· a few seconds ago · 3 min read

Text Classification - A Quick Walkthrough!

INTRODUCTION:

This walkthrough uses the dataset from Kaggle|TopRedditPostsandComments .
The dataset contains Top_Posts.csv and Top_Posts_Comments.csv files. The objective is to build a classifier that can predict the class of the comment. In the dataset, our class is "subreddit" and its values are "MachineLearning" for Machine Learning, "datascience" for Data Science, and "artificial" for Artificial Intelligence. This blog uses three linear text classifiers and two non-linear classifiers. Linear Classifiers include: SVM (Support Vector Machine), Logistic Regression, and Naive Bayes classifiers. Whereas, the Non-Linear Classifiers include Random Forest and K-Nearest Neighbour classifiers.

LOAD DATASET:

Load the data set and merge the two in order to get the comments and subreddits (Classes) in the same dataset. The datasets are merged on the column values of "Post_id".

Top_Posts.csv:
Has 2,987 rows and 10 columns.

Data:

	post_id	post_title	subreddit	post_url	flair_text	score	comments	upvote_ratio	date-time	y
0	gh1dj9	[Project] From books to presentations in 10s w...	MachineLearning	https://v.redd.it/v492uoheuxx41	Project	7798	186	0.99	2020-05-10 13:19:54	2
1	kuc6tz	[D] A Demo from 1993 of 32-year-old Yann LeCun...	MachineLearning	https://v.redd.it/25nxi9ojfha61	Discussion	5851	133	0.98	2021-01-10 10:30:36	2
2	g7nfvb	[R] First Order Motion Model applied to animat...	MachineLearning	https://v.redd.it/rlmmjm1q5wu41	Research	4761	111	0.97	2020-04-25 04:27:23	2
3	lui92h	[N] AI can turn old photos into moving Images ...	MachineLearning	https://v.redd.it/ikd5gjlbi8k61	News	4688	230	0.97	2021-02-28 15:12:28	2
4	ohxnts	[D] This AI reveals how much time politicians ...	MachineLearning	https://i.redd.it/34sgziebfia71.jpg	Discussion	4568	228	0.96	2021-07-11 04:18:59	2
...	...	...	...	...	...	...	...	...	...	
2982	slx33m	We live in beautiful times where you can learn...	artificial	https://github.com/louisfb01/start-machine-lea...	Discussion	84	6	0.90	2022-02-06 13:50:02	2
2983	k9otbj	Yann LeCun's Deep Learning Course Free From NYU	artificial	https://www.i-programmer.info/news/99-professi...	News	78	1	0.97	2020-12-09 09:22:52	2
2984	k2orib	You Can Now Learn for FREE: 9 Courses by Googl...	artificial	https://laconicml.com/free-artificial-intellig...	Self Promotion	80	2	0.95	2020-11-28 14:43:43	2
2985	ex9w4w	Chatbot trained on "public domain social media...	artificial	https://ai.googleblog.com/2020/01/towards-conv...	news	80	10	0.97	2020-02-01 17:55:23	2
2986	efk5n3	Tesla's Neural Net can now identify red and gr...	artificial	https://www.teslarati.com/tesla-holiday-update...	NaN	80	10	0.89	2019-12-25 18:50:50	2

2987 rows × 10 columns

Top_Posts_Comments.csv:
Has 223,174 rows and 2 columns.

Data1:

	post_id	comment
0	gh1dj9	Twitter thread: [https://twitter.com/cyrildiag...
1	gh1dj9	The future 🤯
2	gh1dj9	Simple yet very useful. Thank you for sharing ...
3	gh1dj9	Almost guaranteed, Apple will copy your idea i...
4	gh1dj9	Ohh the nightmare of making this into a stable...
...	...	...
223169	efk5n3	LiDAR is mot powerful sensor for the auto driv...
223170	efk5n3	So it can now idenrify traffic lights? Musk pr...
223171	efk5n3	Hydranet bro!
223172	efk5n3	It even shows flashing yellow turn arrows.
223173	efk5n3	Ya just saw karpathy talk on hydra and pytorch.

223174 rows x 2 columns

Merged Dataset:

