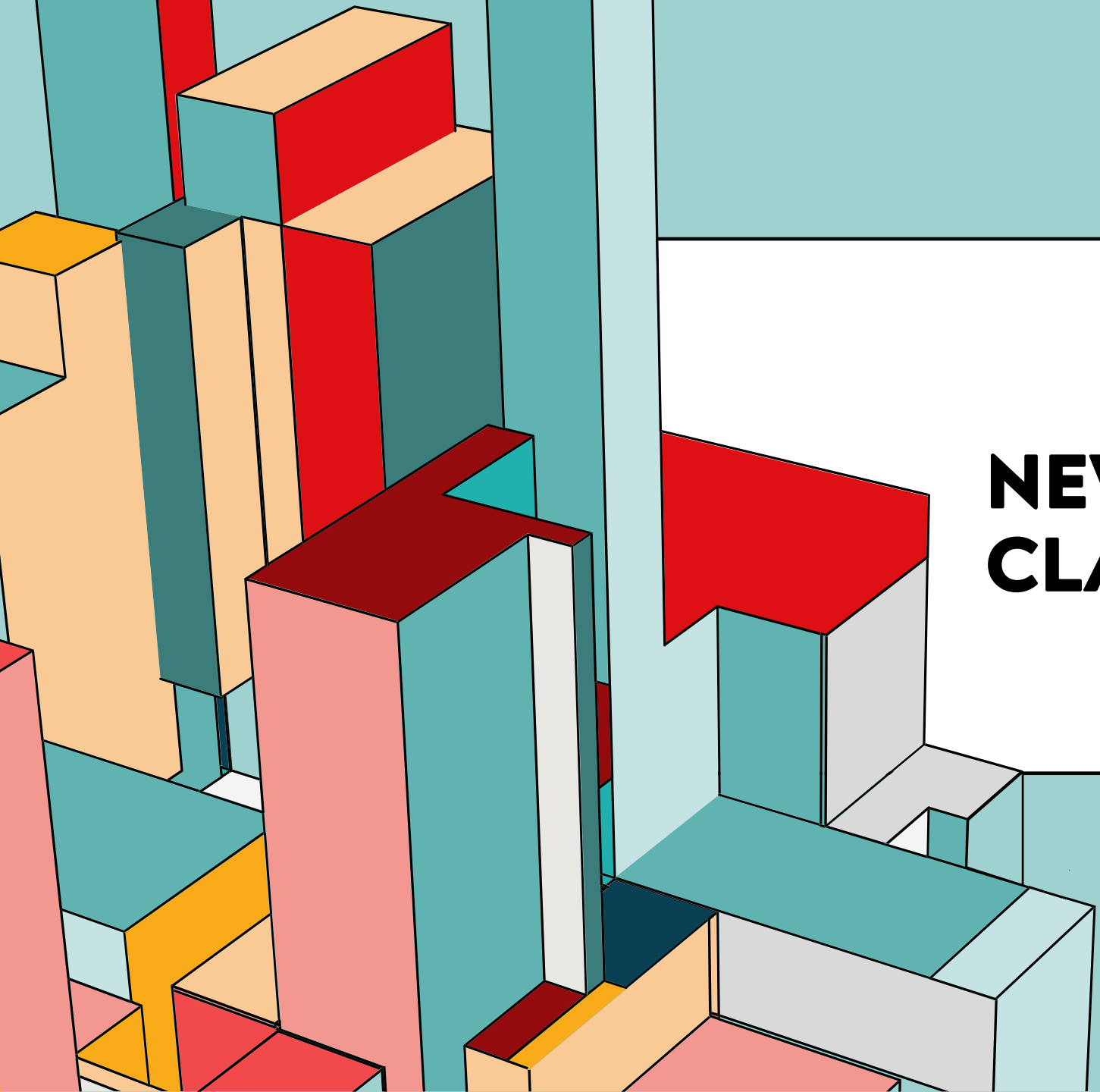
