Automatic Personal Email Organizer

Mohul Maheshwari
2018A7PS0229P

Naman Goenka
2018A7PS0398P

Ayush Sharma
2018A7PS0326P

Data collection Technique :- Firstly we created six different custom labels in our gmail account and added relevant mails in each label. Then, using google takeout we extracted the mbox file for every category. Using a parsing script, we converted the mbox into parsed csv email files.

Details about the Dataset :- Each one of us pooled 1000 mails from their inboxes pertaining to different categories such as

1) **educational promotions -** Mails from educational websites such as newsletters, online software like Pocket, Canva, Finshots, The economic times, hackerearth.
2) **campus events -** Mails from administrations, clubs and departments for giving updates regarding events going on on campus.
3) **Subjects -** Mails from professors regarding information related to enrolled CS and EEE courses such as Machine Learning, microprocessor, etc.
4) **Operational** :- Mails focusing on user gmail account operations such as confirming registration, google activity, google calendar, sign ins and sign ups.
5) **Job / Intern opportunities**. :- Mails updating students about the new job or internship opportunities from Placement unit, Superset, IPCD, etc.

Preprocessing

1) Erroneous, null value containing and duplicate rows are removed.
2) Categorical features such as labels, from are encoded.
3) Text (subject and body) preprocessing : Additional spaces, hyperlinks, punctuations, special characters, numbers are removed. Text is lowercased and stop words and non-english words are removed. Further, lemmatization with the help of position tags is performed.

Problem Modelling:-

We formulated the problem statement as a multi classification problem, after preprocessing we trained and tested the dataset on different combinations of machine learning models and language models such as 1-gram, 2-gram, 3-gram. After training and testing our models we did a comparative analysis using various metrics such as confusion matrix, f1 score, accuracy, etc.

Machine Learning Models Used :-

- **Multinomial Naive Bayes**
- **Logistic Regression**
- **Decision Tree**
- **Linear SVM**
-

Highest macro-weighted f1 score was achieved using Logistic regression followed by SVM. And unigram language models generally performed better than the bigram and trigram models.