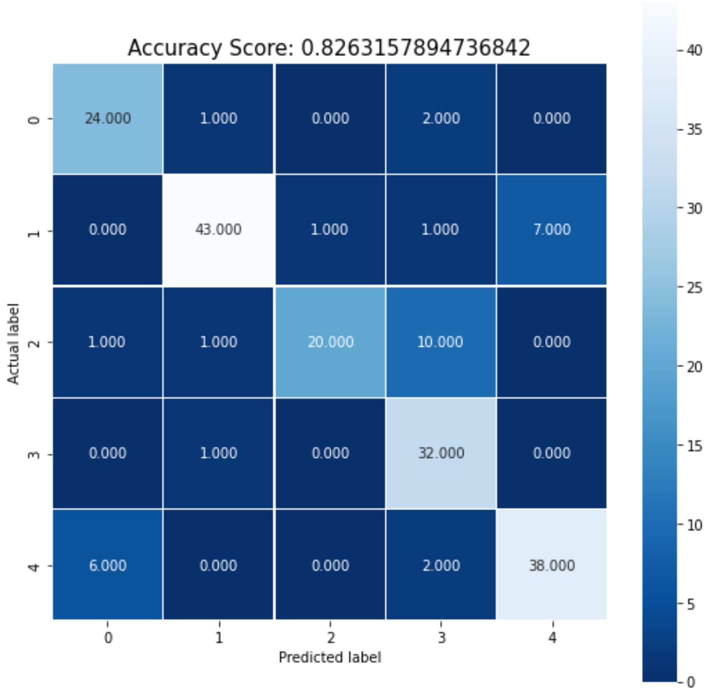
Data 3

MultinomialNB

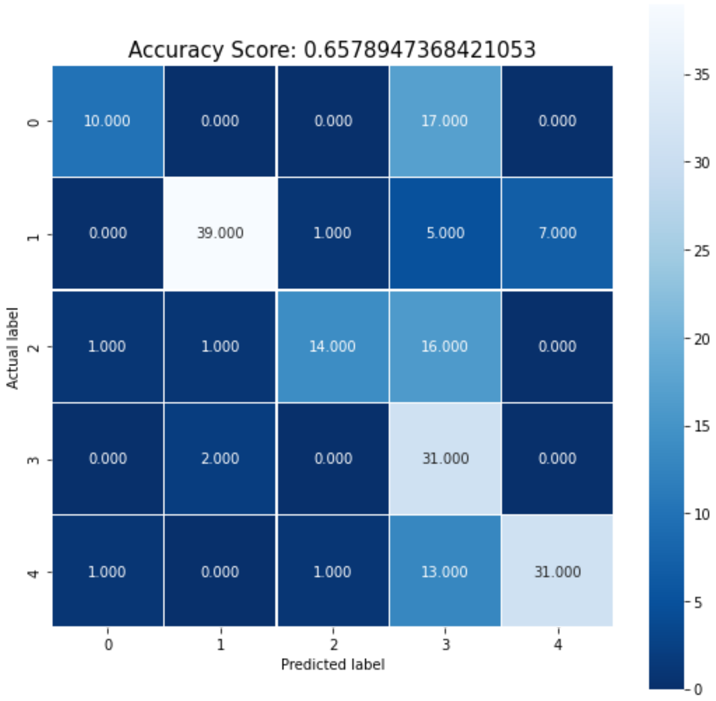
Unigram



Bigram



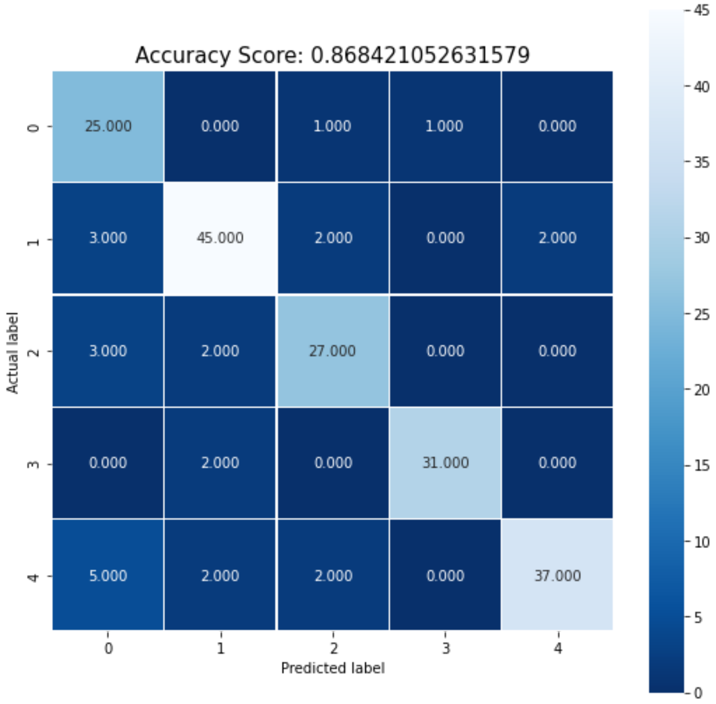
Trigram


Accuracy Score: 0.6578947368421053

Actual label \ Predicted label	0	1	2	3	4
0	10.000	0.000	0.000	17.000	0.000
1	0.000	39.000	1.000	5.000	7.000
2	1.000	1.000	14.000	16.000	0.000
3	0.000	2.000	0.000	31.000	0.000
4	1.000	0.000	1.000	13.000	31.000

