CLASSIFYING AND CLUSTERING THE 20 NEWSGROUPS TEXT DATASET

Sarita Himthani

Pace University, New York

1. ABSTRACT

Classification and Clustering techniques find their usefulness in a variety of text applications. In this project, an algorithmic approach has been leveraged for document classification and document clustering. Various models have been trained for document classification and they all have been evaluated using performance metrics followed by tuning of the model hyper-parameters to reach towards the most accurate classification. Additionally, a model has been trained for document clustering, which is followed by a dimensionality reduction technique to visualize the document clusters in 2D space.

Keywords: Natural Language Toolkit, Vectorization, Document Classification, Classification Model, Naive Bayes, Support Vector Machine, K-Nearest Neighbors, Performance Metric, Hyper-parameter, Grid Search, Cross Validation, Document Clustering, Clustering Model, K-Means, Dimensionality Reduction, Principal Component Analysis, Word Cloud

2. INTRODUCTION

The task of document classification involves classifying a document to one of the categories or classes, where as the task of document clustering involves grouping together the documents that have similar content.

2.1. Research Questions

Following are the research questions that the project attempts to answer:

a. Does the number of documents and the number of classes have an impact on the classification accuracy?
b. Can a document be classified into one of the available classes?
c. Are there clusters which can group the documents belonging to the similar topic?

2.2. Dataset

For document classification and clustering, a favorable dataset is the one that has a huge number of documents spread across a variety of topics. 'The 20 newsgroups text dataset' is one such dataset that has the following:

a. 18846 newsgroups posts
b. 20 topics

The entire dataset is already split into two subsets - a training set and a testing set. The dataset is publicly available under Real world datasets on scikit-learn.

These newsgroup posts belong to topics such as graphics, atheism, hardware, baseball, christian, guns, and so forth. After the thorough analysis of the dataset, it is observed that there are documents that belong to mac, ibm, graphics that come under the umbrella topic of computers. There are posts that belong to hockey, baseball, motorcycle, and autos that come under the umbrella topic of recreation. However, three of the topics (atheism, forsale, and christian) are the most different from the remaining 17 topics. Hence, in this project, these three topics are not considered for the purpose of document classification and document clustering, and eventually, documents from these three topics won't be the candidates that would contribute to answering the research questions.

3. METHODS

To answer the first research question that talks about the impact on the classification accuracy due to the number of documents and number of classes, a decision is made to utilize the dataset in a particular way which is as follows:

a. *More classes with less number of documents per class:* In this approach, the 17 topics have been treated individually. There are an average of 565 documents that belong to each of those 17 topics. The plot in Figure 1. illustrates the number of documents per topic, wherein the topic named rec.sport.hockey has the highest number of documents, while the topic named talk.religion.misc has the lowest number of documents.

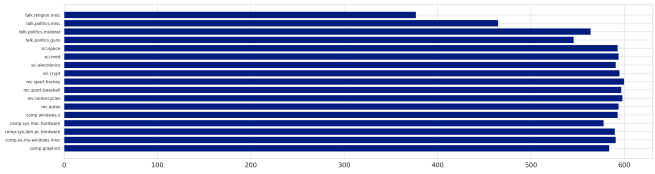

Fig. 1. Number of documents across 17 topics

b. *Less classes with more number of documents per class:* In this approach, documents that come under the same umbrella topic are combined to form a group. This resulted in 17 topics getting reduced to just 4 topics namely - Computer, Recreation, Science, and Talk. This synthesis is done by changing the target labels in both the training and the testing data. Through this synthesis, an average of 2410 documents belonging to each of those 4 new classes is achieved. The plot in Figure 2 illustrates the number of documents per new topic, wherein the topic named computer has the highest number of documents, while the topic named talk has the lowest number of documents.

