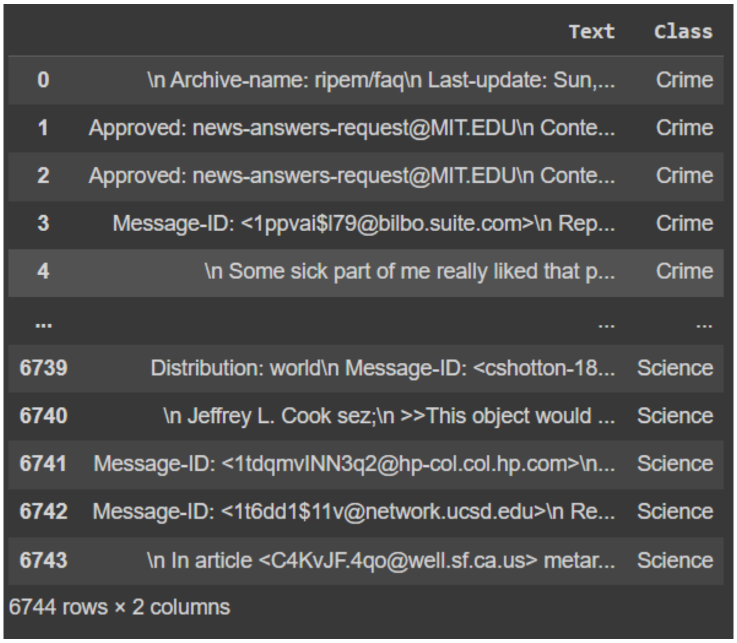
PRML MAJOR PROJECT
Text Classification on Emails

Drithi Davuluri (B21AI055)
Nishtha Karki (B21AI051)

- **Loading Data From Data Folder**
    1) Data was stored in four folders ,each having emails corresponding to that class.
    2) Extracted files from each folder, stored the content of file and name of the folder in a dictionary .
    3) Transformed this dictionary into a pandas dataframe having two columns-Text and Class.



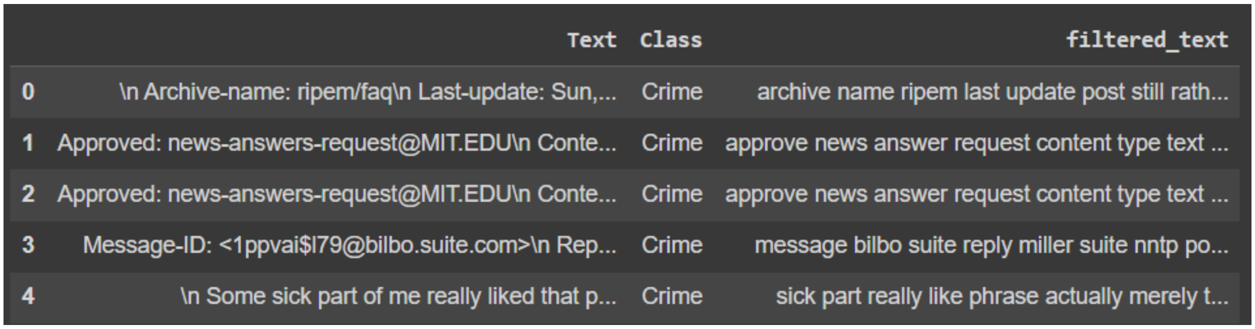
	Text	Class
0	\n Archive-name: ripem/faq\n Last-update: Sun,...	Crime
1	Approved: news-answers-request@MIT.EDU\n Conte...	Crime
2	Approved: news-answers-request@MIT.EDU\n Conte...	Crime
3	Message-ID: <1ppvai$l79@bilbo.suite.com>\n Rep...	Crime
4	\n Some sick part of me really liked that p...	Crime
...	...	...
6739	Distribution: world\n Message-ID: <cshotton-18...	Science
6740	\n Jeffrey L. Cook sez;\n >>This object would ...	Science
6741	Message-ID: <1tdqmvINN3q2@hp-col.col.hp.com>\n...	Science
6742	Message-ID: <1t6dd1$11v@network.ucsd.edu>\n Re...	Science
6743	\n In article <C4KvJF.4qo@well.sf.ca.us> metar...	Science