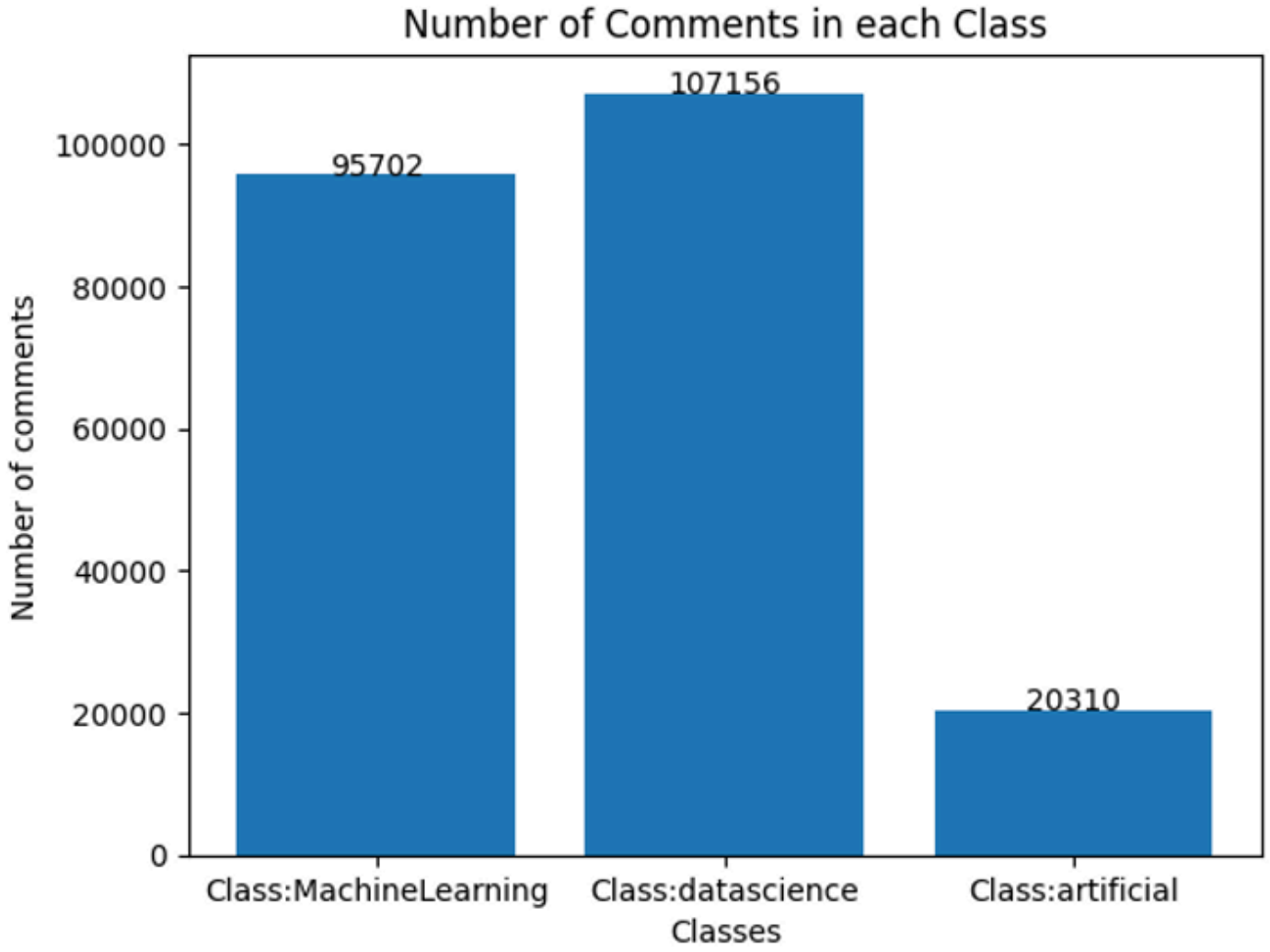
Has 223,174 rows and 11 columns.

Data2:

	post_id	post_title	subreddit	post_url	flair_text	score	comments	upvote_ratio	date-time	year	com
0	gh1dj9	[Project] From books to presentations in 10s w...	MachineLearning	https://v.redd.it/v492uoheuxx41	Project	7798	186	0.99	2020-05-10 13:19:54	2020	Twitter thread: [https://twitter.com/cyrilc
1	gh1dj9	[Project] From books to presentations in 10s w...	MachineLearning	https://v.redd.it/v492uoheuxx41	Project	7798	186	0.99	2020-05-10 13:19:54	2020	The futu
2	gh1dj9	[Project] From books to presentations in 10s w...	MachineLearning	https://v.redd.it/v492uoheuxx41	Project	7798	186	0.99	2020-05-10 13:19:54	2020	Simple yet very useful. Thank you for shar
3	gh1dj9	[Project] From books to presentations in 10s w...	MachineLearning	https://v.redd.it/v492uoheuxx41	Project	7798	186	0.99	2020-05-10 13:19:54	2020	Almost guaranteed, Apple will copy your id
4	gh1dj9	[Project] From books to presentations in 10s w...	MachineLearning	https://v.redd.it/v492uoheuxx41	Project	7798	186	0.99	2020-05-10 13:19:54	2020	Ohh the nightmare of making this into a sta
...	...	...	...	...	...	...	...	...	...	...	
223163	efk5n3	Tesla's Neural Net can now identify red and gr...	artificial	https://www.teslarati.com/tesla-holiday-update...	NaN	80	10	0.89	2019-12-25 18:50:50	2019	LiDAR is mot powerful sensor for the auto
223164	efk5n3	Tesla's Neural Net can now identify red and gr...	artificial	https://www.teslarati.com/tesla-holiday-update...	NaN	80	10	0.89	2019-12-25 18:50:50	2019	So it can now idenrify traffic lights? Mus
223165	efk5n3	Tesla's Neural Net can now identify red and gr...	artificial	https://www.teslarati.com/tesla-holiday-update...	NaN	80	10	0.89	2019-12-25 18:50:50	2019	Hydrane
223166	efk5n3	Tesla's Neural Net can now identify red and gr...	artificial	https://www.teslarati.com/tesla-holiday-update...	NaN	80	10	0.89	2019-12-25 18:50:50	2019	It even shows flashing yellow turn ar
223167	efk5n3	Tesla's Neural Net can now identify red and gr...	artificial	https://www.teslarati.com/tesla-holiday-update...	NaN	80	10	0.89	2019-12-25 18:50:50	2019	Ya just saw karpathy talk on hydra and py

