SMAI

PROJECT

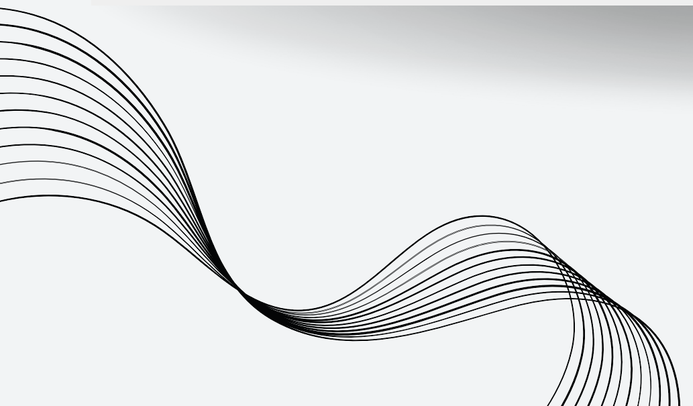
TEAM GALAXY

CONTENT

INTRODUCTION

Some algorithms require a large amount of labeled data to train the classifier. But, in some cases, it is difficult to get a large amount of labeled data. In such cases, the unlabeled data can be used to improve the performance of the classifier.

One such algorithm is the cotrain algorithm which is mentioned in the paper (https://www.cs.cmu.edu/~avrim/Papers/cotrain.pdf).



INTRODUCTION

The Cotrain algorithm is a semi-supervised learning algorithm that uses unlabeled data to improve the performance of the classifier. The algorithm uses two classifiers and two views of the data to train the classifiers.

Co-training assumes that

- Features can be split into two sets.
- Each sub-feature set is sufficient to train a good classifier.
- The two sets are conditionally independent given the class.

COTRAINING

Given:

- a set L of labeled training examples
- a set U of unlabeled examples

Create a pool U' of examples by choosing u examples at random from U

Loop for k iterations:

Use L to train a classifier h_1 that considers only the x_1 portion of x

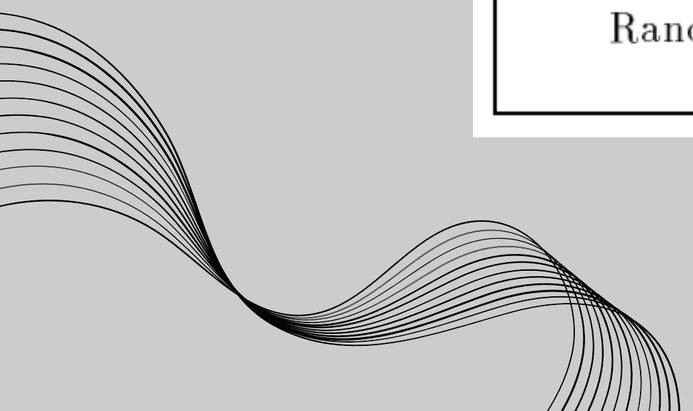
Use L to train a classifier h_2 that considers only the x_2 portion of x

Allow h_1 to label p positive and n negative examples from U'

Allow h_2 to label p positive and n negative examples from U'

Add these self-labeled examples to L

Randomly choose $2p + 2n$ examples from U to replenish U'



DELIVERABLES

01

DELIVERABLE 1

Implement the Cotraining paradigm described in the paper

02

DELIVERABLE 2

Generate, create, and find other relevant datasets which could be used to evaluate this method

03

DELIVERABLE 3

Do a comprehensive evaluation of the method

04

DELIVERABLE 4

Compare against current methods for similar training paradigm which uses both labelled and unlabelled datasets.

DATASETS USED

1

The WebKB Binary dataset comprises 1051 web pages from computer science departments categorized as Course (230) and Non-Course (821). The dataset is organized into Fulltext (web page content) and Inlinks (anchor text on hyperlinks) directories.

WEBKB (BINARY)

2

With 8,282 pages manually categorized into seven classes, including Student, Faculty, Staff, Department, Course, Project, and Other, the dataset features content from four specific universities and additional miscellaneous pages.

WEBKB (MULTICLASS)

DATASETS USED

3

News Category Dataset, encompassing 210k news headlines from HuffPost (2012-2022). With 42 categories, we focus on the top 15, such as POLITICS and WELLNESS. To address data imbalance, we randomly sampled 1000 articles per category for algorithmic analysis, utilizing two views: Headlines and Short Descriptions.

NEWS CATEGORY

4

Rt-polaritydata dataset, consisting of 10,662 snippets, divided equally into positive and negative sentiments. Each line represents a down-cased snippet, serving as the basis for creating an antonymous view. This dual-view dataset enhances our analysis by juxtaposing original and antonymous perspectives.

SENTENCE POLARITY

COMPARISION

Error rate in percent for classifying web pages as course home pages.

Results Shown in The Paper

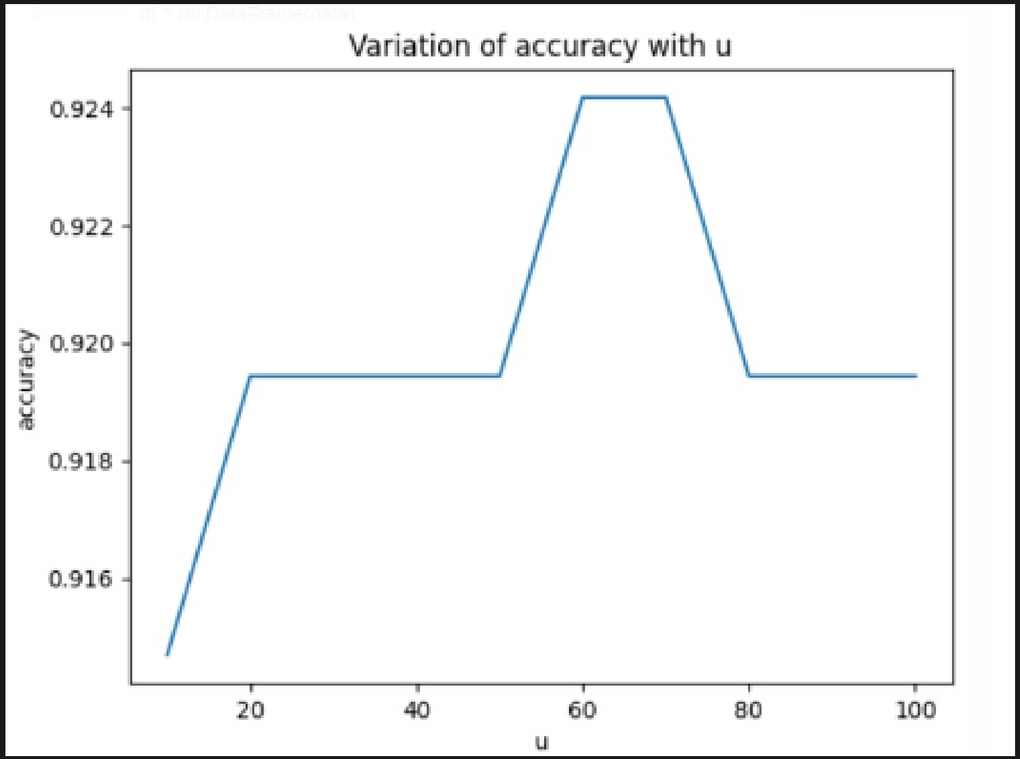
	Page-based classifier	Hyperlink-based classifier	Combined classifier
Supervised training	12.9	12.4	11.1
Co-training	6.2	11.6	5.0

Results we obtained

	Algorithm	page-based	link-based	combined
0	co-training	7.58294	40.2844	8.53081
1	supervised learning	14.6919	14.6919	14.6919