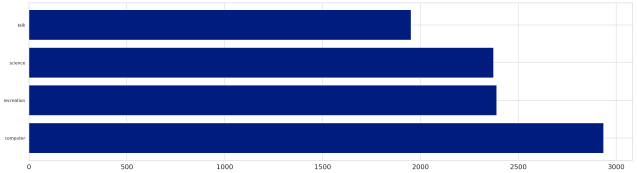

Fig. 2. Number of documents across 4 topics

To answer the second research question that talks about classifying documents, various classification models are trained and tested that lead to a document being assigned to a specific class out of all the available classes. Performance of each of the models is evaluated using the metrics such as Accuracy, Precision, Recall, and F1-score. Additionally, for each of the classifiers, hyper-parameter tuning is done using the grid search cross validation technique to get the best model parameters. The classification models are retrained and reevaluated using the hyper-tuned parameters. Eventually, along with the training and testing accuracy curves, confusion matrix is reported that summarizes the performance of the classifier. All these steps are performed during document classification for both 17 individual groups and 4 clubbed groups. Following are the models that are leveraged for classification:

a. K-Nearest Neighbor (KNN) Classifier
b. Support Vector Machine (SVM) Classifier
c. Multinomial Naive Bayes (MNB) Classifier

To answer the third research question that talks about finding clusters that can group documents belonging to the similar topic, a clustering algorithm named the K-Means algorithm is leveraged. The K corresponds to the number of clusters and the value for it is found using the elbow method. However, even after finding the appropriate number of clusters, all the efforts worth less if there is no way to visualize those clusters. To visualize those clusters, the data in the higher dimension has to be mapped to the data in the lower dimension by a dimensionality reduction technique. The technique of Principal Component Analysis (PCA) has been leveraged to reduce the data to two dimensions so that the clusters can be visualized in a 2D space.

In this project, the data consists of the newsgroup posts. Each of those posts are preprocessed before they can be used for either classification or clustering. The preprocessing of every post involves removing URL or links, numbers, extra whitespaces, and special characters. Another important removal that is taken care of is the removal of stop words because they are the most commonly occurring words, which do not give additional value to the document vector. Preprocessing also involves converting the words into their lowercase representation and once the preprocessing is complete, it is followed by vectorization for word importance. A Term Frequency Inverse Document Frequency (TF-IDF) vectorizer is leveraged to transform the words in the documents into a meaningful representation of numbers, which is used by the classification and clustering models.

Additionally, to visualize the importance or frequency of words in a cluster, a Word Cloud is constructed. For each cluster, a word cloud is created that depicts the words of varying font size belonging to a particular cluster. A word that is more important or frequent in a cluster has a greater font size than a word that is less important or rare.

4. RESULTS

4.1. Document Classification

4.1.1. Classification using 17 Individual Groups

Classification models (along with the hyper-tuned ones) are trained and tested using the 17 individual groups. The table in Figure 3 summarizes the training and testing accuracies of all the models. It can be observed that the testing set accuracy of all the models increases after tuning the hyper-parameters. With a testing set accuracy of 71.62%, the hyper-tuned MNB outperforms all the other models. It is

Algorithm	Training_Accuracy	Testing_Accuracy
Multinomial Naive Bayes Hypertuned	0.924249	0.716265
Multinomial Naive Bayes	0.931710	0.713774
Support Vector Machine Hypertuned	0.960518	0.698366
Support Vector Machine	0.969119	0.695875
K Nearest Neighbor Hypertuned	0.971399	0.118599
K Nearest Neighbor	0.414715	0.090584

Fig. 3. Accuracies of classifiers using 17 individual groups

able to classify the documents better than other classifiers. The accuracies of SVM and its hyper-tuned version closely follow up. The training set accuracy of the hyper-tuned KNN is the best but it doesn't perform well on the testing set. Lastly, performance of the vanilla KNN is not at all impressive as it achieves an accuracy of just 9.05% on the testing set.

The line plot in Figure 4 illustrates the training set and testing set accuracy curves for all the models. The training set accuracy of all the models is greater than the testing set accuracy. The gap between both the curves indicates huge

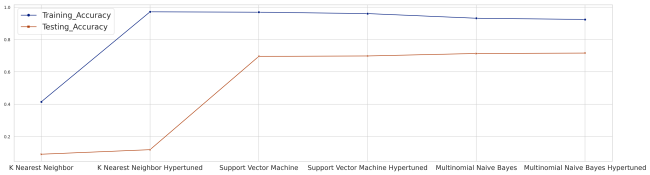

Fig. 4. Accuracy curves for 17 individual groups

overfitting, which means that the models have learned the noise of the training documents very well. The KNN classifier overfits most to the training documents, whereas the MNB classifier overfits the least. The SVM classifier overfits slightly more than the MNB classifier.

