PRML Minor Project

Group:
Ankush Deshmukh(B21CS022)
Deepanshu (B21CS021)
Azhar Khan (B21CS087)

Topic: This dataset contains videos from 4 different YouTubers and all the comments made on those videos. The primary objective of this dataset is to cluster the comments to identify a cluster that contains all the spam comments and fix the issue once and for all.

About dataset: The data set contains videos from 4 different youtubers and a total of 861962 comments and 9 features.

DATA preprocessing:

1.handling null values: We printed the null values of different features and got the following results.

User	0
Video Title	149
Video Description	196767
Video ID	429330
Comment (Displayed)	467375
Comment (Actual)	482881
Comment Author	482909
Comment Author Channel ID	482868
Comment Time	482862

Since the data is all textual we can't replace it with the mean of the data that's why we decided to remove all the null values first.For this we used the

Comment (Actual) feature of the data and removed all the null values .After this almost most of the the null values were handled.

2.Selecting the relevant features: Since feature like comment Time and user ID had no significance we dropped down such columns and were left with:

['Video Title', 'Video Description', 'Video ID', 'Comment (Actual)', 'Comment Author Channel ID']

The final data set had:
Number of rows : 379073
Number of columns : 5
Number of Unique Videos After Preprocessing: 292

3.Textual pre-processing: For textual preprocessing we used NLTK library

In textual pre-processing we removed stop words,punctuations,numeric data,lemmatized the data and tokenized it. For doing all this we made our own function named text_process().

4.Vectorization of text: Using `TfidfVectorizer` we vectorized the actual comment.

5.Dimensionality Reduction : Truncated SVD is often preferred over PCA (Principal Component Analysis) in certain situations because it can handle sparse matrices, whereas PCA requires a dense matrix. This makes it especially useful in natural language processing (NLP), where the data is often represented as sparse matrices.

When we go further to discuss our ideas you will see that we are performing outlier detection hence Truncated SVD is preferred over PCA.

IDEAS: Before performing the problem we thought of various ideas to solve the problem all of them are listed below.

1.Performing basic clustering algorithms like Kmeans and selecting 2 clusters. But as we applied this approach we got to know that the outliers detected are the spam comments.Hence we focused our approach more towards outlier detection rather than clustering problem.

2.Our second Idea was to perform the clustering one video at a time and detect the spam for each and every video.

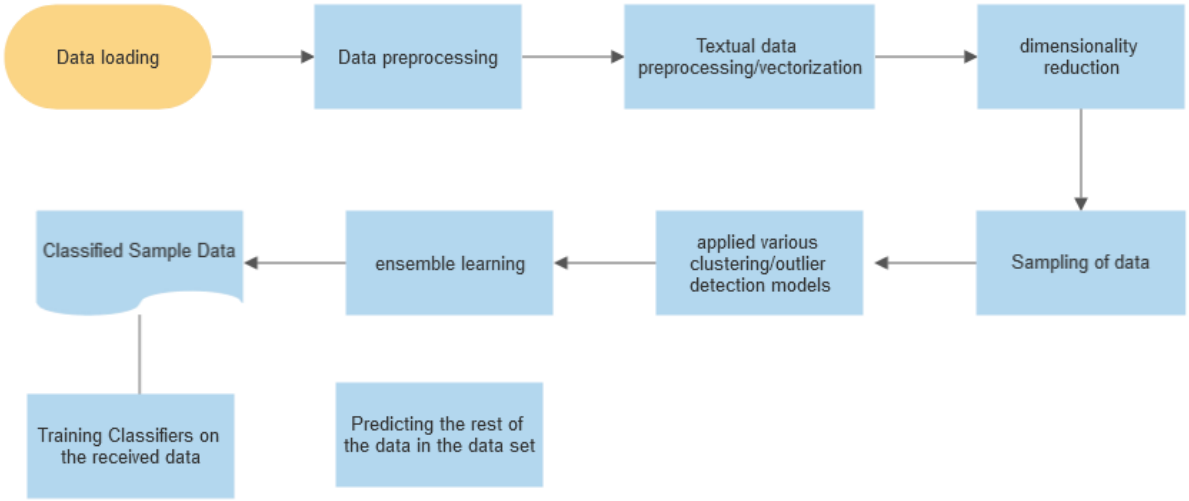
3.another idea we thought of was getting a correlation between vocabulary of the description of the videos and the comment and if some correlation is detected we classify it as a ham. But we found that as length of both description and comment is small it is unlikely that most comments will show any correlation.

When we approached the problem we noticed that due to a very large dataset our collab notebook kept on crashing.So we thought of sampling 1000 points from the data set appling outlier detection through various methods like Z-Score Method, LocalOutlierFactor Model, OneClassSVM Model and performing an **ensemble** of the models by voting. After classifying those 1000 comments we trained a **Decision Tree classifier** on it and obtained the trained classification model. We then labeled the rest of the comments using this model, somewhat of a **Semi-supervised learning approach.**

Similar problem was encountered when we performed for one video at a time that our collab notebook kept on crashing but with this method we are able to get a more accurate prediction of the spam as repeated comments were reported as outliers.

4.After the google collab kept on crashing we decided to label the data on 3 different machines by dividing the data into 3 different data with the respective videos so that we can label the data much faster as collab has limitations on hardware requirements.

PipeLine:



Process:

1) **Z-Score Method :** The z-score method is a statistical technique that is often used in machine learning to identify outliers in a dataset. This method measures the deviation of each data point from the mean of the dataset in terms of the number of standard deviations. We tried different values of threshold to get optimal spam comments.

2) **LocalOutlierFactor Model :** The Local Outlier Factor (LOF) unsupervised machine learning algorithm used for outlier detection. The LOF model identifies outliers by calculating the local density deviation of a given data point with respect to its neighbors. The intuition behind this approach is that outliers are typically located in sparse regions of the dataset, whereas inlier points tend to be clustered together. We set parameters n_neighbours and contamination which states a fraction of expected outlier points.
3) **OneClassSVM Model :** OneClassSVM, short for One-Class Support Vector Machine, is a popular unsupervised machine learning algorithm used for outlier detection. It belongs to the family of Support Vector Machines (SVMs) and is particularly useful for detecting outliers in high-dimensional datasets.
4) **Ensemble Method :** We got spam comments from all three above Models. Then we applied the Voting Method and got final spam comments detected by at least two models.
5) **Decision Tree Classification :** We trained the classifier on samples and outputs we got from above models and classified the rest of comments using this model. Finally we printed all spam comments. We also tried different classifiers like the Kth Nearest Neighbour, etc. But it didn't provide better results.