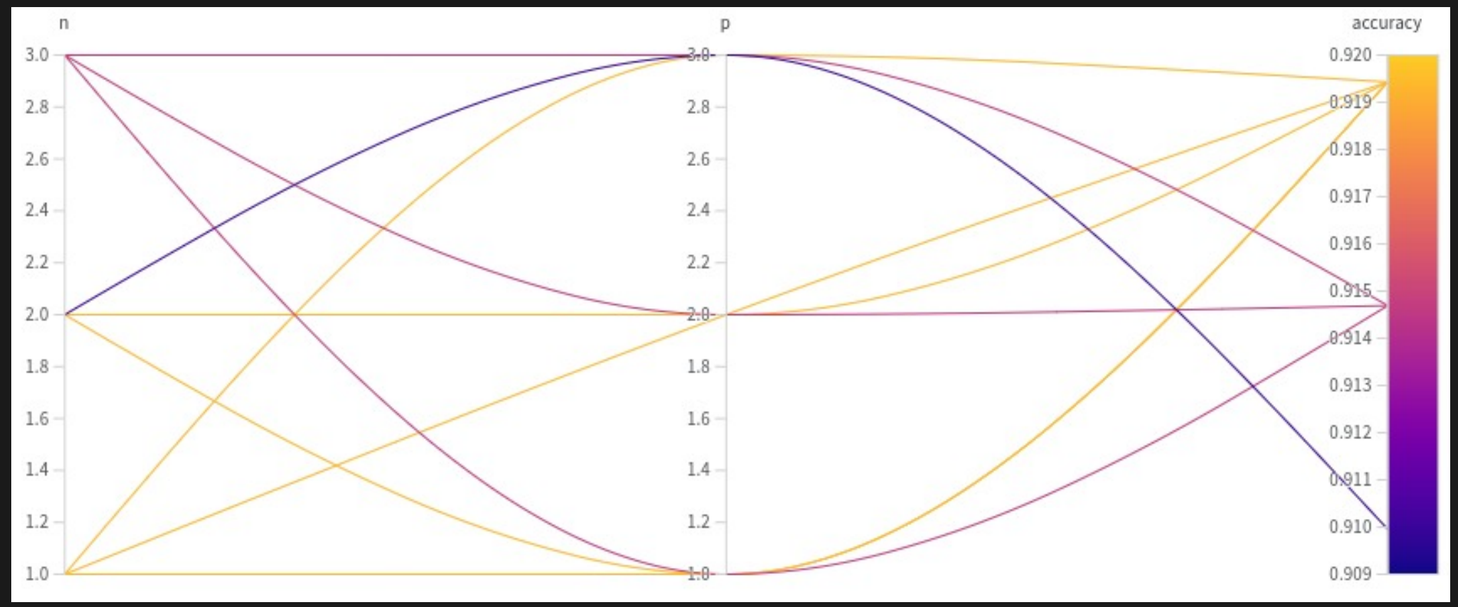
The difference in the error rates might be due to different methods of preprocessing data (tokenization, lemmatization, etc).

Variation of Accuracy with U



- The graph shows the variation of accuracy with ‘u’, where u is the number of unlabelled examples.
- From the above graph, we can observe that as the value of u increases at first accuracy increases reaches a maximum and then decreases.

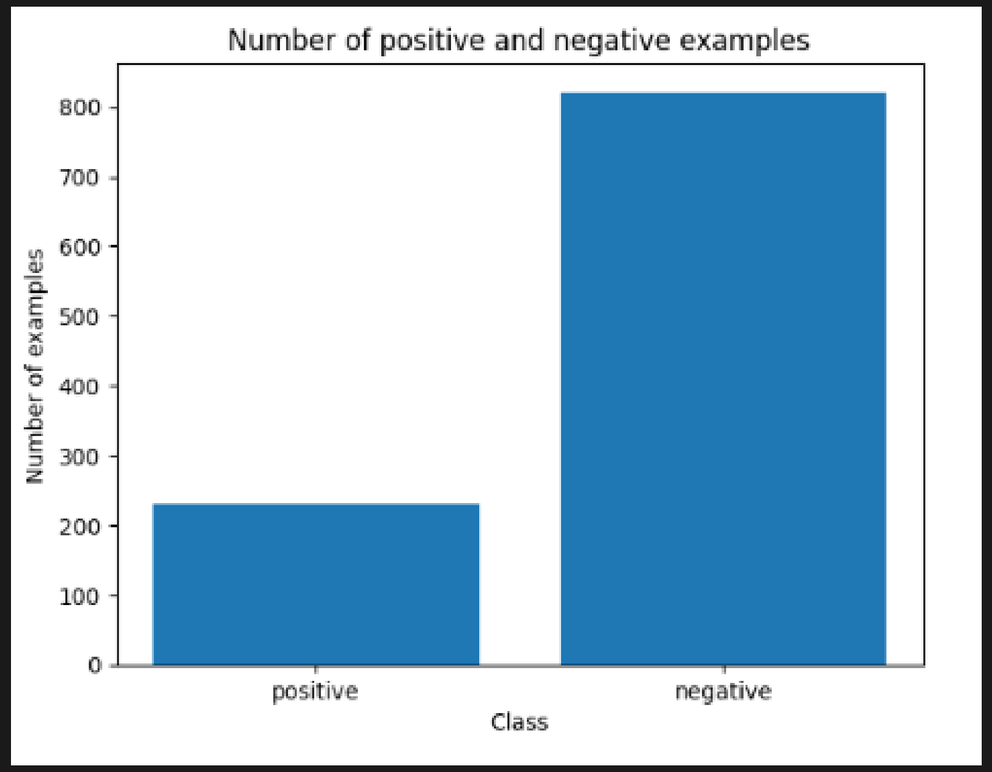
WandB sweep varying n and p



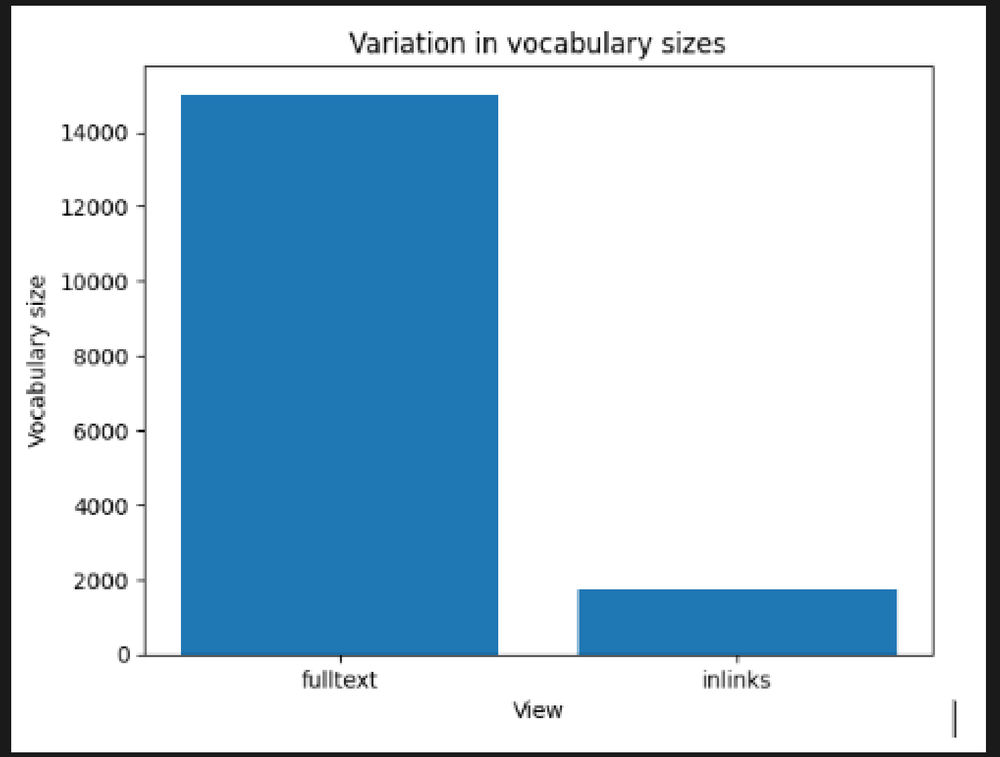
- The accuracy is maximum for the following (n,p) pairs: (1,3); (1,2); (2,2); (1,1)