6744 rows × 2 columns

- **Cleaning of the Text**
    1) The text of emails contains a lot of unnecessary information. Hence, proper analysis requires cleaning of the text
    2) Imported nltk (Natural Language Toolkit) and regex (Regular Expression) libraries.
    3) Defined a list of stopwords so that we can remove words like and , or ,with etc.
    4) Defined a function which:
        - ❖ tokenizes the text into individual words using the WordPunctTokenizer module from the nltk library and converts them to lowercase for consistency.

- Uses regex to remove any non-alphabetic characters from the text and also removes any words that are less than 4 characters long.
- lemmatizes the remaining words using the WordNetLemmatizer class and filters out any stop words.
- Joined those individual words into a string

5)Applied the function to each row in the Text column of the data, and stored the filtered text in a new column called filtered_text.



	Text	Class	filtered_text
0	\n Archive-name: ripem/faq\n Last-update: Sun,...	Crime	archive name ripem last update post still rath...
1	Approved: news-answers-request@MIT.EDU\n Conte...	Crime	approve news answer request content type text ...
2	Approved: news-answers-request@MIT.EDU\n Conte...	Crime	approve news answer request content type text ...
3	Message-ID: <1ppvai$l79@bilbo.suite.com>\n Rep...	Crime	message bilbo suite reply miller suite nntp po...
4	\n Some sick part of me really liked that p...	Crime	sick part really like phrase actually merely t...

- **Raw Data Analysis**

1)found null values from the dataset to be zero

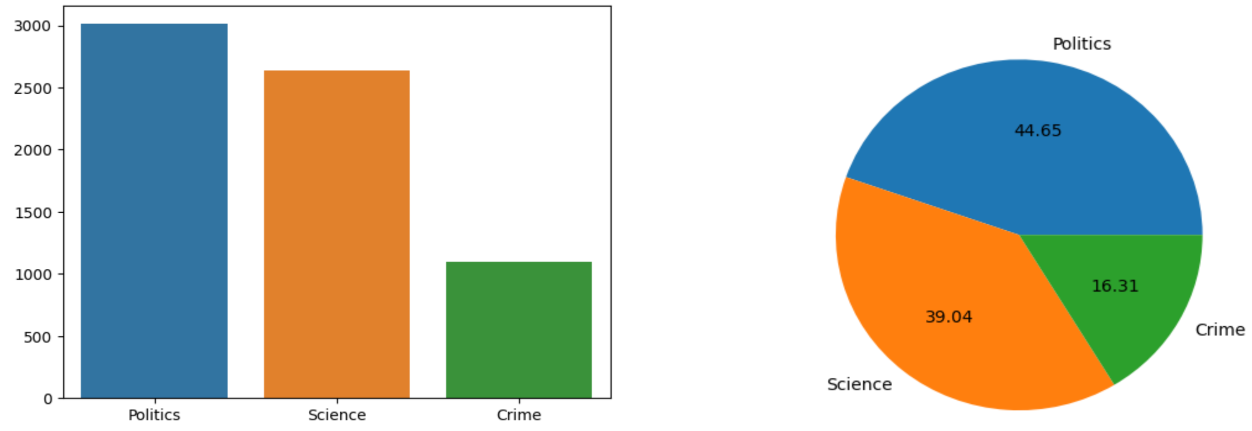
2)found duplicate values in the dataset.

3)decided a priority order for dropping duplicate rows as :'Crime','Politics','Science','Entertainment'

4)dropped the column having duplicate values and hence all the data points of the entertainment class were dropped.