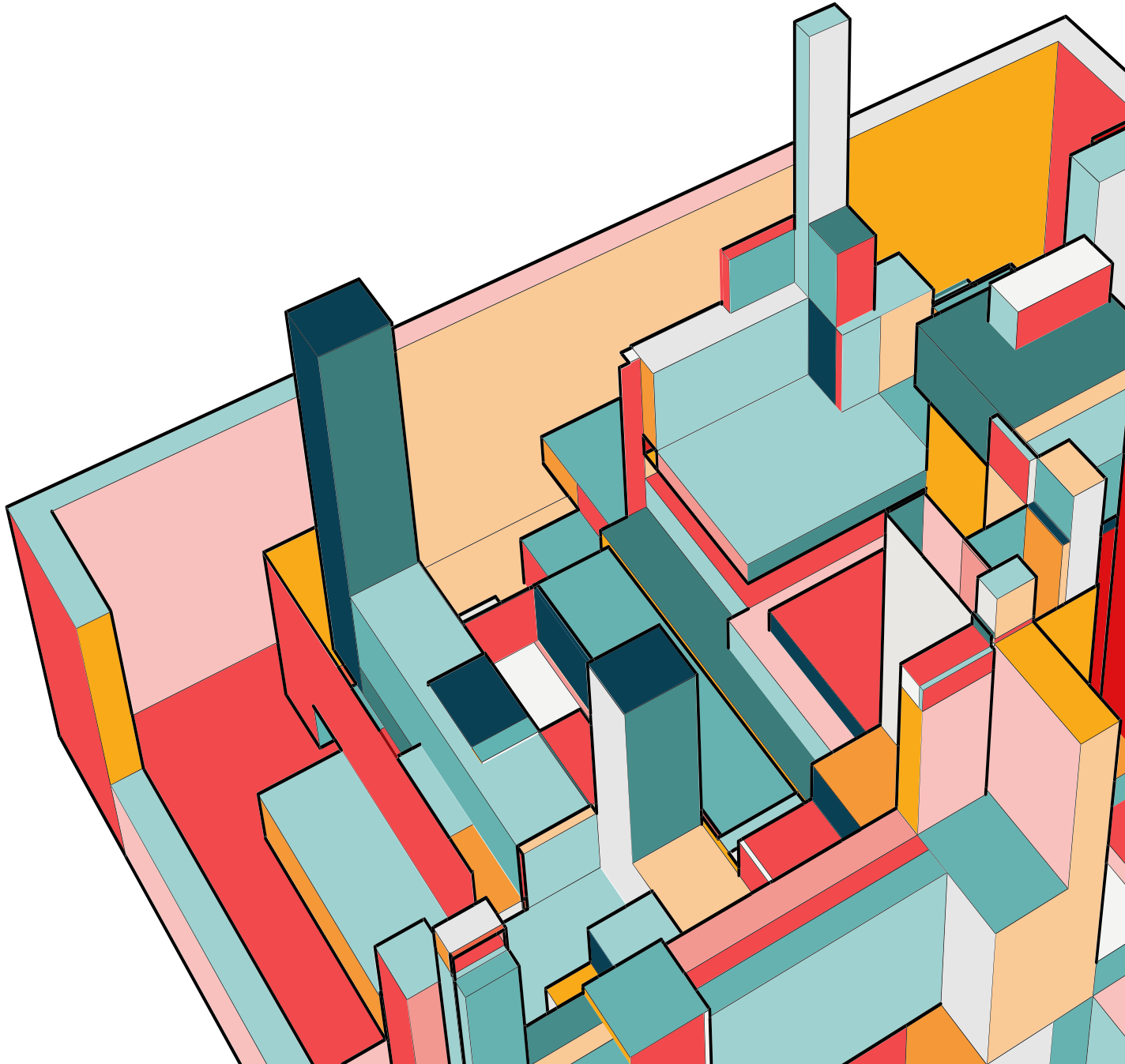