OBSERVATIONS AND ANALYSIS

WEBKB BINARY

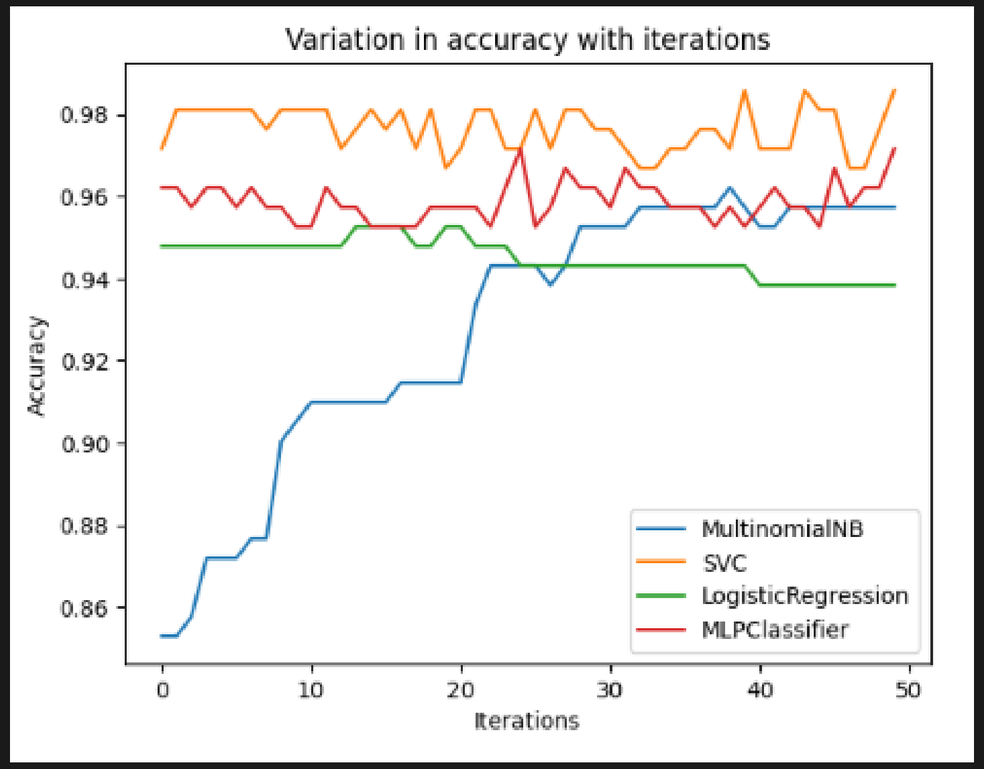


- The above figure shows the distribution of data samples between the two classes.
- From the graph, we can observe that the number of negative examples is 4 times that of the positive examples.

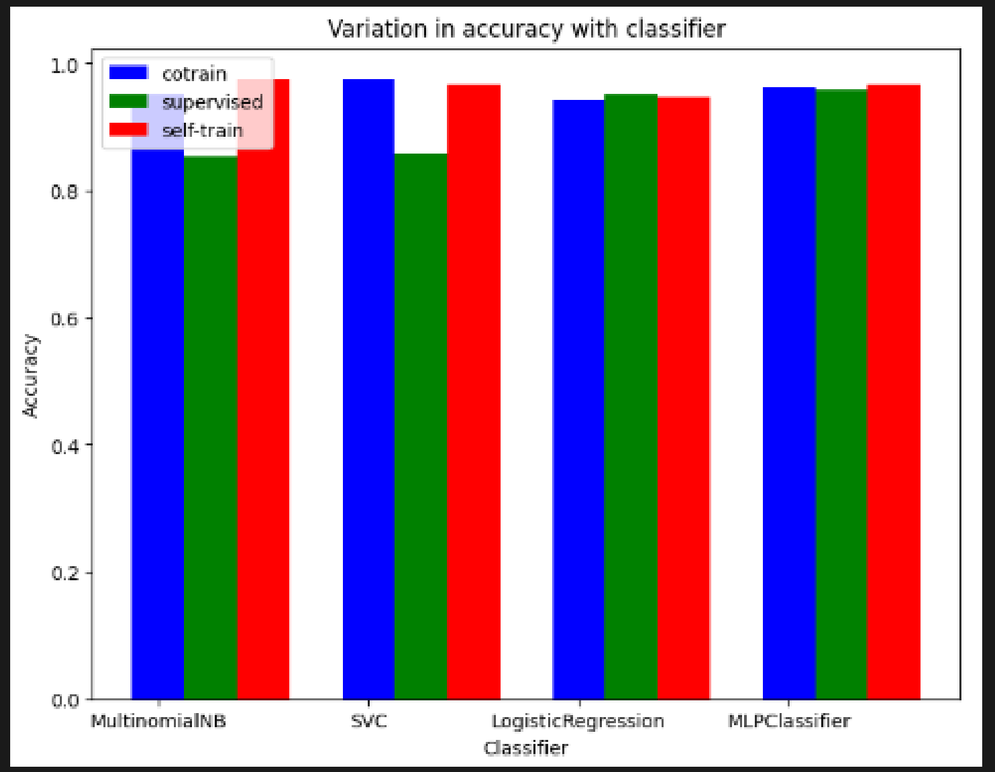


- The graph below shows the variation in vocabulary sizes of the two different views (FullText and Inlinks).
- The vocabulary size of the full-text view is much bigger compared to that of the in-links as the number of words in the webpage is more compared to that of the words used in hyperlinks pointing to that of the webpage.

WEBKB BINARY



From the above graph, as the number of iterations increases the performance of Multinomial Naive Bayes increases whereas, for Multinomial Logistic Regression at first, it increases and then starts decreasing. For the other two classifiers, there is no significant change in performance by increasing or decreasing the number of iterations.



- Using the Naive Bayes classifier and SVM, co-training and self-training outperform supervised training. In the case of MLP and logistic regression, the three methods are almost the same.
- By increasing the number of iterations co-training may outperform self-training in the case of Naive Bayes.

COMPARING THE THREE METHODS

COTRAINING

	Classifier	Accuracy	Precision	Recall	F1 Score
0	MultinomialNB	0.952607	0.973684	0.83871	0.890333
1	SVC	0.976303	0.938352	0.97276	0.954536
2	LogisticRegression	0.943128	0.894562	0.873208	0.883425
3	MLPClassifier	0.962085	0.915815	0.937724	0.926327

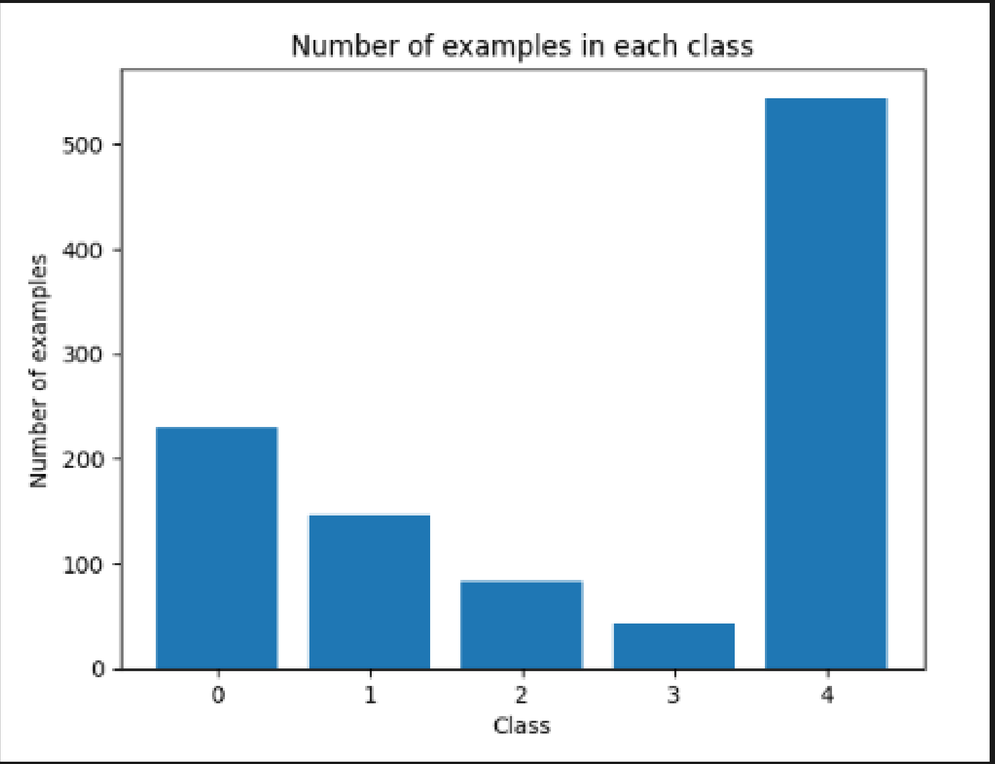
SELF-TRAINING

	Classifier	Accuracy	Precision	Recall	F1 Score
0	MultinomialNB	0.976303	0.947538	0.959409	0.953354
1	SVC	0.966825	0.920821	0.953853	0.93635
2	LogisticRegression	0.947867	0.909446	0.875986	0.891628
3	MLPClassifier	0.966825	0.92912	0.940502	0.934695