223168 rows x 11 columns

The below graph shows the count of comments in each class:

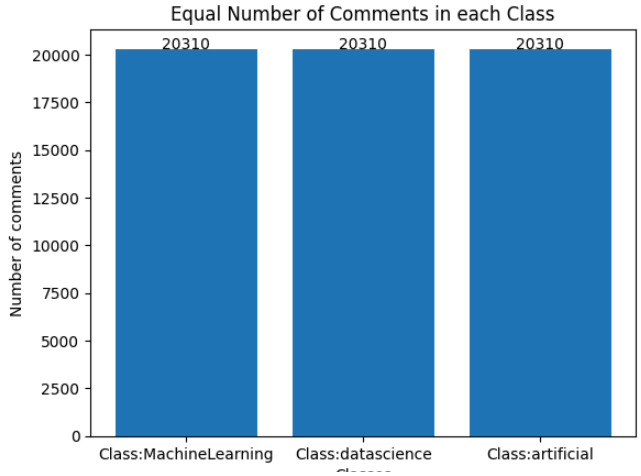


DATA PREPROCESSING:

Since we have 20,310 number of comments under artificial class, we consider it as max value of data set. This will make the number of comments in each class balanced.

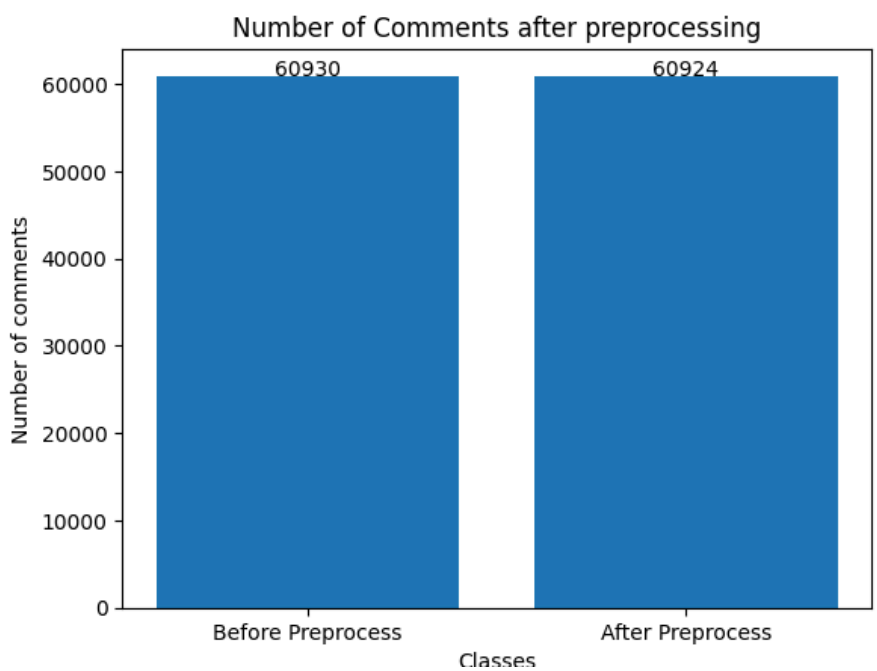
```
#Equalizing the count of class data
adata=data2[data2["subreddit"]=="artificial"]
L = len(adata)
mdata=data2[data2["subreddit"]=="MachineLearning"][:L]
ddata=data2[data2["subreddit"]=="datascience"][:L]
print(len(adata),len(mdata),len(ddata))
data_dict = {"Class:MachineLearning": len(mdata),"Class:datascience": len(ddata),"Class:artificial": len(adata)}
# Create bar chart
plt.bar(range(len(data_dict)), list(data_dict.values()), align='center')

# Set the x-ticks to be the keys of the dictionary
plt.xticks(range(len(data_dict)), list(data_dict.keys()))

# Display the values on top of each bar
for i, v in enumerate(data_dict.values()):
    plt.text(i, v + 1, str(v), ha='center')

# Add axis labels and title
plt.xlabel('Classes')
plt.ylabel('Number of comments')
plt.title('Equal Number of Comments in each Class')

# Show the plot
plt.show()
```

```
20310 20310 20310
```