4.1.2 Classifying under 4 Clubbed Groups

Classification models (along with the hyper-tuned ones) are trained and tested using the 4 clubbed groups. The table in Figure 5 summarizes the training and testing accuracies of all the models. It can be observed that the testing set accuracy of all the models increases after tuning the hyper-parameters. With a testing set accuracy of 85.01%, the hyper-tuned MNB outperforms all the other models. It is

Algorithm	Training_Accuracy	Testing_Accuracy
Multinomial Naive Bayes Hypertuned	0.928601	0.850117
Multinomial Naive Bayes	0.932539	0.848249
Support Vector Machine Hypertuned	0.966943	0.837510
Support Vector Machine	0.976269	0.836732
K Nearest Neighbor Hypertuned	0.978342	0.295564
K Nearest Neighbor	0.594301	0.289027

Fig. 5. Accuracies of classifiers using 4 clubbed groups

able to classify the documents better than other classifiers. The accuracies of SVM and its hyper-tuned version closely follow up. The training set accuracy of the hyper-tuned KNN is the best but it doesn't perform well on the testing set. Lastly, the performance of the vanilla KNN lags far behind as it achieves an accuracy of just 28.9% on the testing set.

The line plot in Figure 6 illustrates the training set and testing set accuracy curves for all the models. The training set accuracy of all the models is greater than the testing set accuracy. The gap between both the curves indicates overfitting, which means that the models have learned the noise of the training documents very well. The KNN classifier overfits most to the training documents,

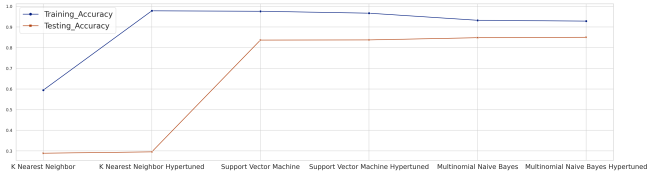

Fig. 6. Accuracy curves for 4 clubbed groups

whereas the MNB classifier overfits the least. The SVM classifier overfits slightly more than the MNB classifier.

4.1.3 Accuracy: 17 Individual Groups vs 4 Clubbed Groups

All the models trained and tested using the 4 clubbed groups perform better than their 17 individual group counterparts.

Algorithm	17_Individual_Groups_Testing_Accuracy	4_Clubbed_Groups_Testing_Accuracy
Multinomial Naive Bayes Hypertuned	0.716265	0.850117
Multinomial Naive Bayes	0.713774	0.848249
Support Vector Machine Hypertuned	0.698366	0.837510
Support Vector Machine	0.695875	0.836732
K Nearest Neighbor Hypertuned	0.118599	0.295564
K Nearest Neighbor	0.090584	0.289027

Fig. 7. Accuracy comparison: 17 individual groups vs 4 clubbed groups

The table in Figure 7 illustrates that for this dataset, all the models are able to classify the documents with greater accuracy when the documents are clubbed under 4 umbrella topics than when there are 17 individual topics. The hyper-tuned MNB classifier and the hyper-tuned SVM classifier achieved a significant accuracy gain (approximately 19%) when it was trained and tested using the documents in 4 clubbed groups. Even though the vanilla KNN classifier is the least impressive model, it achieves a massive accuracy gain (approximately 220%) when it is trained and tested using the documents in 4 clubbed groups.

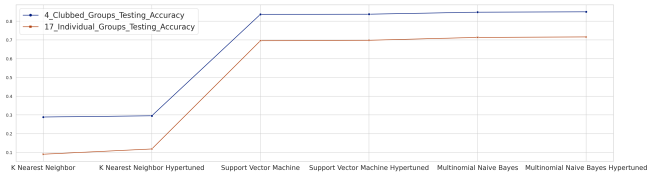

Fig. 8. Testing accuracy curves: 17 individual groups vs 4 clubbed groups

The line plot in Figure 8 illustrates the testing set accuracy curves for all the models including their hyper-tuned versions.The testing accuracy of the 4 clubbed groups follow the same trend as that of the 17 individual groups.

However, the accuracies of the former are significantly better. These accuracy gains signify the importance of more number of documents spread across reduced number of topics.

4.1.4 Metrics: 17 Individual Groups vs 4 Clubbed Groups

Performance metrics such as Precision, Recall, and F1 Score are calculated. The table in Figure 9 illustrate these metrics for both the 17 individual groups and the 4 clubbed groups.