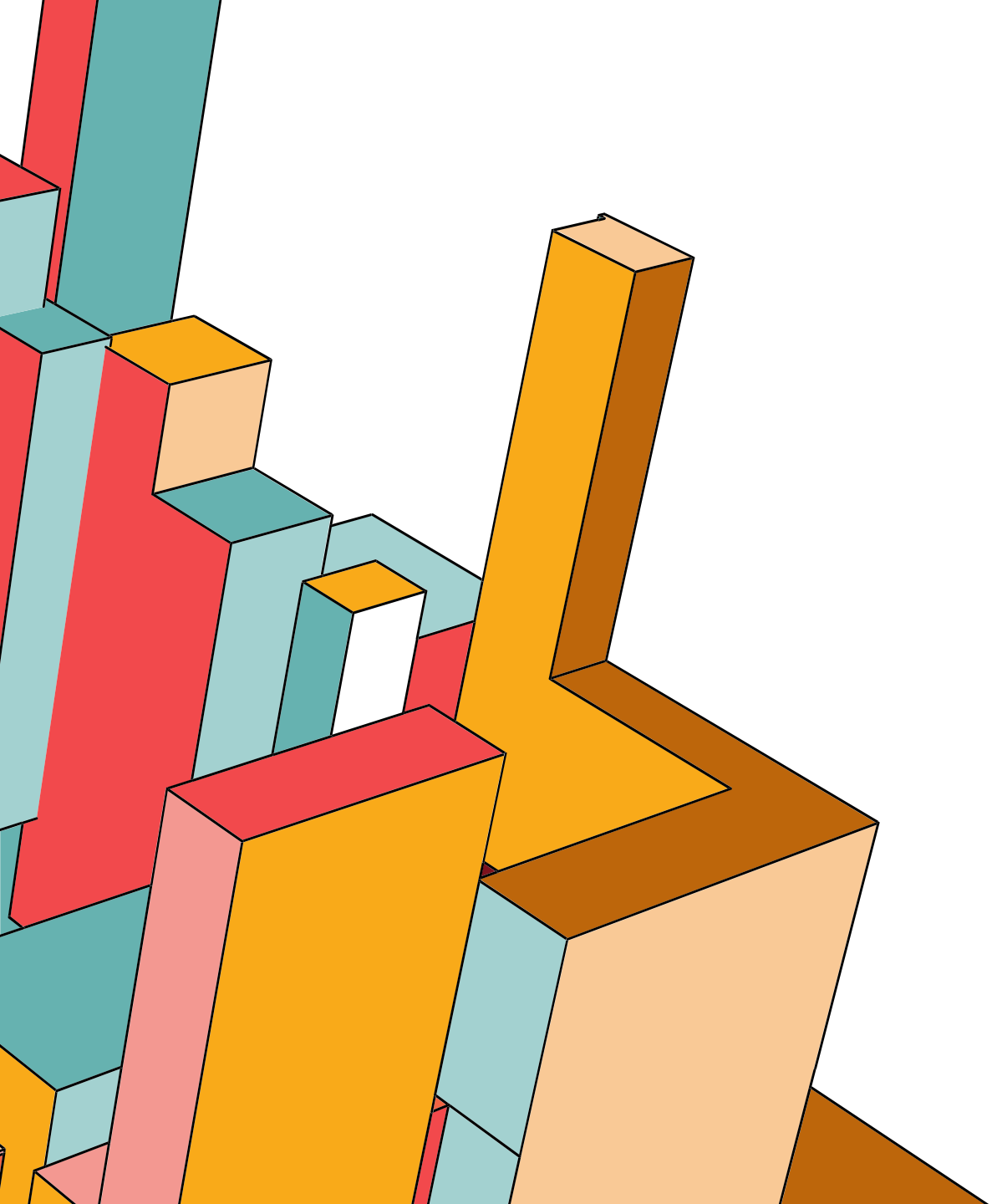