DATA PREPROCESSING RESULT:

```
DF = pd.concat([mdata, adata, ddata])
DF = shuffle(DF, random_state=42)
DF
DF_before = len(DF)

# Preprocess data
DF.dropna(subset=['comment'], inplace=True)
vectorizer = TfidfVectorizer(stop_words='english')
X = vectorizer.fit_transform(DF['comment'])
y = DF['subreddit']

# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=47)
DF_after = len(DF)

data_dict = {"Before Preprocess": DF_before,"After Preprocess": DF_after}
# Create bar chart
plt.bar(range(len(data_dict)), list(data_dict.values()), align='center')

# Set the x-ticks to be the keys of the dictionary
plt.xticks(range(len(data_dict)), list(data_dict.keys()))

# Display the values on top of each bar
for i, v in enumerate(data_dict.values()):
    plt.text(i, v + 1, str(v), ha='center')

# Add axis labels and title
plt.xlabel('Classes')
plt.ylabel('Number of comments')
plt.title('Number of Comments after preprocessing')

# Show the plot
plt.show()
```



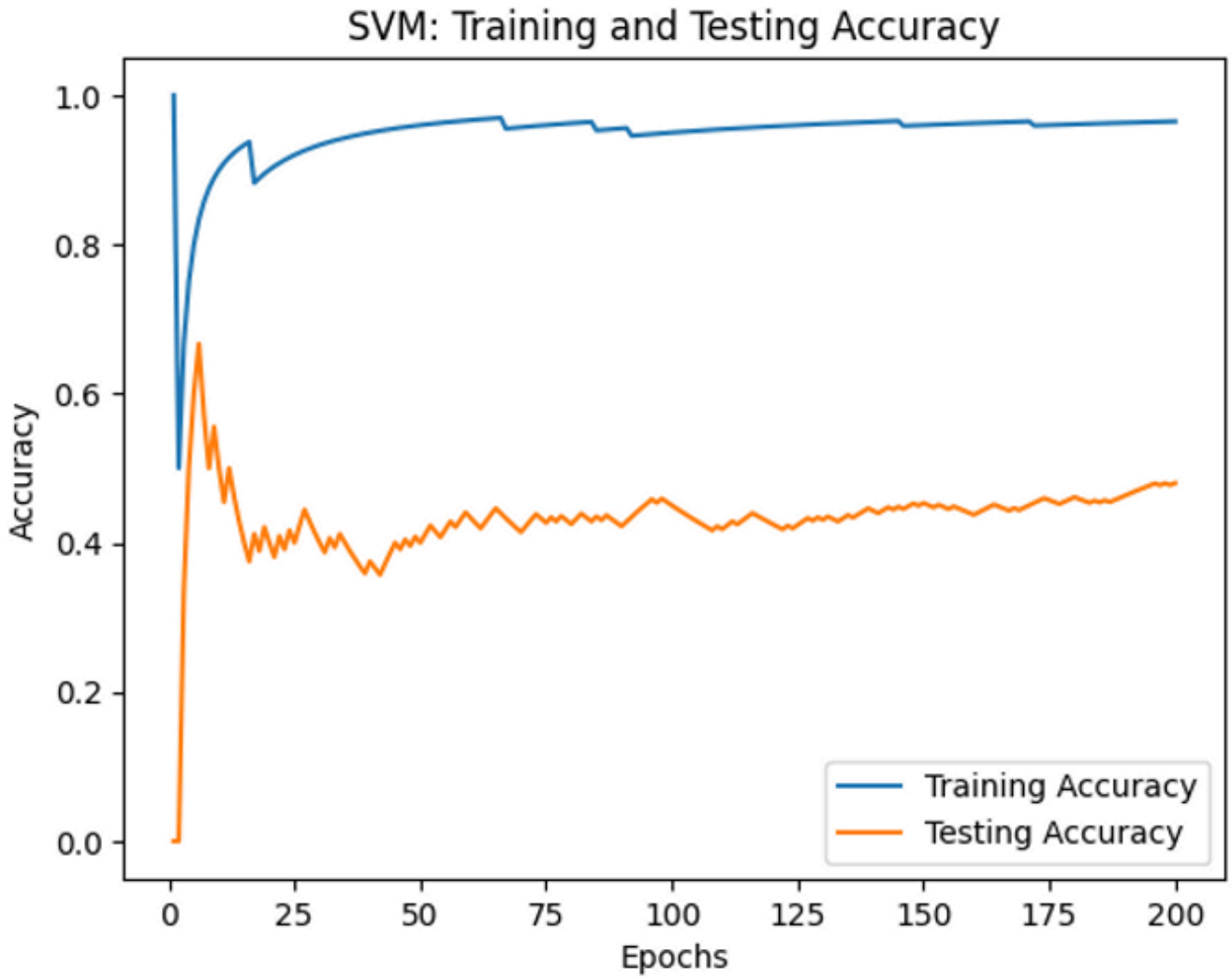
LINEAR CLASSIFIERS:

1. SVM (Support Vector Machines):

SVM is used to determine the hyperplane that separates the data. The data points close to the hyperplane are known as the Support Vectors. SVM uses kernels: Linear and Radial Basis Function rbf; used depending on the type of dataset; linearly separable or not. SVM has high cost of computation and the training time increases when a model is trained with large data. In SVM, we basically have five commonly used hyper parameters. i) C - for regularisation, ii) kernel - for transforming the input data, iii) gamma, iv) Degree - to specify the degree of polynomial kernel function and v) Coefficient - coeffficient parameter for polynomial or sigmoid kernal function.

The below graph shows the accuracy of the SVM model for the dataset:
Shows Overfitting: The Training accuracy > Test Accuracy.

```
204 204
```

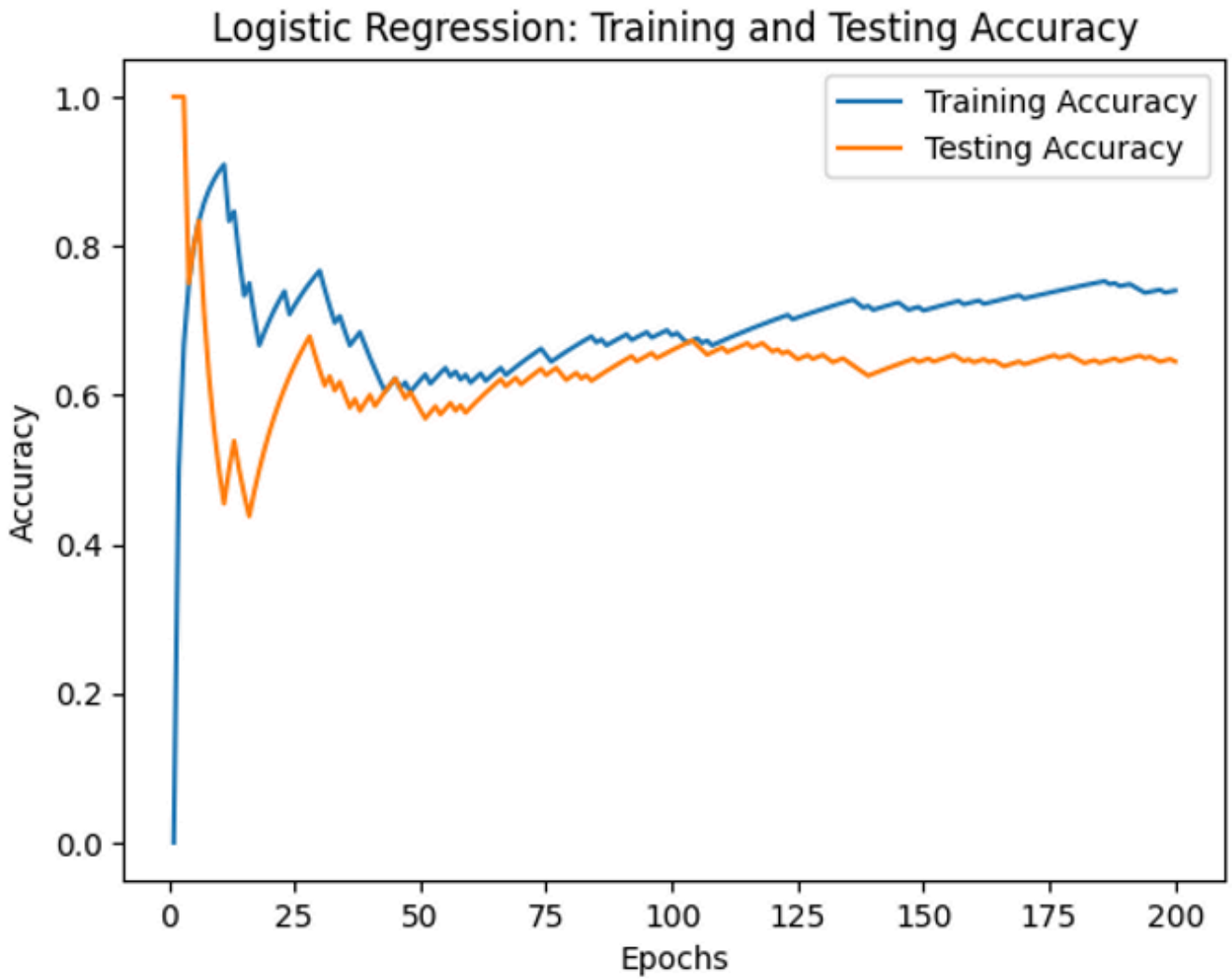


2. Logistic Regression:

Logistic Regression is used to predict the values based on more than one or even one predictor values. The output is a categorical value. The commonly tuned hyper parameters are: i) C - for regularisation, ii) max_iter - 100 by default, can be set according to the performance, iii) Class Weights and iv) Penalty.