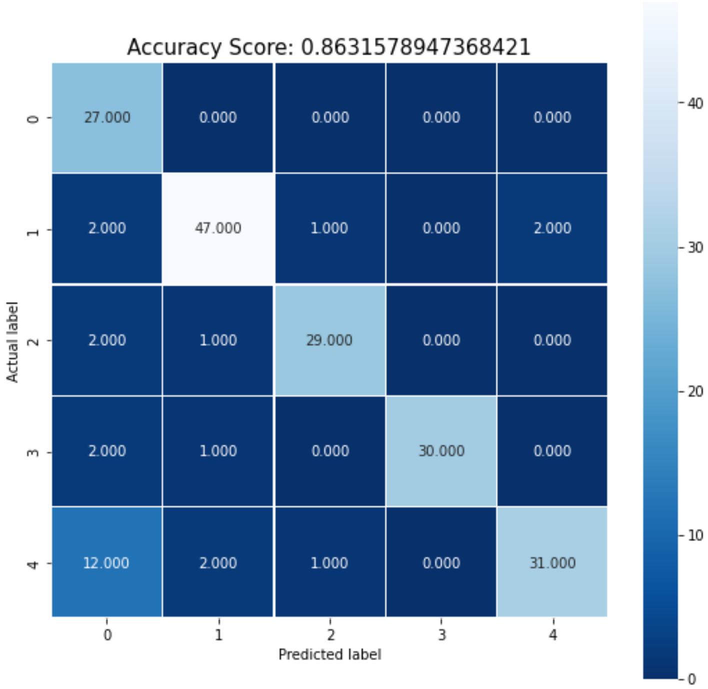
SVM

Unigram


Accuracy Score: 0.868421052631579

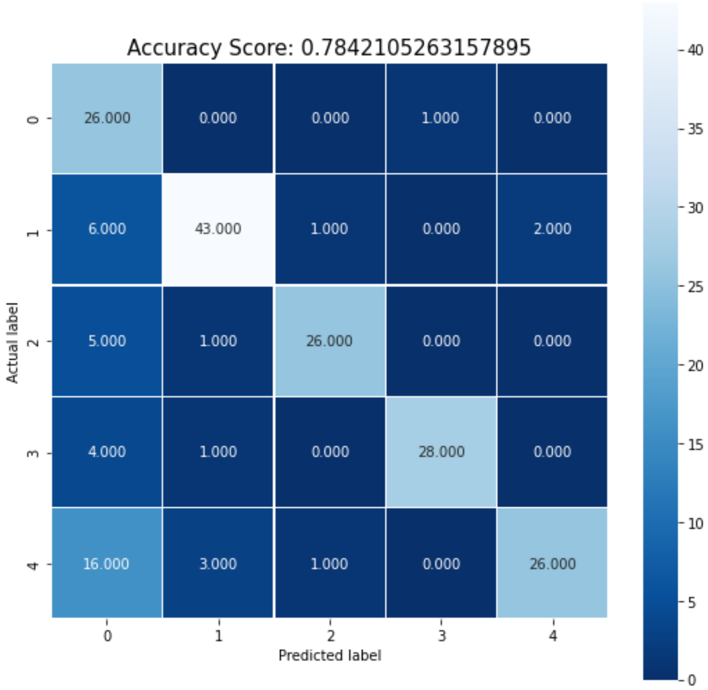
Actual label \ Predicted label	0	1	2	3	4
0	25.000	0.000	1.000	1.000	0.000
1	3.000	45.000	2.000	0.000	2.000
2	3.000	2.000	27.000	0.000	0.000
3	0.000	2.000	0.000	31.000	0.000
4	5.000	2.000	2.000	0.000	37.000