Class distribution of dataset

Frequency wise **Percentage wise**



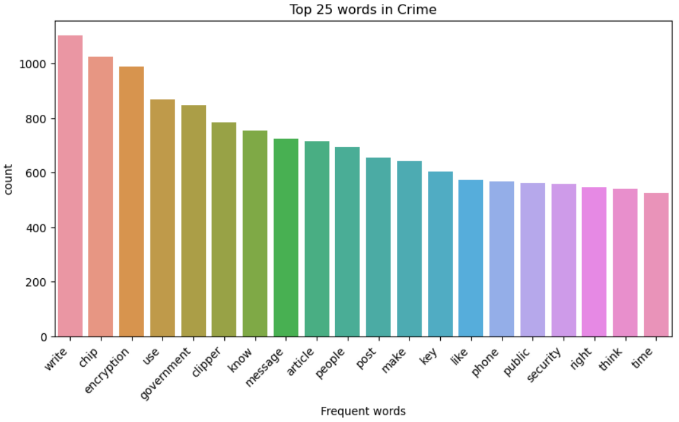
5) Downloaded the dataset from one notebook and loaded it on other because the RAM was insufficient to run the whole project in one notebook.

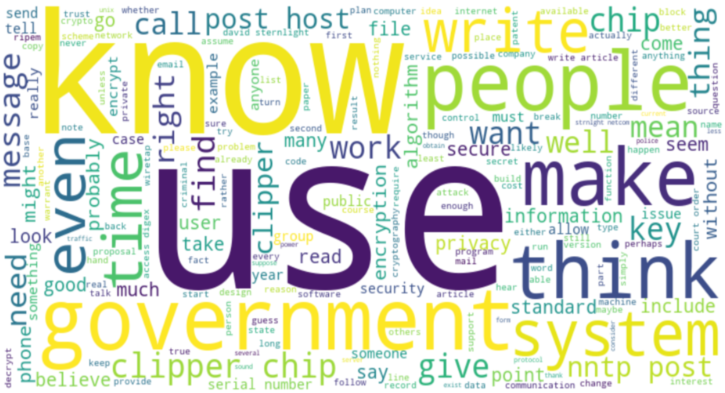
6) Formed a new dataset having Class values names as class_data.

7) Separated the filtered text column according to the class values and joined them all to form three paragraphs corresponding to each class.

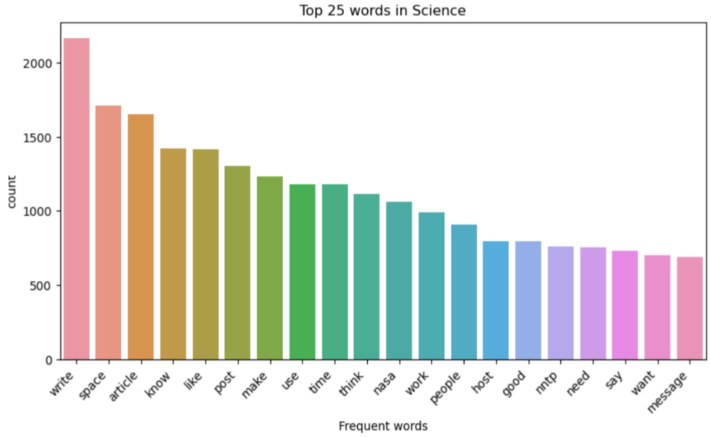
8) Found top 25 most occurring words in each class.