The below graph shows the accuracy of the LR model for the dataset:
Shows the best fit: The Training accuracy is mostly similar to Test accuracy.

```
204 204
```

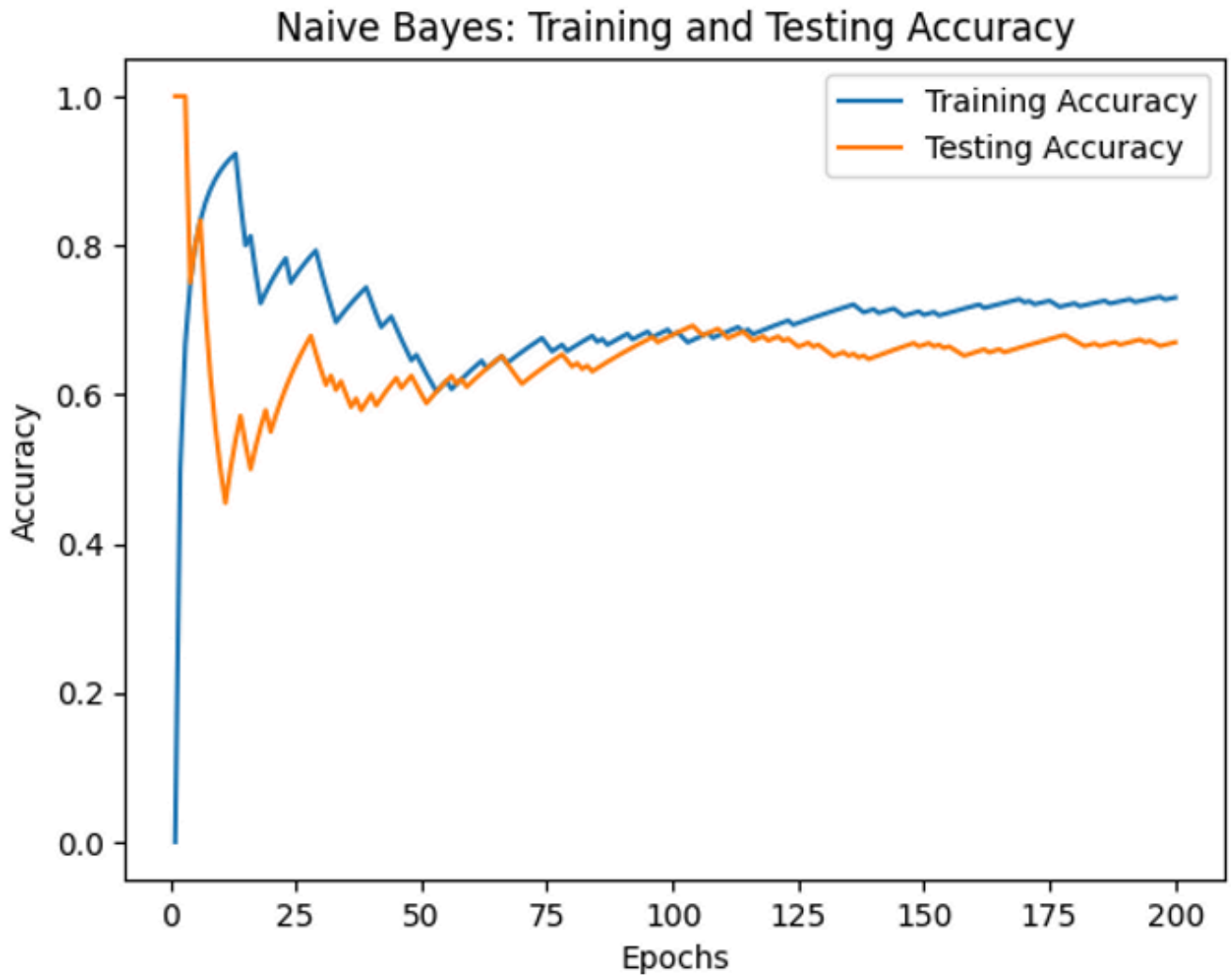


3. Naive Bayes Classifier:

Naive Bayes Classifier uses probability calculations and predicts the classes. There are three types: Bernoulli, Multinomial and Gaussian. The common hyper parameters are:
i) alpha, ii) Fit_prior and iii) Class_Probability. Only Multinomial Naive Bayes Classifier is implemented.

The below graph shows the accuracy of the NB model for the dataset:
Shows the best fit: The Training accuracy is mostly similar to Test accuracy.

