prepare_dataset

March 30, 2025

[317]:
```
# main_df = None

main_train_df = None
main_test_df = None
```

[318]:
```
import os
import kagglehub
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split

def prepare_dataset(dataset, label, text):
  dataset = dataset.copy()

  encoder = LabelEncoder()
  dataset[label] = encoder.fit_transform(dataset[label])

  dataset = dataset.rename(columns={label: 'label', text: 'text'})
  dataset = dataset.drop_duplicates(keep = 'first')
  dataset = dataset.dropna()

  dataset = dataset[['label', 'text']]

  print(dataset.head())
  print(dataset.isnull().sum())
  print(dataset['label'].unique())
  print(dataset.shape)

  return dataset

# def concat_with(dataset):
#   global main_df

#   if main_df is None:
#     main_df = dataset
#   else:
#     dataset = dataset[['label', 'text']]
#     main_df = pd.concat([main_df, dataset], ignore_index=True)
#    main_df = main_df.drop_duplicates(subset=['label', 'text'], keep='first')
#    main_df = main_df.dropna()

#    print(main_df.head())
#    print(main_df.isnull().sum())
#    print(main_df['label'].unique())
#    print(main_df.shape)

#    return main_df

main_train_df = None
main_test_df = None

def concat_with(dataset):
    global main_train_df, main_test_df

    dataset = dataset[['label', 'text']].drop_duplicates(subset=['label', 'text'], keep='first').dropna()

    if main_train_df is not None:
        dataset = dataset[~dataset.apply(lambda row: ((row['label'], row['text']) in zip(main_train_df['label'], main_train_df['text'])), axis=1)]

    if main_test_df is not None:
        dataset = dataset[~dataset.apply(lambda row: ((row['label'], row['text']) in zip(main_test_df['label'], main_test_df['text'])), axis=1)]

    if dataset.empty:
        return main_train_df, main_test_df

    train_df, test_df = train_test_split(dataset, test_size=0.2, random_state=42)

    if main_train_df is None:
        main_train_df = train_df
    else:
        main_train_df = pd.concat([main_train_df, train_df], ignore_index=True)

    if main_test_df is None:
        main_test_df = test_df
    else:
        main_test_df = pd.concat([main_test_df, test_df], ignore_index=True)

    print("Train shape:", main_train_df.shape)
    print("Test shape:", main_test_df.shape)
    return main_train_df, main_test_df
```

```
[319]: # Download latest version
path = kagglehub.dataset_download("abdallahwagih/spam-emails")

files = os.listdir(path)
print("Files in dataset directory:", files)

csv_file_path = os.path.join(path, "spam.csv")
df = pd.read_csv(csv_file_path)

print(df.head())
```

```
Files in dataset directory: ['spam.csv']
  Category                                            Message
0      ham  Go until jurong point, crazy.. Available only …
1      ham                      Ok lar… Joking wif u oni…
2     spam  Free entry in 2 a wkly comp to win FA Cup fina…
3      ham  U dun say so early hor… U c already then say…
4      ham  Nah I don't think he goes to usf, he lives aro…
```

```
[320]: df = prepare_dataset(df, "Category", "Message")
```

```
   label                                               text
0      0  Go until jurong point, crazy.. Available only …
1      0                      Ok lar… Joking wif u oni…
2      1  Free entry in 2 a wkly comp to win FA Cup fina…
3      0  U dun say so early hor… U c already then say…
4      0  Nah I don't think he goes to usf, he lives aro…
label    0
text     0
dtype: int64
[0 1]
(5157, 2)
```

```
[321]: concat_with(df)
```

```
Train shape: (4125, 2)
Test shape: (1032, 2)
```

```
[321]: (      label                                               text
        2598      0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
        5418      0  So how are you really. What are you up to. How…
        99        0                     I see a cup of coffee animation
        2321      0      This pain couldn't have come at a worse time.
        2388      0                              Also where's the piece
        …       …                                                …
        4750      0  Thanx u darlin!im cool thanx. A few bday drink…
        474       1  Want 2 get laid tonight? Want real Dogging loc…
        3273      0  MOON has come to color your dreams, STARS to m…
        4022      0              We have to pick rayan macleran there.
        882       0  see, i knew giving you a break a few times wou…

        [4125 rows x 2 columns],
              label                                               text
        3031      0  Also sir, i sent you an email about how to log…
        495       0                    Are you free now?can i call now?
        2942      0  My supervisor find 4 me one lor i thk his stud…
        3911      0  She.s good. She was wondering if you wont say …
        3360      1  Sorry I missed your call let's talk when you h…
        …       …                                                  …
        2513      0  Hiya , have u been paying money into my accoun…
        1943      0     K…k:)why cant you come here and search job:)
        3038      0     Cos darren say ü considering mah so i ask ü…
        3340      0  Babe !!!! I LOVE YOU !!!! *covers your face in…
        5058      0  Hey next sun 1030 there's a basic yoga course…

        [1032 rows x 2 columns])
```