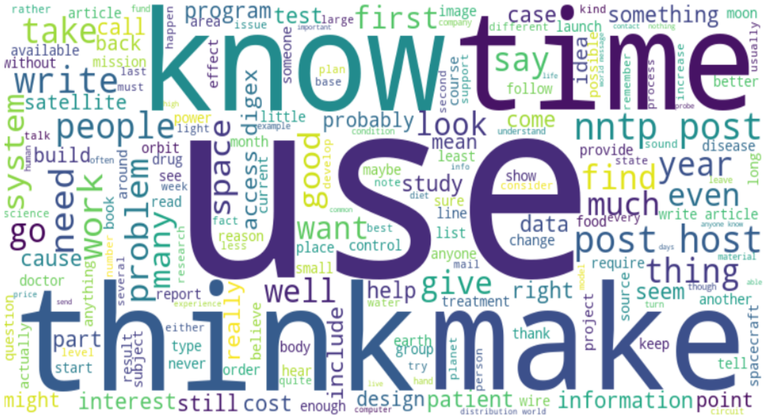
CRIME



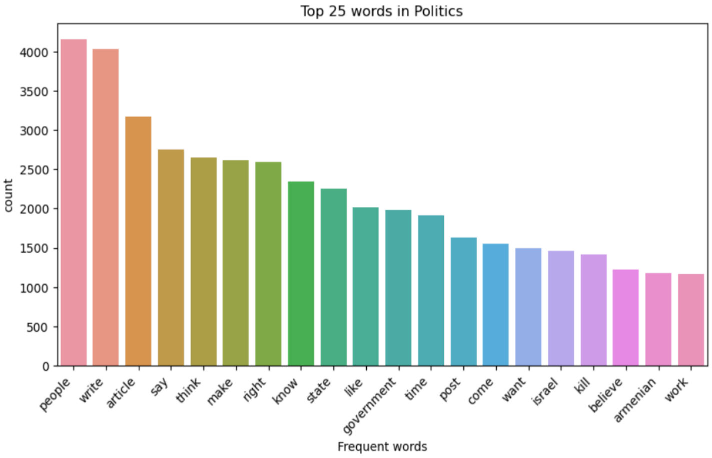


SCIENCE





POLITICS

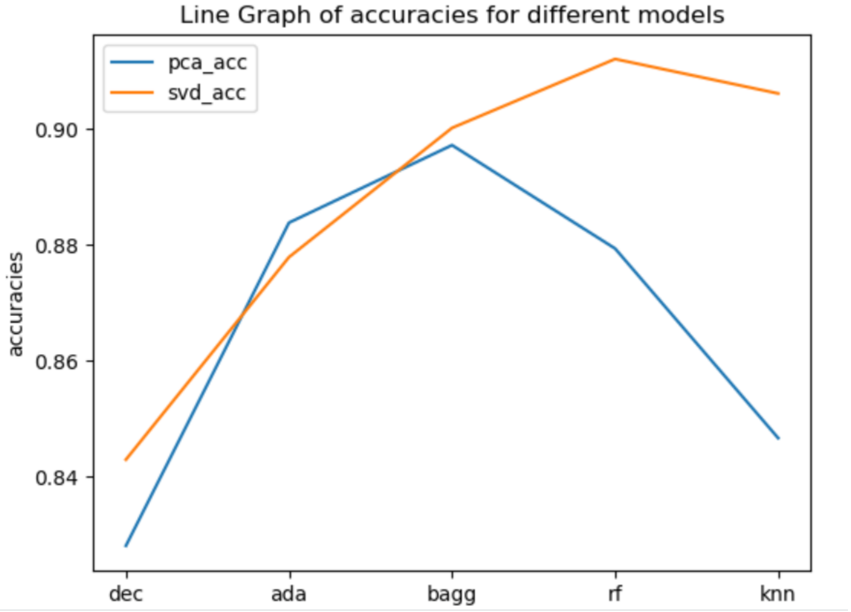


people
say
think
know
make
go
come
time
government
want
even
armenian
israel
write article
take
right
give
point
thing
state
tell
mean
find
year
kill
well
fact
look
still
believe
claim
use
need
leave
work
call
talk
first
back
start
report
arab
many
seem
fire
good
post host
issue
force
much
show
case
keep
hear
read
jews
israeli
someone
control
something
another
anything
anyone
question
problem
support
country
place
part
help
president
attack
person
really
american
happen
nntp post
children
include
follow
consider
group
number

- **Transformation into Numerical Data**
  1) Used TFIDF Vectorizer to transform the data into a matrix of TF-IDF features.
  2) Data is transformed into dataset having 43,129 features which has very high computational cost
- **Data Reduction**
  1) Reduced the dataset using two methods - PCA and SVD
  2) Selected top 400 components in PCA.
  3) Selected top 50 components in SVD and applied feature selection to reduce the number of features to 15.
  4) Divided the dataset into train and test
  5) Applied decision tree , random forest, knn and adaboost for checking which model works the best.