	Model	Precision	Recall	F1	Test_Accuracy
4 Clubbed Groups	Multinomial Naive Bayes Hypertuned	85.186507	85.011673	84.888599	0.850117
	Multinomial Naive Bayes	85.012906	84.824903	84.707742	0.848249
	Support Vector Machine Hypertuned	83.773891	83.750973	83.672813	0.837510
	Support Vector Machine	83.636989	83.673152	83.555545	0.836732
	K Nearest Neighbor Hypertuned	71.928645	29.556420	19.509310	0.295564
	K Nearest Neighbor	28.249116	28.902724	27.535885	0.289027
17 Indivudial Groups	Multinomial Naive Bayes Hypertuned	73.158085	71.626459	72.004476	0.716265
	Multinomial Naive Bayes	72.087955	71.377432	71.299498	0.713774
	Support Vector Machine Hypertuned	70.363198	69.836576	69.839174	0.698366
	Support Vector Machine	70.204894	69.587549	69.651118	0.695875
	K Nearest Neighbor Hypertuned	75.020130	11.859922	11.300345	0.118599
	K Nearest Neighbor	13.632067	9.058366	8.285497	0.090584

Fig. 9. Performance metrics: 17 individual groups vs 4 clubbed groups

Apart from the hyper-tuned KNN classier, the precision increases by approximately 12% for all the remaining classifiers. Hence, the number of total positive predictions that are correct increases by 12% for the remaining classifiers when the documents are clubbed into 4 groups.

The recall score improves for all the classifiers in 4 clubbed groups. It has shown significant improvement from 8% to 27% for the KNN classifier. Also, for other classifiers, the recall score improves by approximately 13%. It means that the number of positive cases correctly predicted over actual positive cases increases by 13%.

The F1 score has shown similar trends as that of the recall score. The KNN classifier has shown good improvement and the score of the remaining classifiers improves approximately by 13%. The F1 score entirely depends on Precision and Recall. If any of the values are low, the F1 score gets affected accordingly. As the Precision and Recall scores increases for the 4 clubbed groups, the F1 score also increases.

The accuracy score for all the classifiers has been improved significantly. The new model with 4 clubbed groups is more likely to predict correctly than the 17 Individual groups model.

The performance metrics of all the hyper-tuned classifiers are better than the performance metrics of their vanilla counterparts. And the performance metrics of the 4 clubbed groups are better than that of the the 17 individual groups.

4.2. Document Clustering

4.2.1. K-Means Algorithm

Elbow method is leveraged to choose the value of K that corresponds to the number of clusters. The plot in Figure 10 illustrates the line that helps in choosing the elbow point. It can be observed that the sum of squared errors (inertia) does

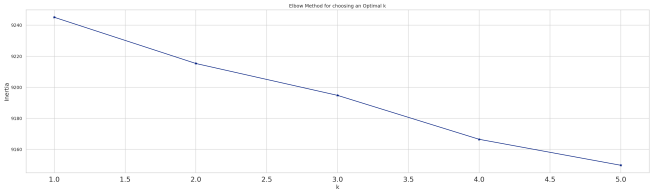

Fig. 10. Elbow Method for choosing an optimal K

not improve much after point K = 4. Hence, 4 is chosen to be the number of clusters and each document is assigned to one of the clusters using the K-Means algorithm. This value of K matches the number of topics that are there after clubbing i.e. 4.

4.2.2. Principal Component Analysis

The dimensions of the dataset is reduced by performing the Principal Component Analysis and the clusters are visualized in a 2D space as shown in Figure 11. The 4 clusters correspond to the 4 clubbed groups i.e. Talk, Computer, Science, and Recreation.

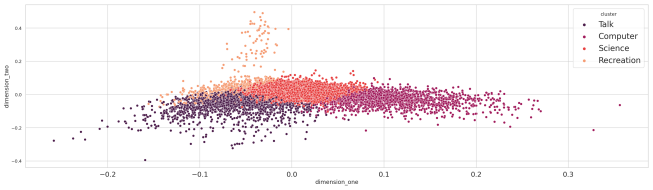

Fig. 11. Clusters of 4 clubbed groups

There is some overlap between the clusters of different topics. This is because there are common words that are present in the documents that belong to different topics. For instance, considering Science and Recreation clusters, words such as *good*, *think*, and *one* are present in the documents that belong to both the topics. Another instance is that the word *think* appears in documents that belong to three of the topics i.e. Talk, Science, and Recreation.