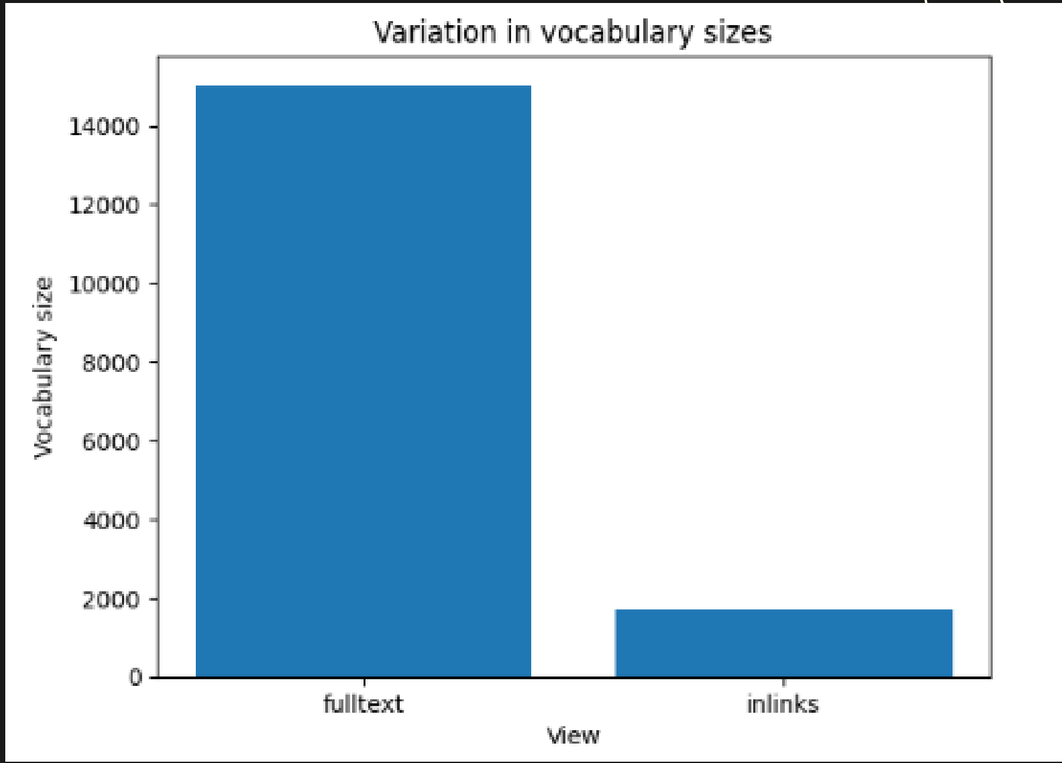
SUPERVISED LEARNING

	Classifier	Accuracy	Precision	Recall	F1 Score
0	MultinomialNB	0.853081	0.42654	0.5	0.460358
1	SVC	0.85782	0.928571	0.516129	0.492788
2	LogisticRegression	0.952607	0.914551	0.892115	0.902855
3	MLPClassifier	0.957346	0.957776	0.86819	0.90592

WEBKB MULTICLASS

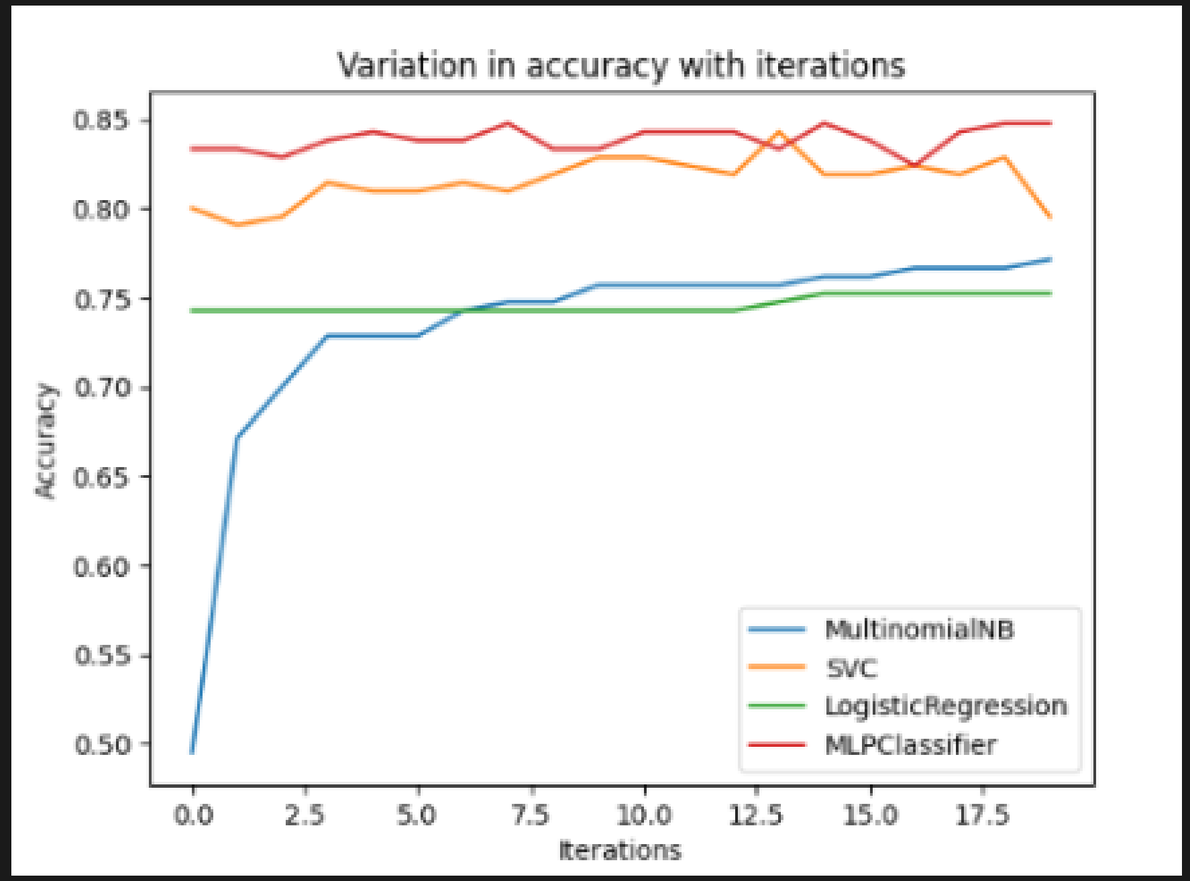


- The above figure shows the distribution of data samples between different classes.



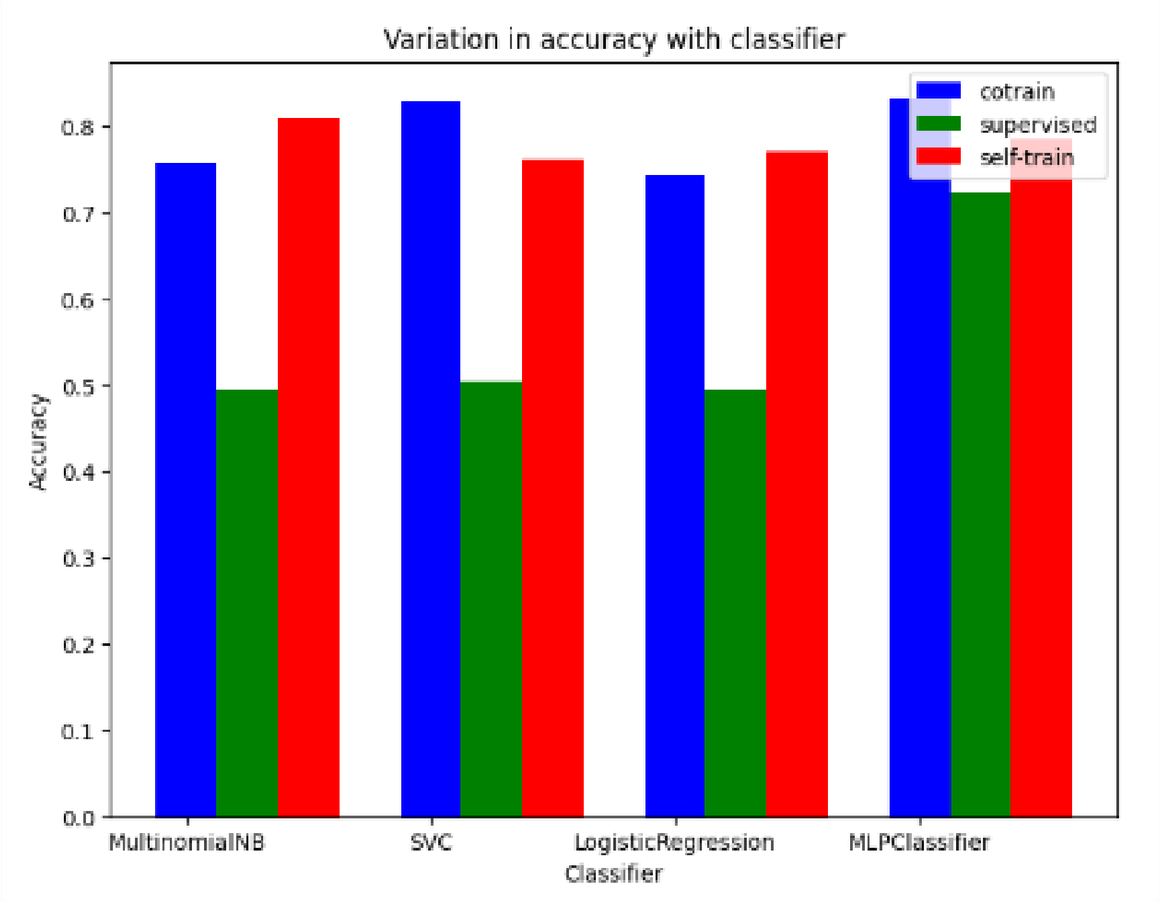
The vocabulary size of the full-text view is much bigger compared to that of the in-links as the number of words in the webpage is more compared to that of the words used in hyperlinks pointing to that of the webpage.