Classifier	PCA	SVD
Decision Tree Classifier	0.828	0.845
Adaboost with Decision tree as Base classifier	0.885	0.878
Bagging with Decision Tree	0.892	0.893
Random Forest	0.874	0.908
KNN with Voting Classifier	0.845	0.911

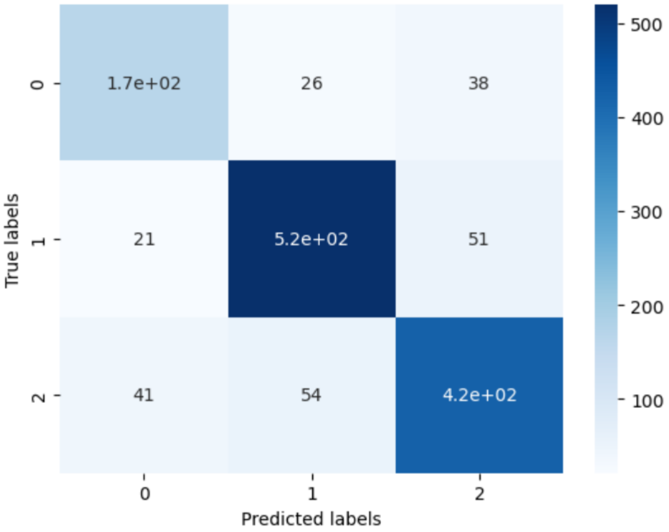
- **Accuracy Plot**

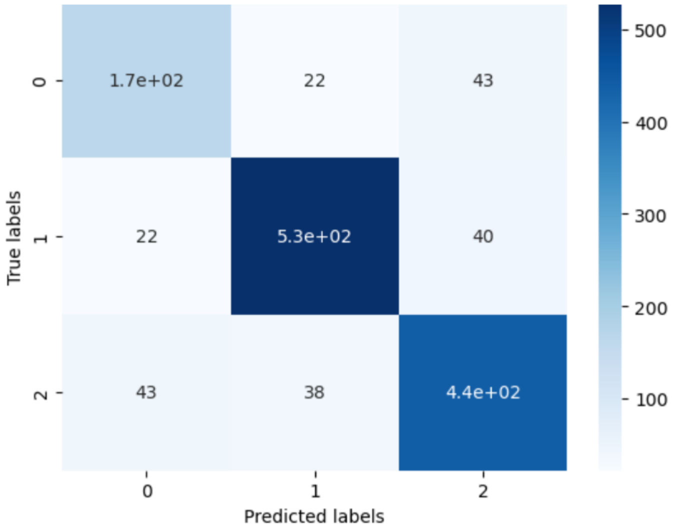


- **Confusion Matrix**

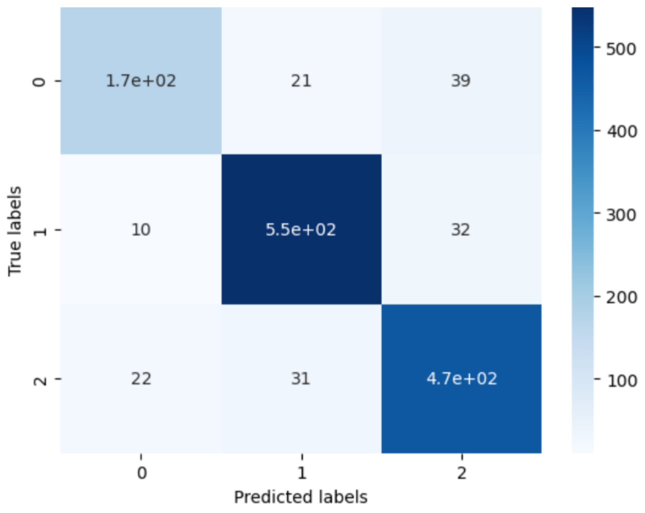
PCA **SVD**

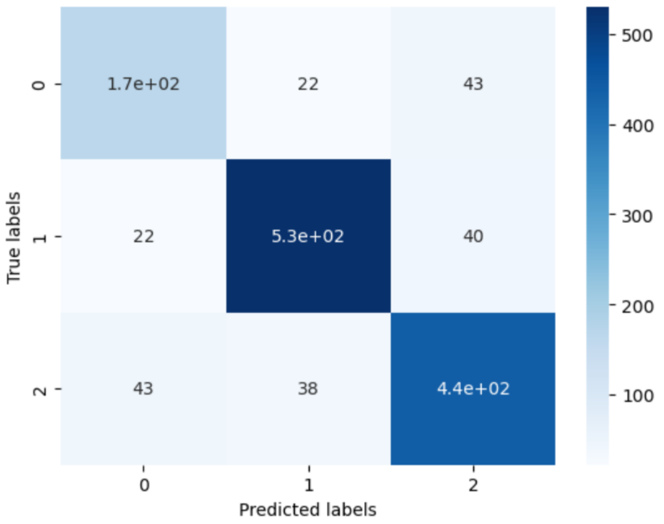
Decision Tree



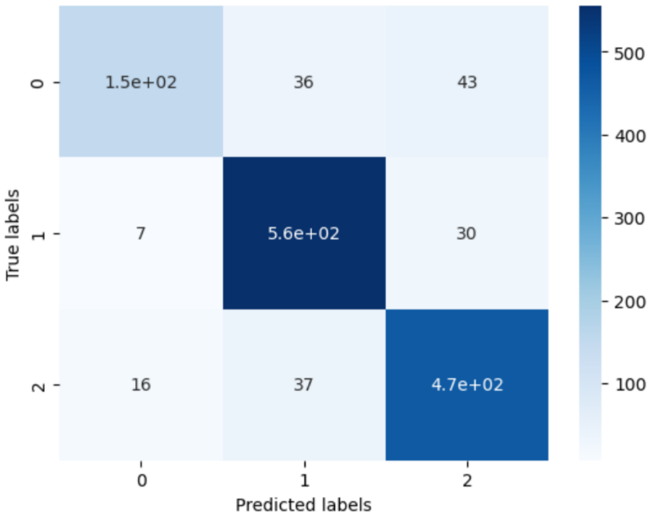


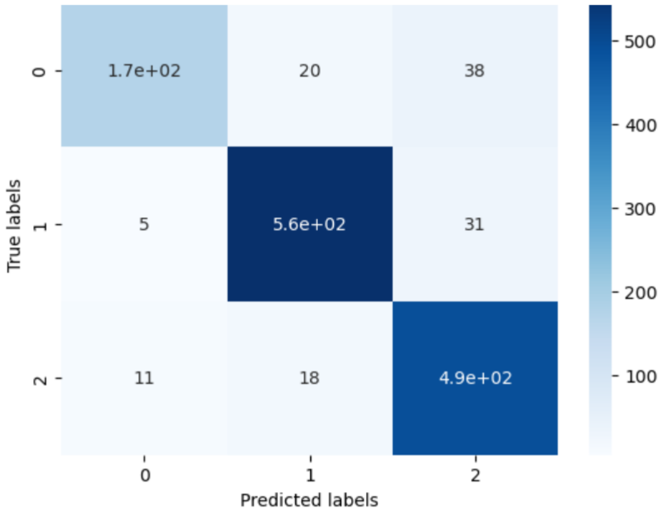
Adaboost



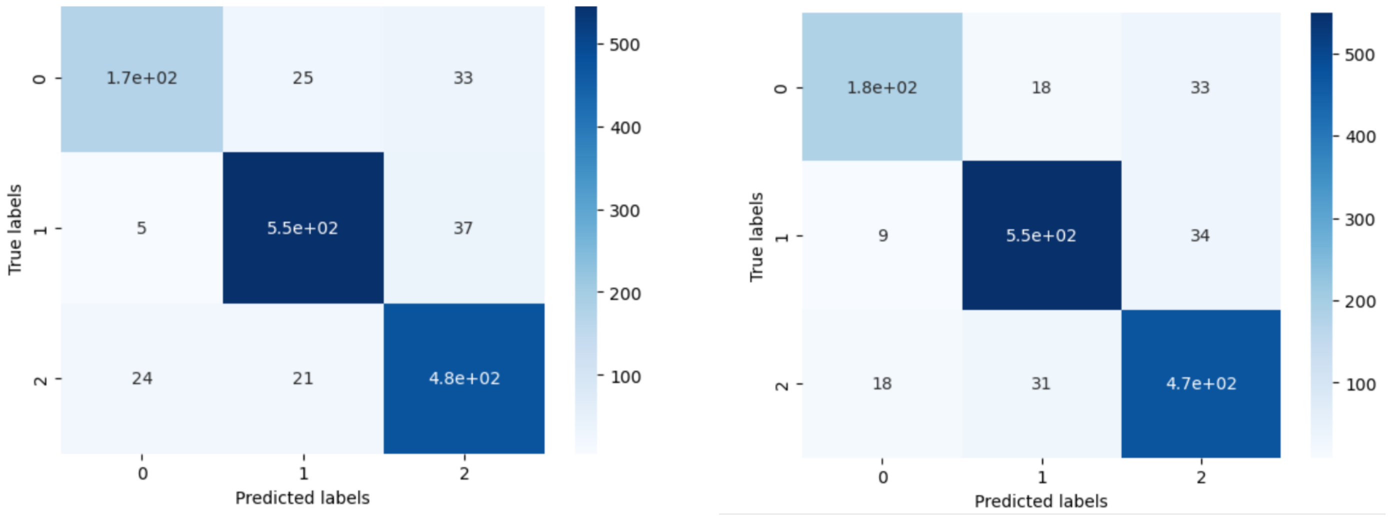


Random Forest

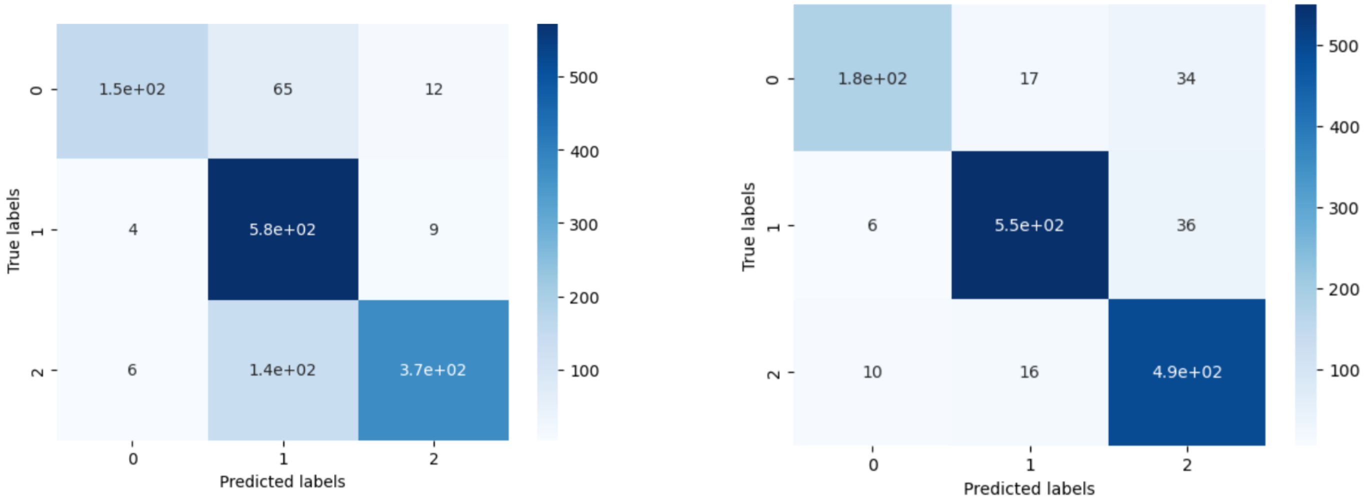




Bagging



KNN



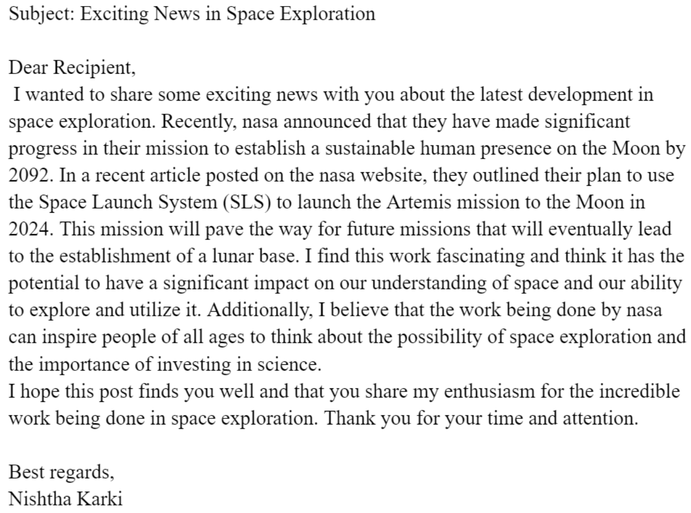
- **Observation**