Bigram



Accuracy Score: 0.8631578947368421

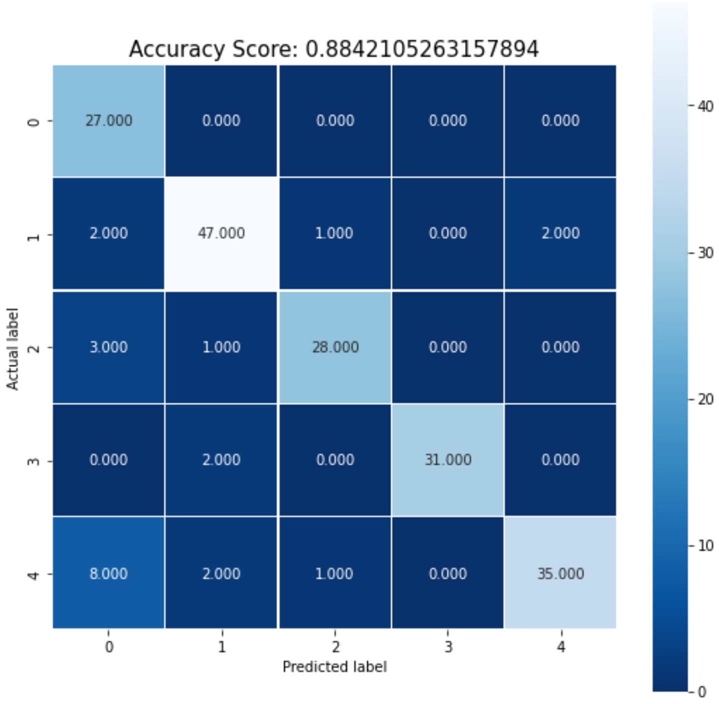
Trigram



Accuracy Score: 0.7842105263157895

Linear regression

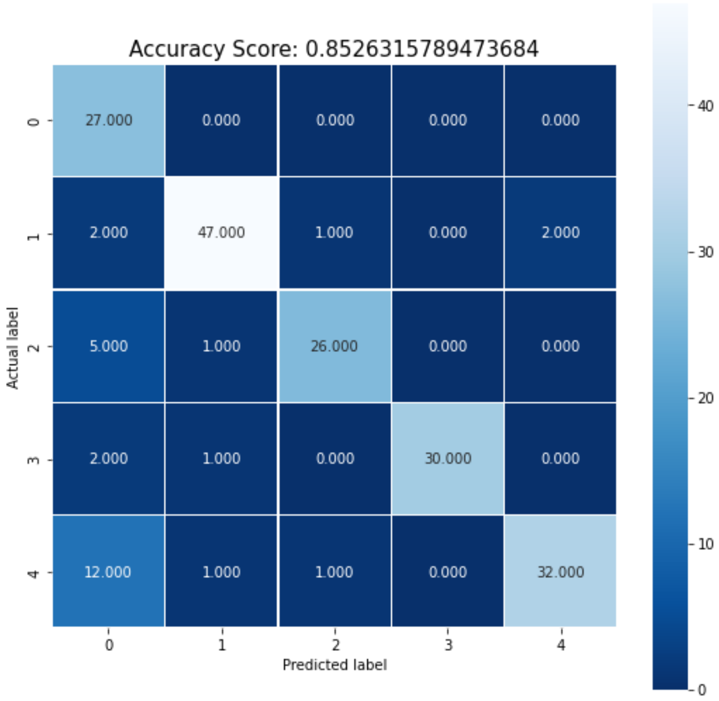
Unigram



Accuracy Score: 0.8842105263157894

	0	1	2	3	4
0	27.000	0.000	0.000	0.000	0.000
1	2.000	47.000	1.000	0.000	2.000
2	3.000	1.000	28.000	0.000	0.000