WEBKB MULTICLASS



With the increasing number of iterations, the performance of the Naive Bayes classifier increases. But, the performance of the other three classifiers is almost constant.

WEBKB MULTICLASS



- From the graph, we can observe that the semi-supervised training methods co-training and self-training outperform the supervised training. In Multinomial Naive Bayes and Logistic Regression, self-training outperforms co-training. But in the case of SVM and MLP co-training outperforms the self-training.
- In the previous graph, we observed that with increasing the number of iterations performance of Multinomial Naive Bayes using the co-training method can be improved.
- So in the case of the Naive Bayes classifier, by increasing the number of iterations co-training method may outperform self-training.

COMPARING THE THREE METHODS

COTRAINING

	Classifier	Accuracy	Precision	Recall	F1 Score
0	MultinomialNB	0.757143	0.534194	0.418259	0.402486
1	SVC	0.828571	0.731677	0.52376	0.535759
2	LogisticRegression	0.742857	0.731646	0.426377	0.443526
3	MLPClassifier	0.833333	0.680593	0.615158	0.634605

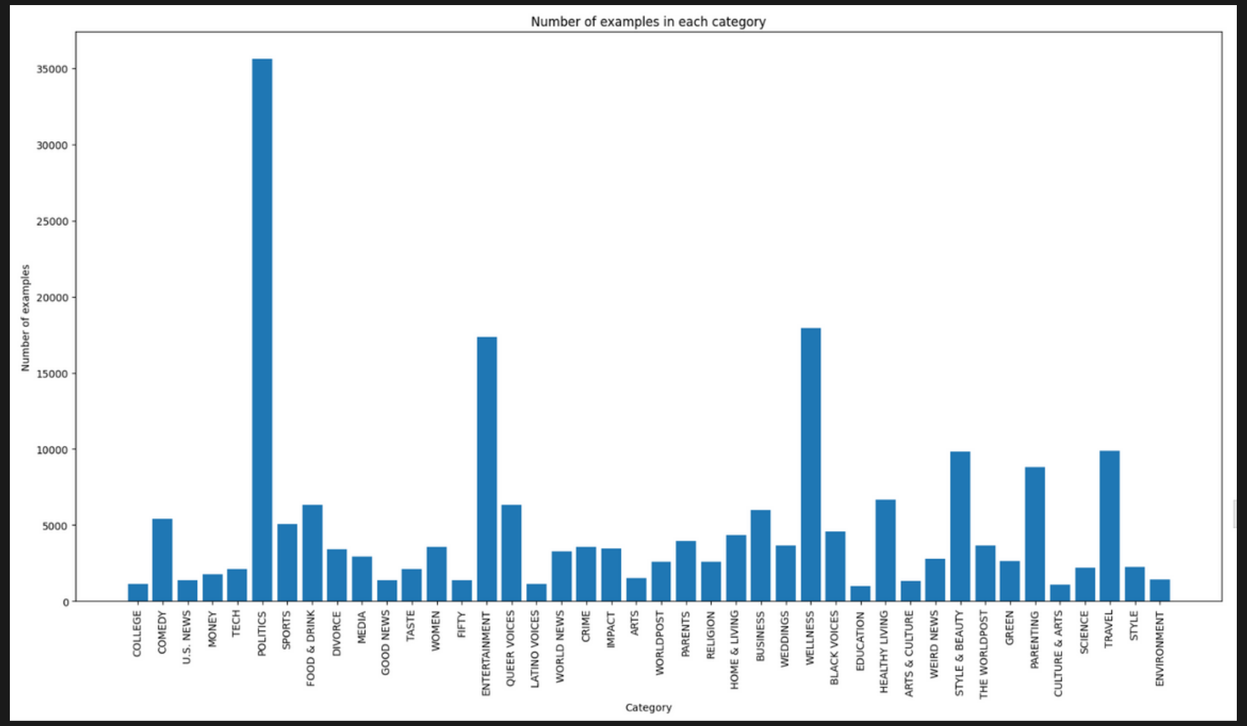
SELF-TRAINING

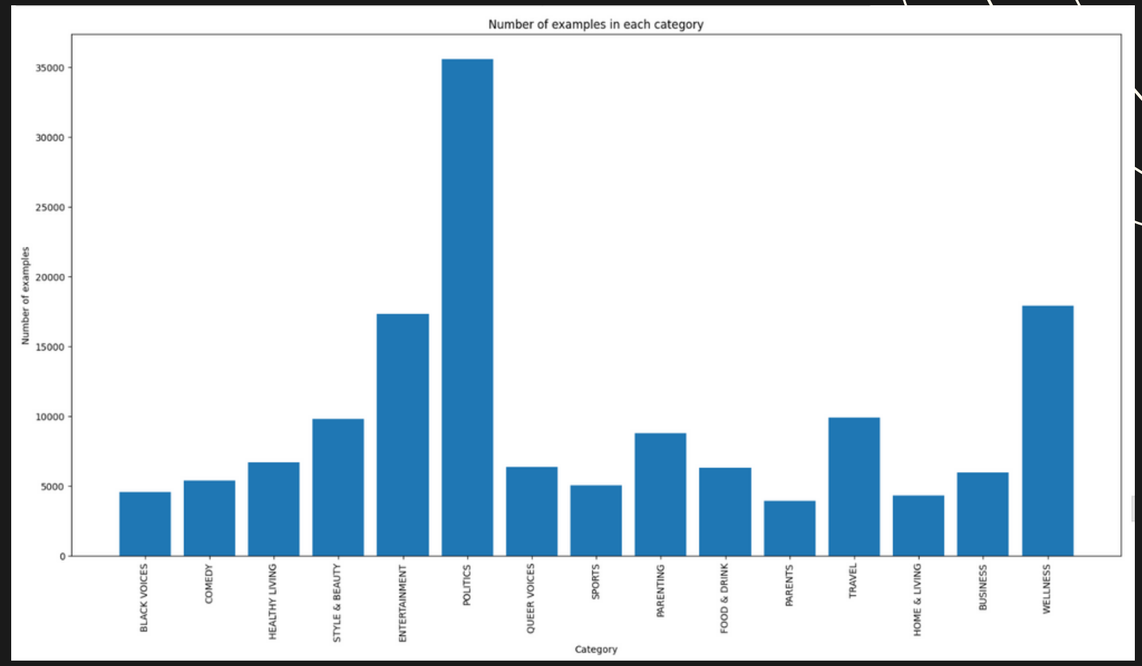
	Classifier	Accuracy	Precision	Recall	F1 Score
0	MultinomialNB	0.809524	0.693143	0.513751	0.534795
1	SVC	0.761905	0.731866	0.438793	0.447317
2	LogisticRegression	0.771429	0.736842	0.446341	0.453888
3	MLPClassifier	0.785714	0.70648	0.478123	0.49235

SUPERVISED LEARNING

	Classifier	Accuracy	Precision	Recall	F1 Score
0	MultinomialNB	0.495238	0.099048	0.2	0.132484
1	SVC	0.504762	0.3	0.207547	0.147879
2	LogisticRegression	0.495238	0.099048	0.2	0.132484
3	MLPClassifier	0.72381	0.521049	0.393695	0.382368