NEWS GROUP CLASSIFICATION

ABOUT US

- اسراء علي رياض عبد الحافظ
  2021170068
- بسمله نعيم عبد الحكيم عبد الوهاب
  2021170116
- الاء علاء عاشور محمد
  2021170078
- احمد هيثم احمد السرسي
  2021170052
- احمد اشرف رفاعي عبد السلام
  2021170012





STEPS

Read Data

Data preprocessing

Features extraction

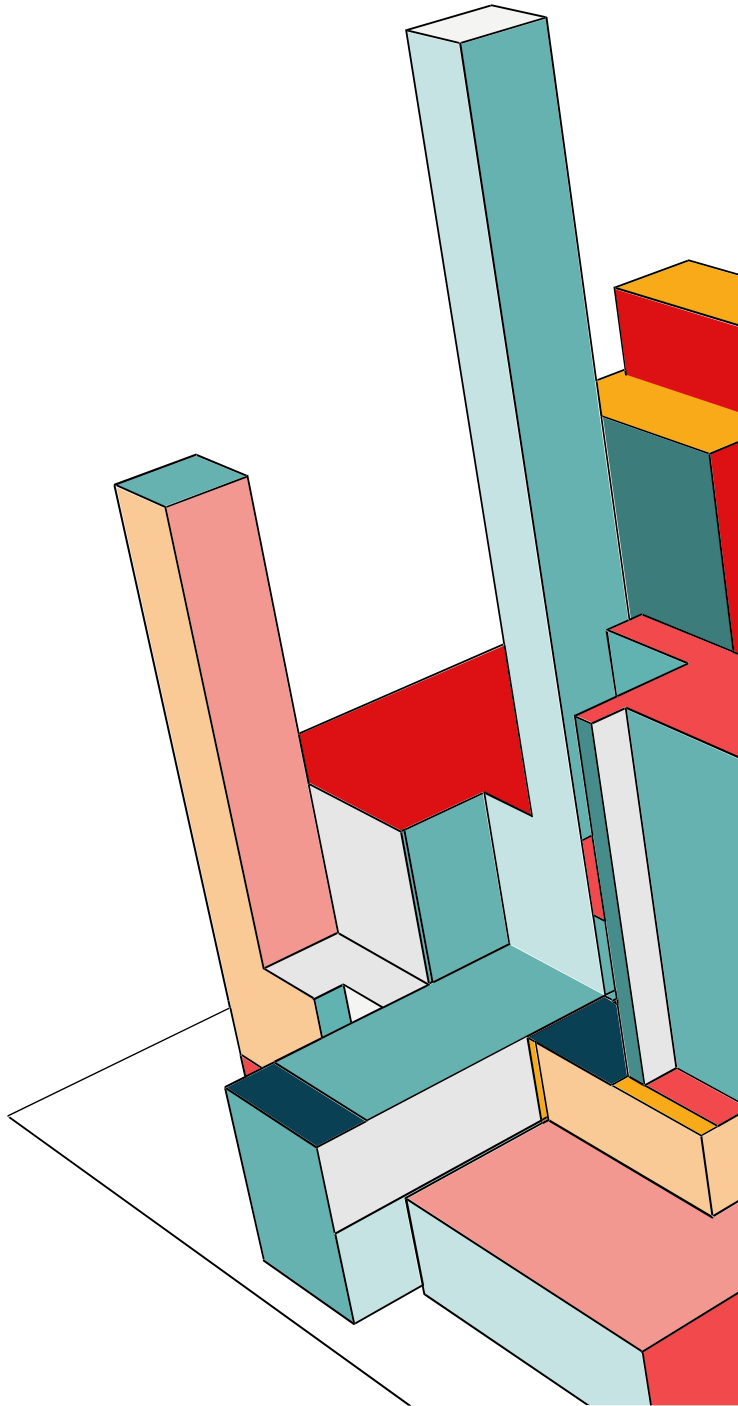
Model training and testing

Visualizing Results

Saving the models

READ DATA

Make a loop on the folder then make another loop inside each folder that loop in each file then read the content , write the target (folder name) , put them in a list then put them in larger list and finally convert the bigger list to DataFrame



DATA PREPROCESSING

Remove Stop Words

Make the article in a lower case then remove stop words

Remove lines that ending with words like : writes , wrote

Remove lines that starting with words like : from , subject , archive-name , last-modified , version , > in article , alt-atheism-archive-name: , -----begin pgp signed message-----