3	0.000	2.000	0.000	31.000	0.000
4	8.000	2.000	1.000	0.000	35.000

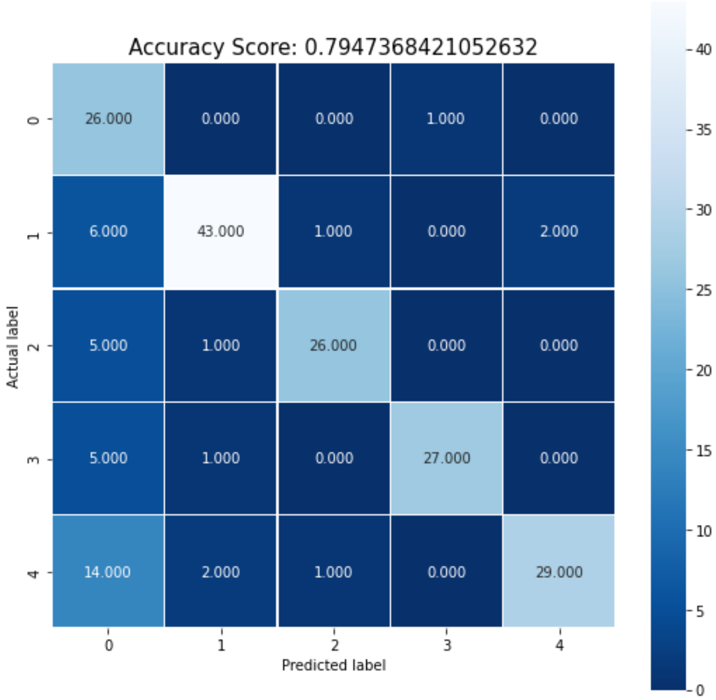
Bigram



Accuracy Score: 0.8526315789473684

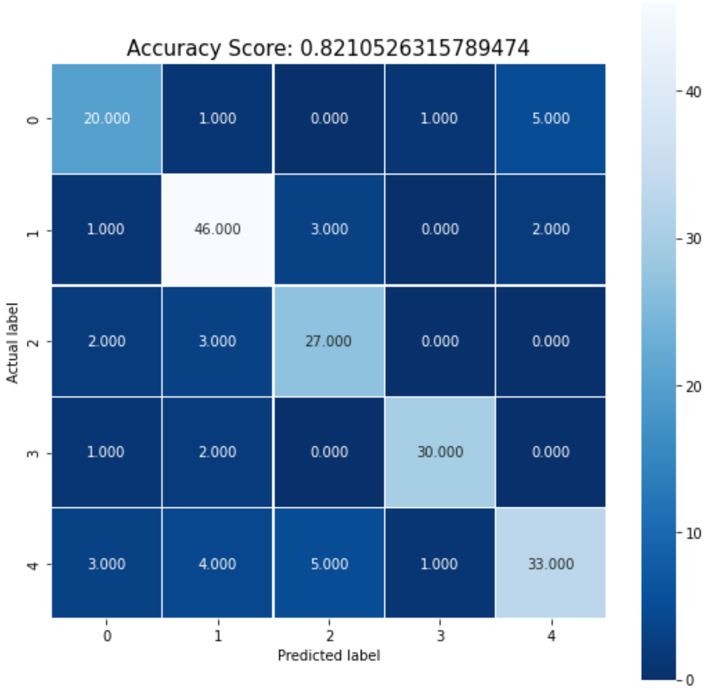
	0	1	2	3	4
0	27.000	0.000	0.000	0.000	0.000
1	2.000	47.000	1.000	0.000	2.000
2	5.000	1.000	26.000	0.000	0.000
3	2.000	1.000	0.000	30.000	0.000
4	12.000	1.000	1.000	0.000	32.000