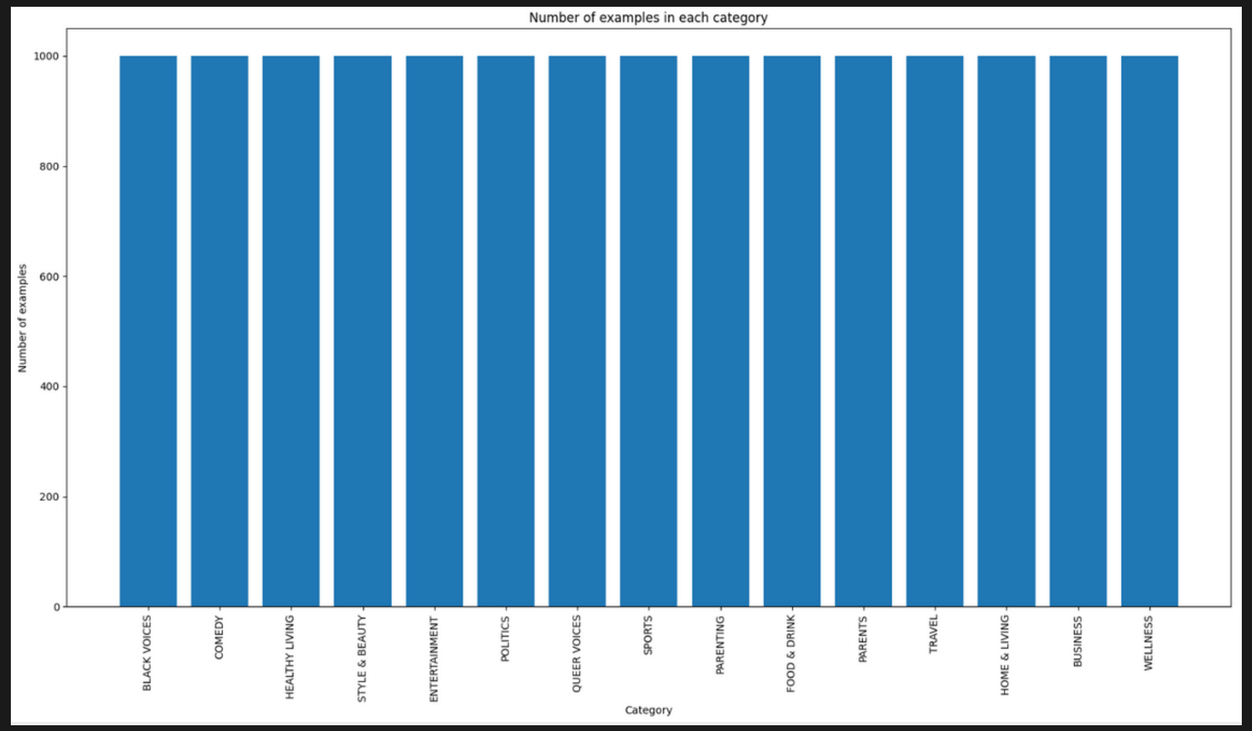
NEWS CATEGORY

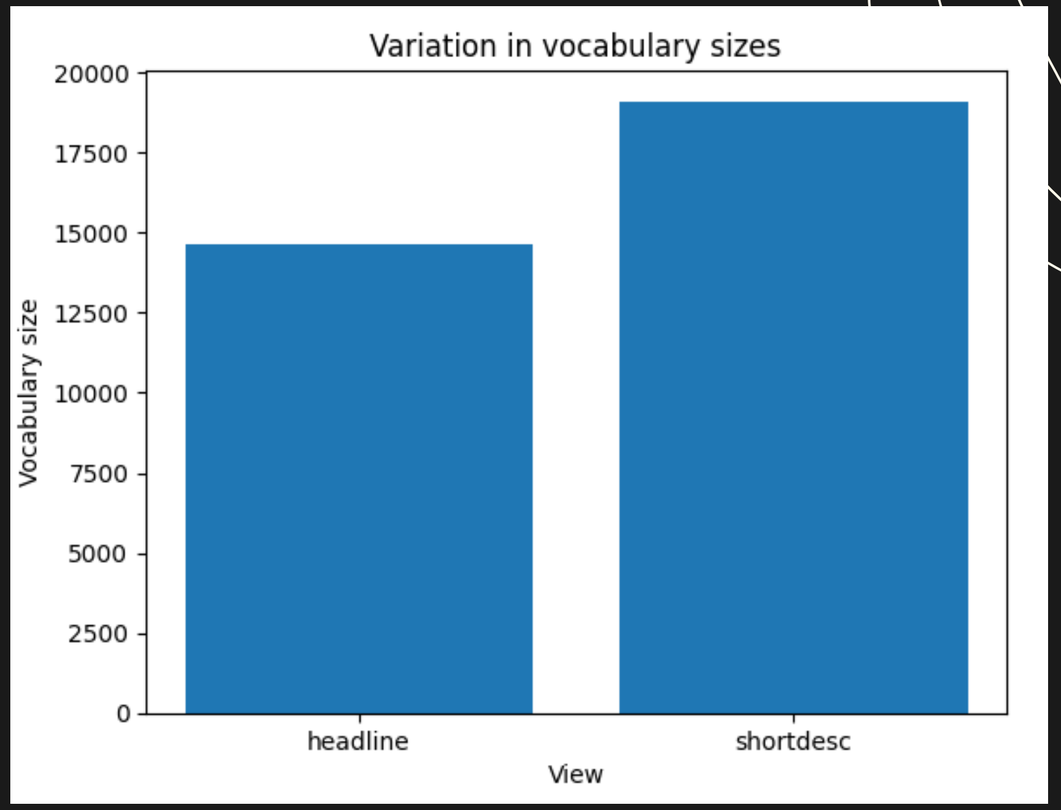




The above data contains 42 classes. We have taken top 15 classes from the data.