Remove lines after pgp

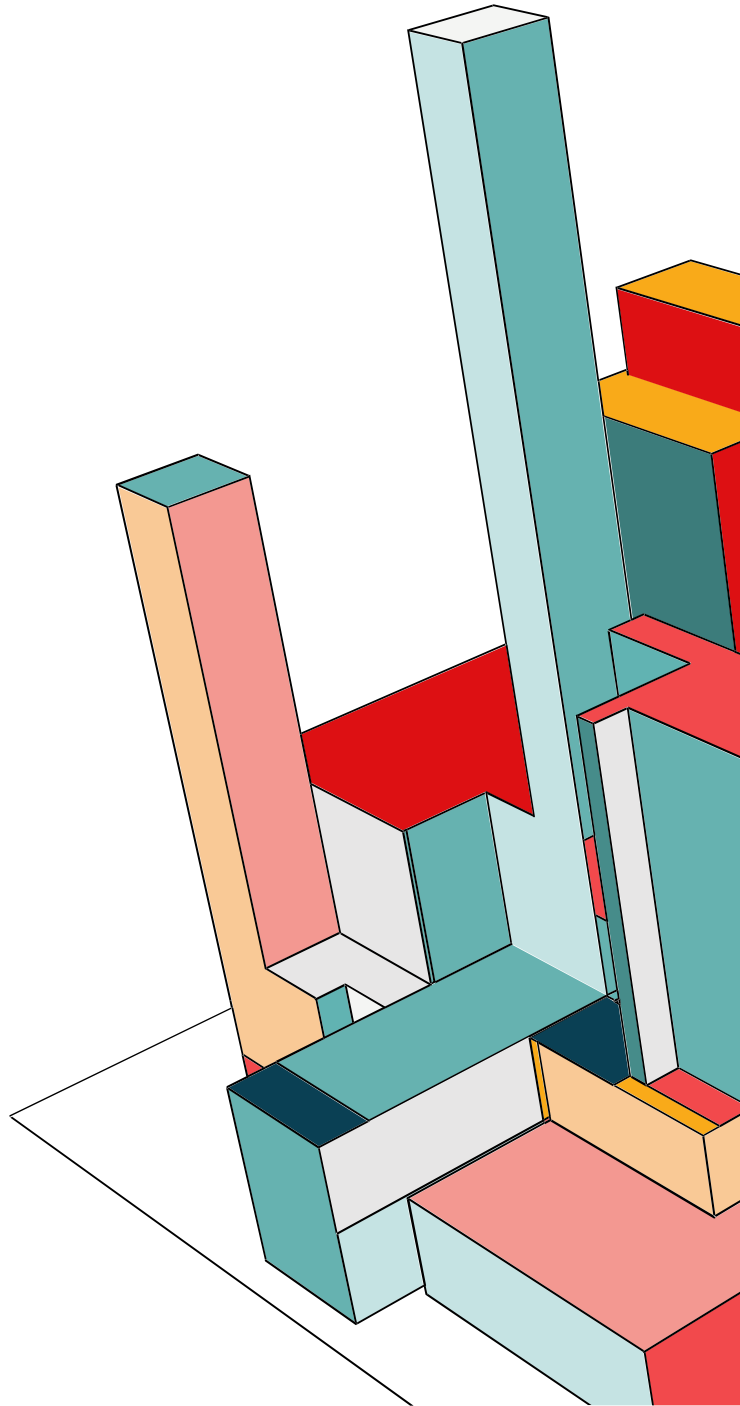
Remove lines that starting with " -----BEGIN PGP SIGNATURE----- " and remove the lines that follow it

Remove line that starting with regular expression : ' .*@,* '