204 204



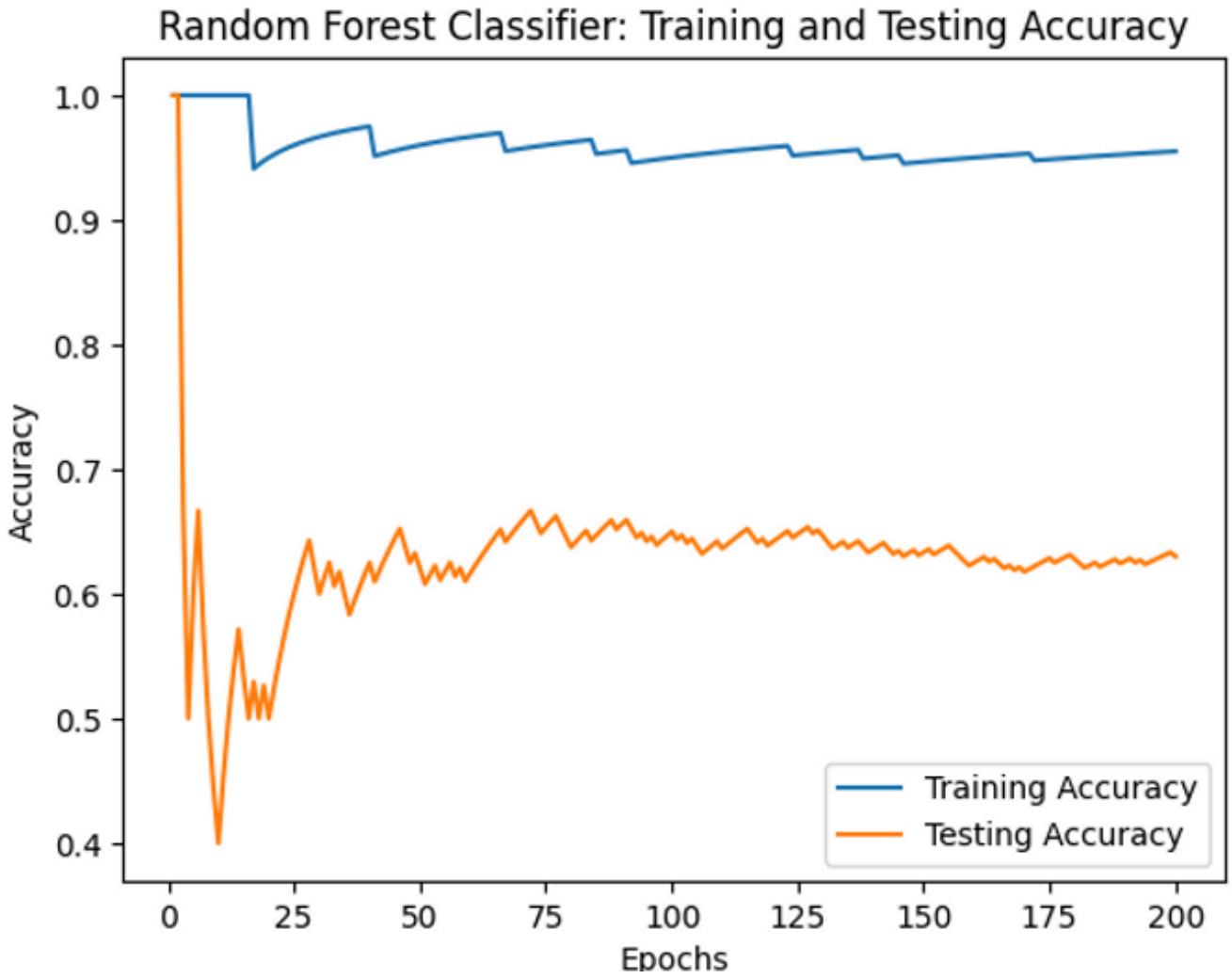
NON-LINEAR CLASSIFIERS:

1. Random Forest Classifier:

Uses decision trees to determine the classes. Its computational time is high. The hyper parameters are: i) max_depth, ii)n_jobs, etc.

The below graph shows the accuracy of the Random Forest Classifier model for the dataset:
Shows Overfitting, since the training accuracy is much greater than the testing accuracy.