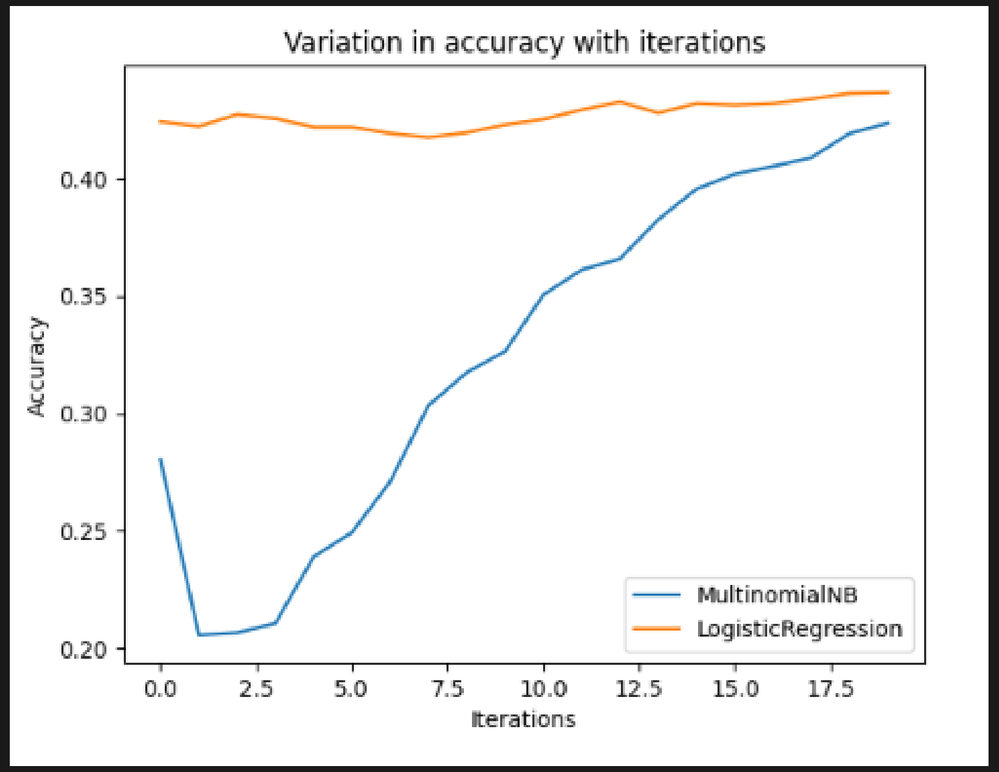
NEWS CATEGORY



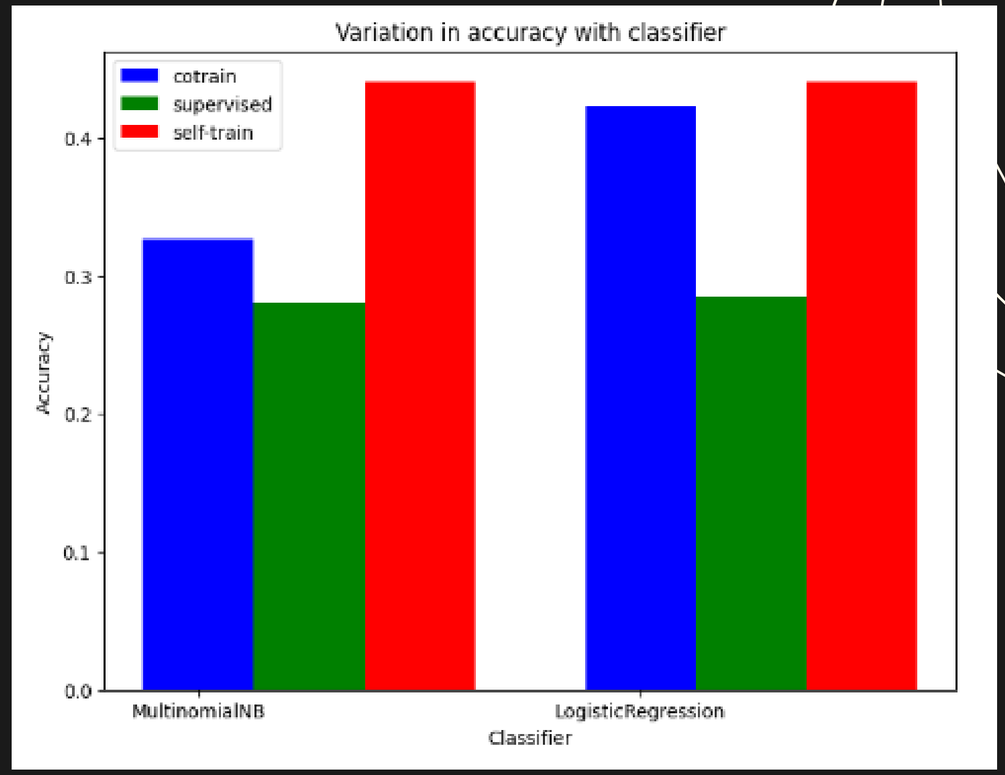


We have balanced the dataset by taking 1000 samples from each class.

NEWS CATEGORY



- The graph shows the variation in accuracy with the number of iterations for co-train.
- By increasing the number of iterations the performance of Naive Bayes is increasing significantly while the performance of Logistic regression is slightly increasing.



- The graph shows the variation in accuracy with the Classifier.
- From the above graph, we can observe that the semi-supervised training methods co-training and self-training outperform the supervised training. In the case of Naive Bayes self-training outperforms co-training significantly and in the case of Logistic regression both are similar.
- Based on the previous graph, the performance of co-training can be increased for both classifiers by increasing the number of iterations.

COMPARING THE THREE METHODS

COTRAINING

	Classifier	Accuracy	Precision	Recall	F1 Score
0	MultinomialNB	0.326333	0.503227	0.326108	0.298407
1	LogisticRegression	0.422667	0.449802	0.42193	0.408237

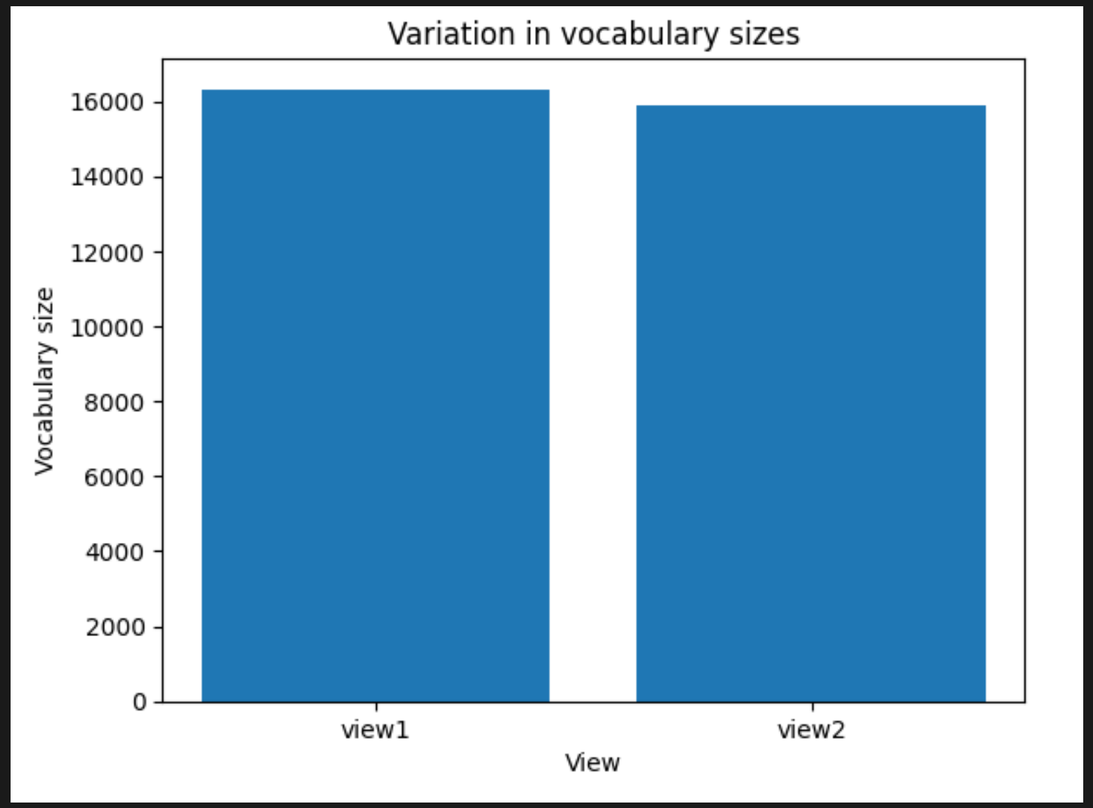
SELF-TRAINING

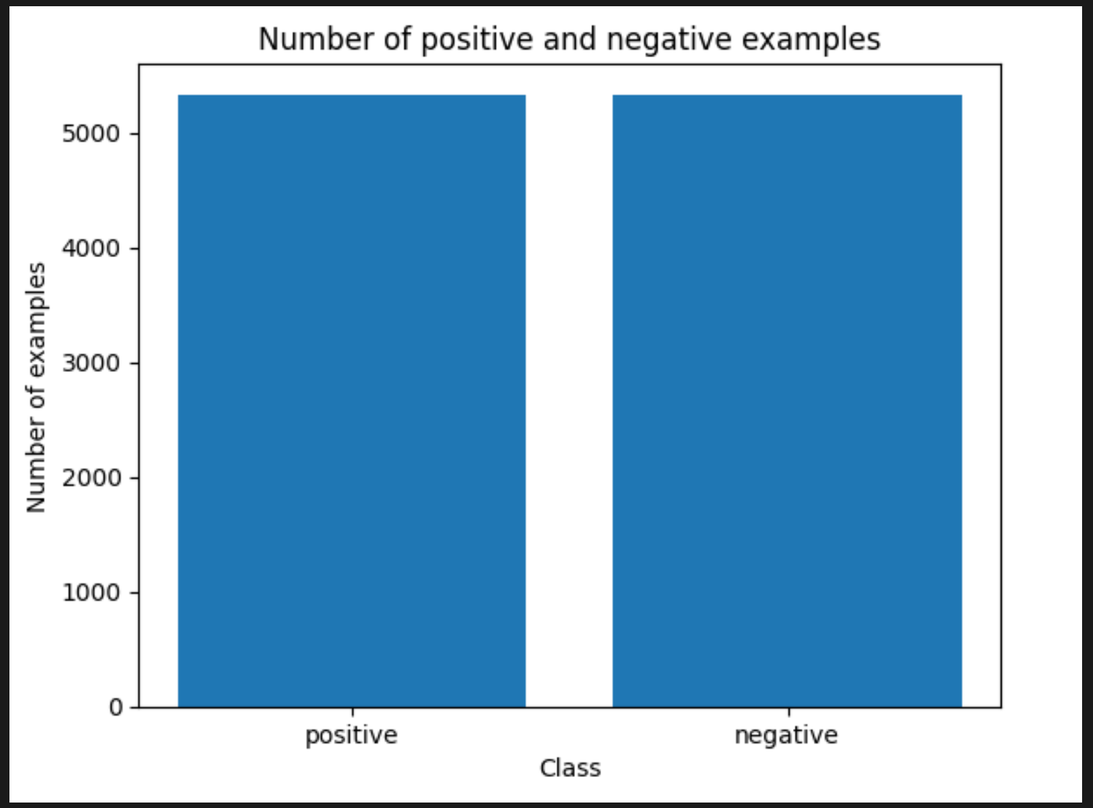
	Classifier	Accuracy	Precision	Recall	F1 Score
0	MultinomialNB	0.474667	0.525153	0.474262	0.462332
1	LogisticRegression	0.437667	0.475013	0.436982	0.424423