This pattern essentially matches any string that contains the "@" symbol

Remove non_alphanumeric

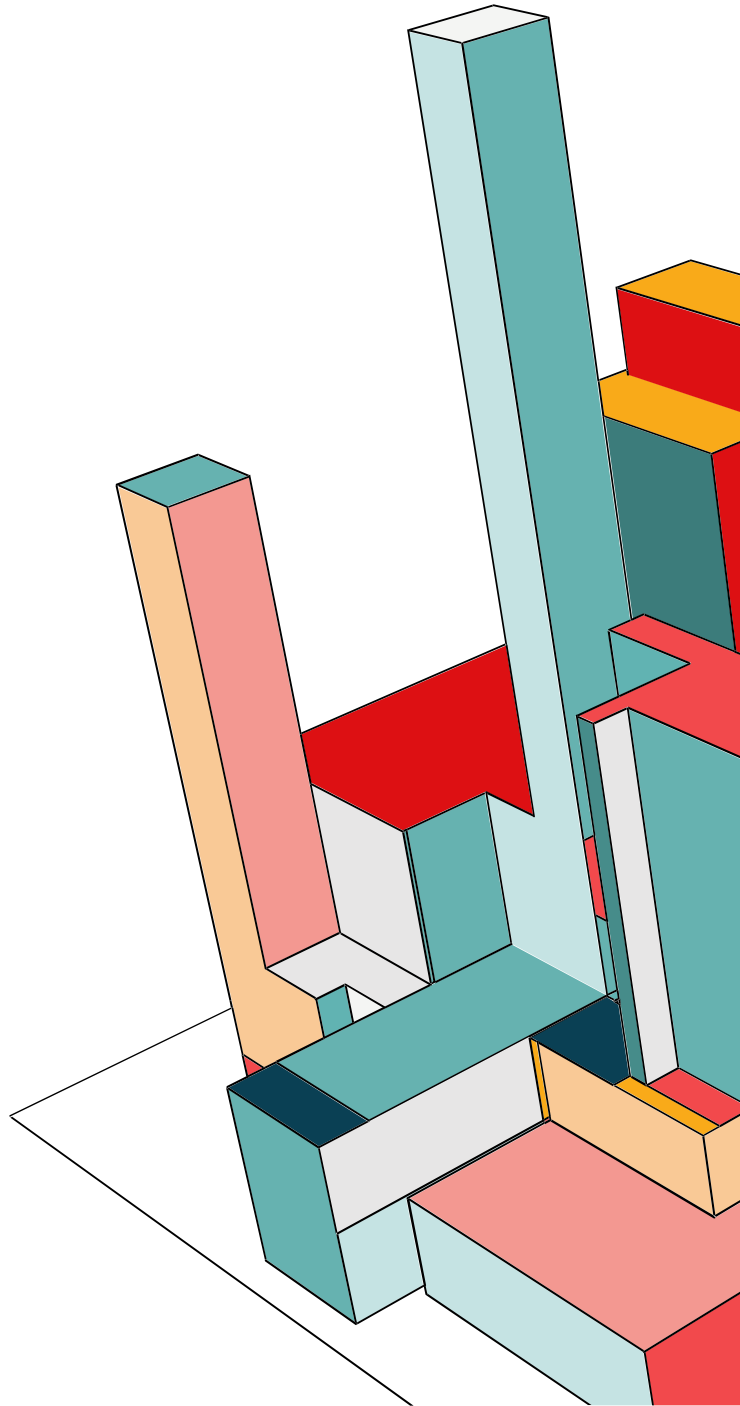
Also Use PorterStemmer



FEATURE EXTRACTION

TF-IDF

Appling TF-IDF algorithm to extract the words that belong to a specific topic and give them weight for each word



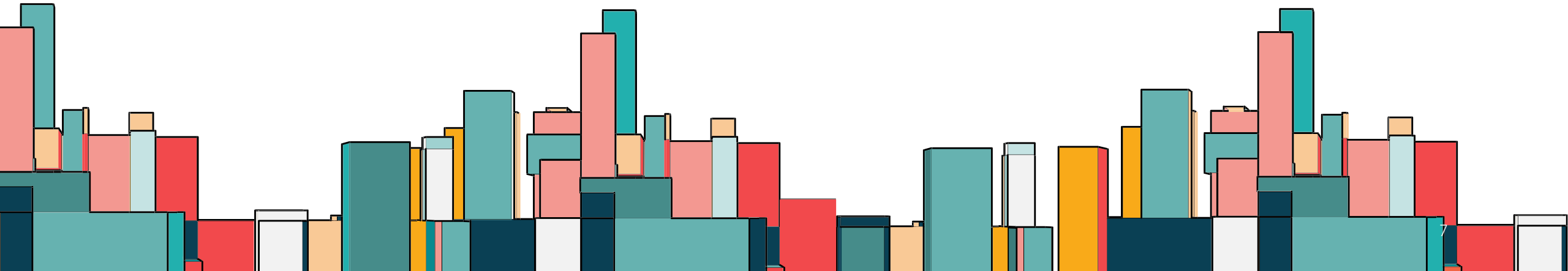
MODEL TRAINING AND TESTING

Logistic Regression

Train Accuracy : 0.9521280362580554 Test Accuracy : 0.8512853197365625

Naïve bayes

Train Accuracy : 0.9509949720274768 Test Accuracy : 0.8502230720203952

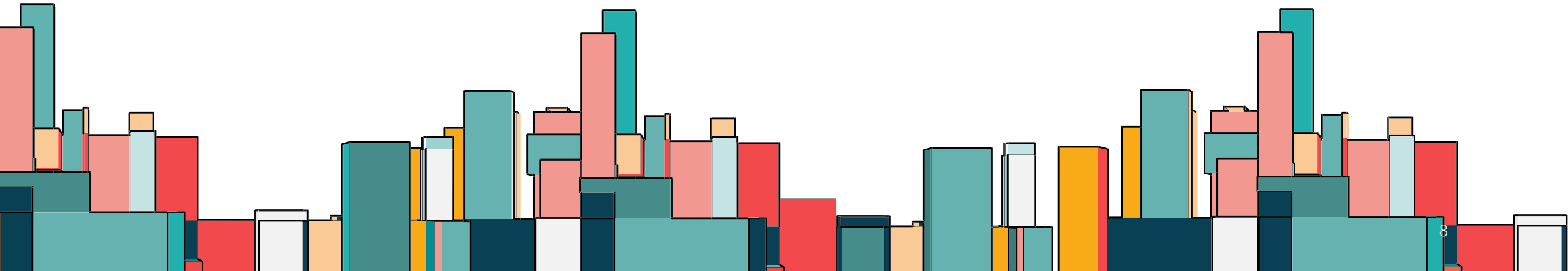


VISUALIZING RESULTS

Confusion Matrix
Bar plot

SAVING THE MODELS

Saving logistic regression and naïve bayes models



THANK YOU