4.2.3. Word Cloud

The Word Cloud is created and visualized for each of the clusters. The words which are appearing repeatedly are big in size and the occurrence or importance of the word is directly proportional to the size of the word. Figure 12, 13, 14, and 15 illustrate the Word Clouds corresponding to the Talk, Computer, Science, and Recreation, respectively.

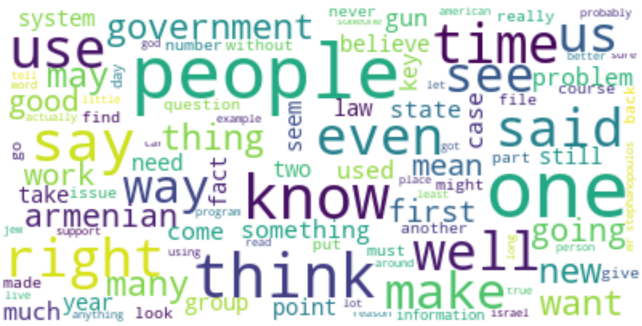

Fig. 12. Word Cloud for topic named 'Talk'

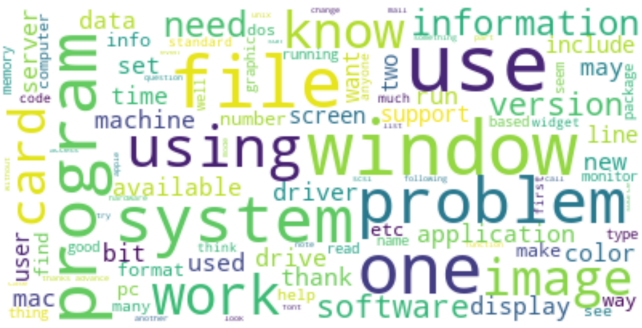

Fig. 13. Word Cloud for topic named 'Computer'

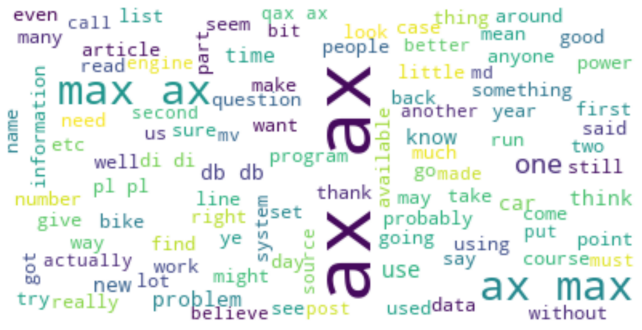

Fig. 14. Word Cloud for topic named 'Science'

Fig. 15. Word Cloud for topic named 'Recreation'

5. DISCUSSION

Through the series of experiments, the three research question are answered.

Does the number of documents and the number of classes have an impact on the classification accuracy? Yes, the number of training documents and the number of classes does impacts the classification accuracy. It is seen that when the documents belonging to the 17 individual groups are clubbed to belong to just 4 groups, the classification model's accuracy increases. The classification accuracy of MNB classifier and SVM classifier increases by approximately 19% and the KNN classifier increases by 220%.

Can a document be classified into one of the available classes? Yes, by using the classification models such as KNN, SVM, and MNB, it is possible to classify the document to belong to one of the available classes. However, all the models overfit the training documents with KNN classifier overfitting the most and MNB classifier overfitting the least. SVM classifier's performance is comparable to the MNB classifier but MNB classifier is able to generalize better on the testing set documents. All the hyper-tuned models classified the documents with higher accuracy than their vanilla counterparts.

Are there clusters which can group the documents belonging to the similar topic? Yes, K-Means algorithm helps finding the clusters and the number of clusters matches with the 4 clubbed groups i.e. Talk, Computer, Science, and Recreation. Using PCA, it is possible to visualize those clusters in the 2D space. Also, words from the documents in each cluster are visualized through a Word Cloud and on closely observing the words, it is seen that words are related to the topic or cluster to which they belong.

6. REFERENCES

Following are the references:

a. https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html

b. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
c. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
d. https://towardsdatascience.com/k-means-explained-10349949bd10
e. https://www.datacamp.com/community/tutorials/wordcloud-python
f. https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558
g. https://medium.com/criteo-engineering/hyper-parameter-optimization-algorithms-2fe447525903
h. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
i. https://www.researchgate.net/publication/317173563_Bayesian_Multinomial_Naive_Bayes_Classifier_to_Text_Classification
j. https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/
k. https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html
l. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html