Trigram

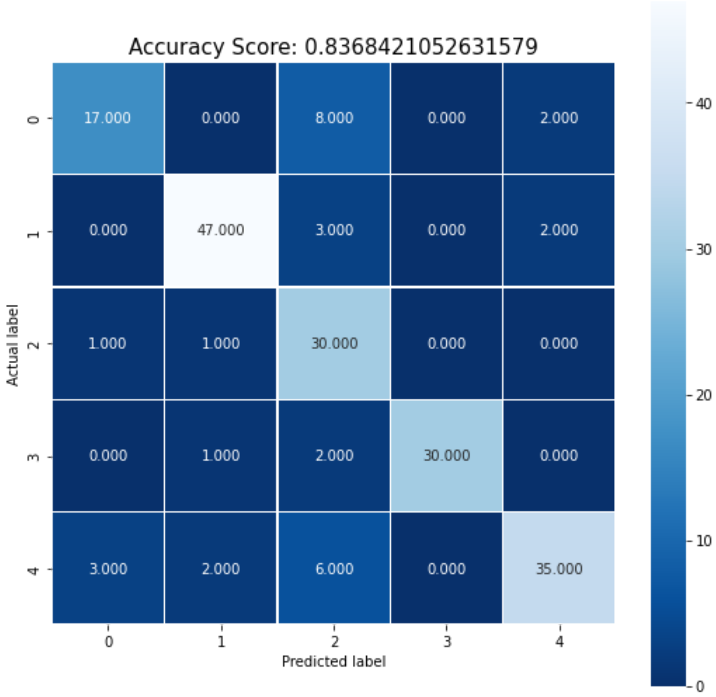


Decision Tree

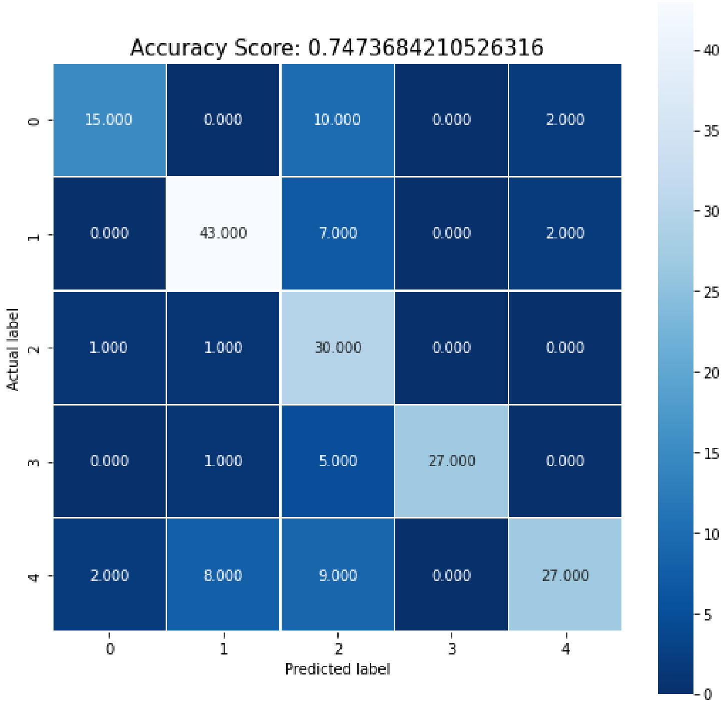
Unigram



Bigram



Trigram


Accuracy Score: 0.7473684210526316
Actual label
Predicted label