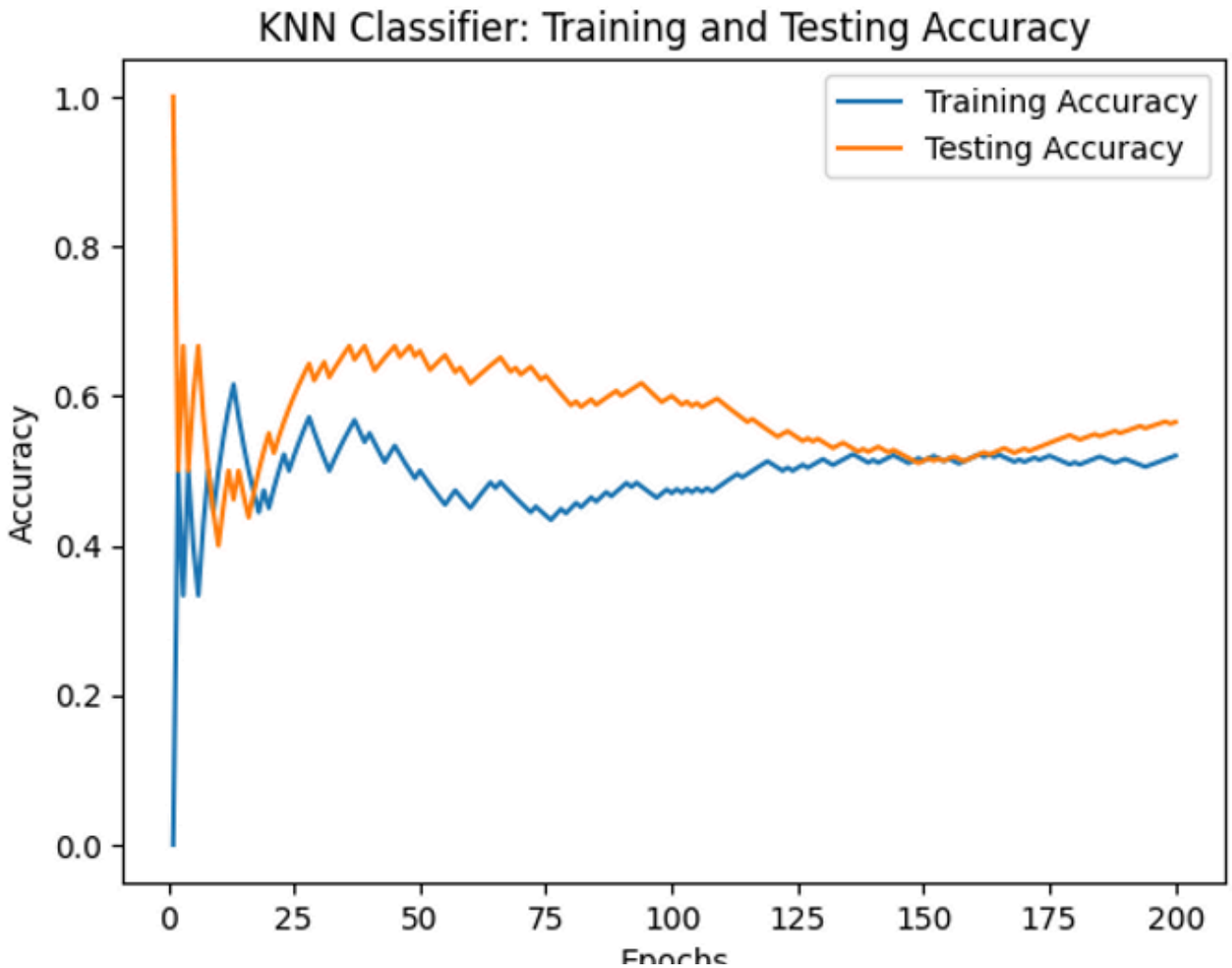
204 204



2. K-Neighbour Classifier:

Has K nearest neighbours to predict the class. It's used for both classification and regression. It's a slow training model and takes time to determine the best K-values for the optimal accuracy.
The below graph shows the accuracy of the KNN Classifier model for the dataset:
Shows the best fit: The Training accuracy is mostly similar to Test accuracy.

```
204 204
```



CHALLENGES:

Had to merge the dataset, remove the stopwords, NaN values. The dataset mostly preprocessed, but the Class: Artificial Intelligence had least number of Comments which would have created an imbalance in the training. Determining the K-value was time taking, K value selected was integer value of 25000/500+1.

CONTRIBUTIONS:

Implemented simple models using sklearn and used hyper parameter tuning to determine the best fit. Represented the training and testing accuracies in a graph for all the classification models used with the epoch size of 200.

CONCLUTION:

It can be concluded that withe hyper parameter tuning we get the best fit, we had the best performance with Naive Bayes Classifier and KNN Classifier. However, the accuracy was higher in Logistic Regression and Random Forest Classifier.

DOWNLOAD SOURCES:

DM_1002027304.ipynb
Download IPYNB • 429KB

REFERENCES:

1. "Introduction to Information Retrieval" by Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze
2. "Speech and Language Processing" by Daniel Jurafsky and James H. Martin
3. "Python Machine Learning" by Sebastian Raschka and Vahid Mirjalili
4. "Applied Text Analysis with Python" by Benjamin Bengfort, Tony Ojeda, and Rebecca Bilbro
5. "Text Analytics with Python" by Dipanjan Sarkar
6. "Natural Language Processing with Python" by Steven Bird, Ewan Klein, and Edward Loper
7. "Scikit-learn Documentation" by Scikit-learn contributors
8. "TensorFlow Documentation" by TensorFlow contributors
9. "PyTorch Documentation" by PyTorch contributors.