  1) From here, we can observe that KNN with Voting classifier works the best for our dataset which has been reduced using SVD.

  2) Hence, to test a new email , we will use the same.

- **Testing on Unknown Emails**

1) Here we wrote a random email belonging to the Science Category as follows.



Subject: Exciting News in Space Exploration

Dear Recipient,
I wanted to share some exciting news with you about the latest development in space exploration. Recently, nasa announced that they have made significant progress in their mission to establish a sustainable human presence on the Moon by 2092. In a recent article posted on the nasa website, they outlined their plan to use the Space Launch System (SLS) to launch the Artemis mission to the Moon in 2024. This mission will pave the way for future missions that will eventually lead to the establishment of a lunar base. I find this work fascinating and think it has the potential to have a significant impact on our understanding of space and our ability to explore and utilize it. Additionally, I believe that the work being done by nasa can inspire people of all ages to think about the possibility of space exploration and the importance of investing in science.
I hope this post finds you well and that you share my enthusiasm for the incredible work being done in space exploration. Thank you for your time and attention.

Best regards,
Nishtha Karki

2) Predicted the Class of this dataset using the model built - SVD - feature selection and then applying KNN using Voting Classifier.

3) Actual class -Science, Predicted Class- Science. Hence, our model is able to classify a random email correctly.

❖ Two ipynb notebooks have been submitted because of the very large computation due to which it was not running completely in one file.

❖ Dataset has been made and downloaded in one and loaded and modeled on another.

- **Contribution-**

Both of us worked together until importing files and making a dataset with actual text and filtered text columns, and visualizing frequently occurring words.
Drithi worked on vectorization and training models with dataset reduced using PCA where n_components=400.
Nishtha worked on training models with dataset reduced using SVD, feature selection, where the features were reduced to 15.
Both worked on testing a new mail using the model that was giving a good accuracy.