SUPERVISED LEARNING

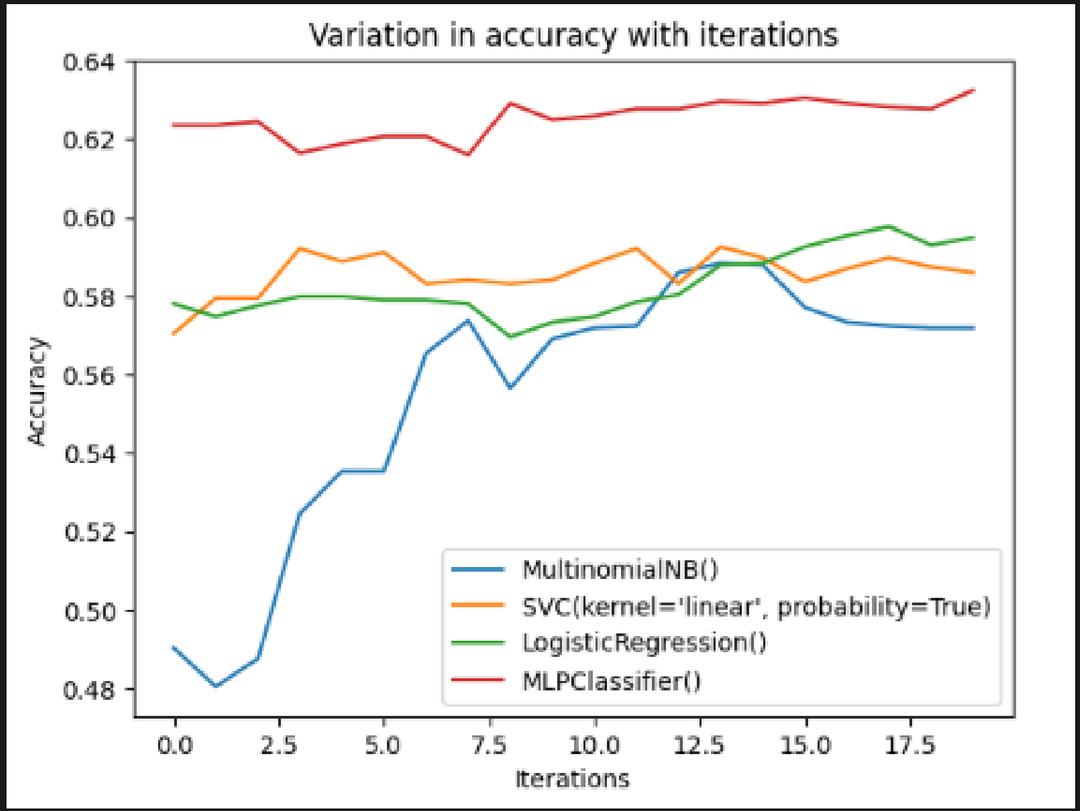
	Classifier	Accuracy	Precision	Recall	F1 Score
0	MultinomialNB	0.280333	0.280497	0.27808	0.274581
1	LogisticRegression	0.284667	0.283885	0.282199	0.277191

SENTENCE POLARITY



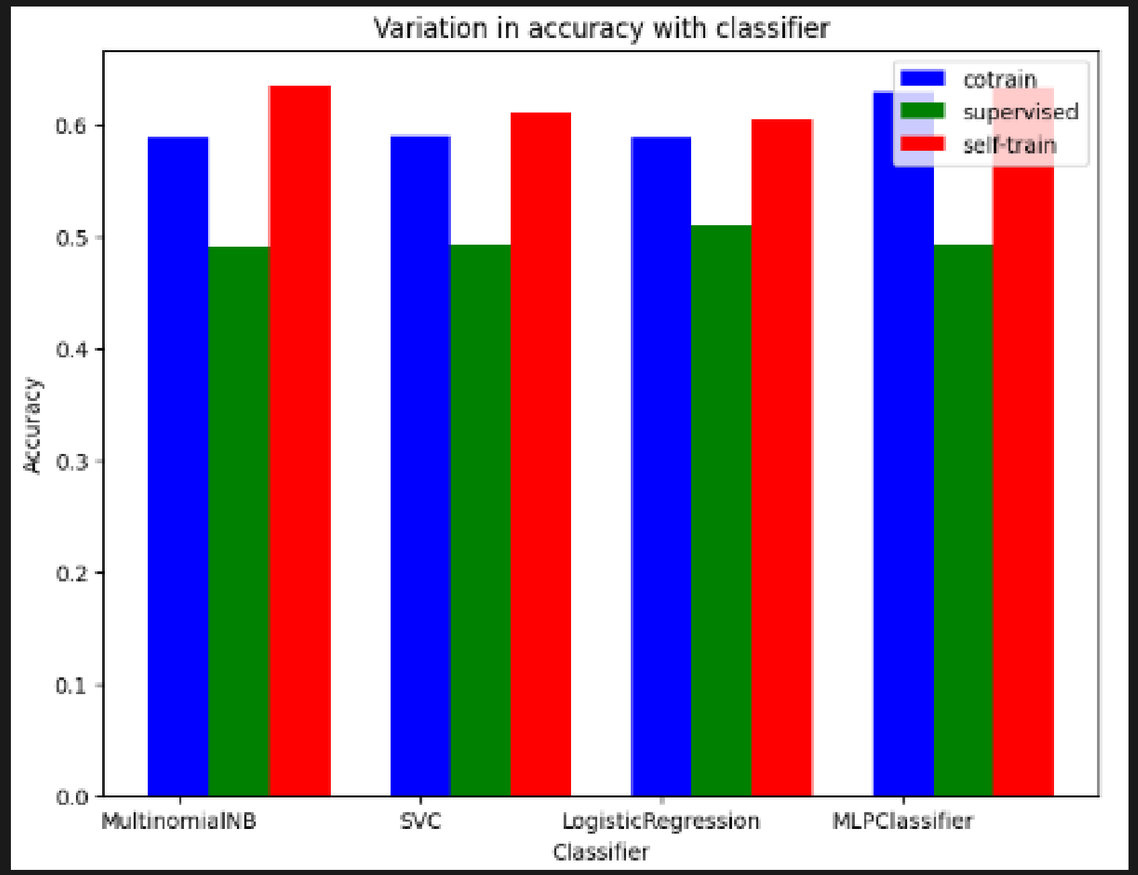


SENTENCE POLARITY



From the above graph, we can observe that with increasing the number of iterations, the performance of MLP and Logistic regression slightly increases. And the performance of Naive Bayes increases with iterations and then slightly decreases. The performance of SVM doesn’t vary much.

SENTENCE POLARITY



- From the above graph, we can observe that the semi-supervised training methods co-training and self-training outperform the supervised training. In Multinomial Naive Bayes, Logistic Regression, and SVM, self-training outperforms co-training. In the case of MLP co-training and self-training are almost similar.
- In the previous graph, we observed that by increasing the number of iterations performance of Multinomial Naive Bayes, MLP and logistic regression can be increased using the co-training method.
- So co-training method may outperform self-training by increasing the number of iterations in the case of MLP, Logistic Regression, and Naive Bayes.

COMPARING THE THREE METHODS

COTRAINING

	Classifier	Accuracy	Precision	Recall	F1 Score
0	MultinomialNB	0.587904	0.600184	0.593515	0.582884
1	SVC	0.58978	0.609931	0.597118	0.580213
2	LogisticRegression	0.588373	0.625061	0.598237	0.56911
3	MLPClassifier	0.629161	0.632046	0.631238	0.628957

SELF-TRAINING

	Classifier	Accuracy	Precision	Recall	F1 Score
0	MultinomialNB	0.634787	0.638434	0.637187	0.634437
1	SVC	0.609939	0.625712	0.616	0.604283
2	LogisticRegression	0.603376	0.632563	0.611855	0.590445
3	MLPClassifier	0.63338	0.635536	0.635092	0.633303

SUPERVISED LEARNING

	Classifier	Accuracy	Precision	Recall	F1 Score
0	MultinomialNB	0.490389	0.492066	0.492107	0.490172
1	SVC	0.492733	0.476555	0.482325	0.452982
2	LogisticRegression	0.509611	0.499513	0.499624	0.473999
3	MLPClassifier	0.492733	0.497126	0.497257	0.488804

THANK YOU