[322]:

```
# Download latest version
path = kagglehub.dataset_download("uciml/sms-spam-collection-dataset")

files = os.listdir(path)
print("Files in dataset directory:", files)

csv_file_path = os.path.join(path, "spam.csv")
df = pd.read_csv(csv_file_path, encoding='latin1')

df.drop(columns = ['Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4'], inplace = True)

print(df.head())
```

```
Files in dataset directory: ['spam.csv']
     v1                                                 v2
0   ham  Go until jurong point, crazy.. Available only …
1   ham                      Ok lar… Joking wif u oni…
2  spam  Free entry in 2 a wkly comp to win FA Cup fina…
3   ham  U dun say so early hor… U c already then say…
4   ham  Nah I don't think he goes to usf, he lives aro…
```

[323]:

```
df = prepare_dataset(df, "v1", "v2")
```

```
   label                                               text
0      0  Go until jurong point, crazy.. Available only …
1      0                      Ok lar… Joking wif u oni…
2      1  Free entry in 2 a wkly comp to win FA Cup fina…
3      0  U dun say so early hor… U c already then say…
```

```
4      0  Nah I don't think he goes to usf, he lives aro…
label    0
text     0
dtype: int64
[0 1]
(5169, 2)
```

[324]:
```
concat_with(df)
```

```
Train shape: (4694, 2)
Test shape: (1175, 2)
```

[324]:
```
(      label                                               text
 0         0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
 1         0  So how are you really. What are you up to. How…
 2         0                    I see a cup of coffee animation
 3         0      This pain couldn't have come at a worse time.
 4         0                             Also where's the piece
 …       …                                                  …
 4689      1  our mobile number has won å£5000, to claim cal…
 4690      1  5 Free Top Polyphonic Tones call 087018728737,…
 4691      0  Beautiful Truth against Gravity.. Read careful…
 4692      0  'An Amazing Quote'' - \Sometimes in life its d…
 4693      0  Wishing you and your family Merry \X\" mas and…

 [4694 rows x 2 columns],
       label                                               text
 0         0  Also sir, i sent you an email about how to log…
 1         0                  Are you free now?can i call now?
 2         0  My supervisor find 4 me one lor i thk his stud…
 3         0  She.s good. She was wondering if you wont say …
 4         1  Sorry I missed your call let's talk when you h…
 …       …                                                  …
 1170      0  Yup i shd haf ard 10 pages if i add figures…
 1171      1  500 New Mobiles from 2004, MUST GO! Txt: NOKIA…
 1172      0  \Petey boy whereare you me and all your friend…
 1173      0                        I havent add Ì_ yet right..
 1174      0  Well done, blimey, exercise, yeah, i kinda rem…

 [1175 rows x 2 columns])
```

[325]:
```
# Download latest version
path = kagglehub.dataset_download("ashfakyeafi/spam-email-classification")

files = os.listdir(path)
print("Files in dataset directory:", files)

csv_file_path = os.path.join(path, "email.csv")
```

```
df = pd.read_csv(csv_file_path)

print(df.head())
```

```
Files in dataset directory: ['email.csv']
  Category                                            Message
0      ham  Go until jurong point, crazy.. Available only …
1      ham                      Ok lar… Joking wif u oni…
2     spam  Free entry in 2 a wkly comp to win FA Cup fina…
3      ham  U dun say so early hor… U c already then say…
4      ham  Nah I don't think he goes to usf, he lives aro…
```

[326]:
```
df["Category"].unique()
```

[326]:
```
array(['ham', 'spam', '{"mode":"full"'], dtype=object)
```

[327]:
```
df = df[df['Category'] != '{"mode":"full"']
df["Category"].unique()
```

[327]:
```
array(['ham', 'spam'], dtype=object)
```

[328]:
```
df = prepare_dataset(df, "Category", "Message")
```

```
   label                                               text
0      0  Go until jurong point, crazy.. Available only …
1      0                      Ok lar… Joking wif u oni…
2      1  Free entry in 2 a wkly comp to win FA Cup fina…
3      0  U dun say so early hor… U c already then say…
4      0  Nah I don't think he goes to usf, he lives aro…
label    0
text     0
dtype: int64
[0 1]
(5157, 2)
```

[329]:
```
concat_with(df)
```

[329]:
```
(      label                                               text
 0         0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
 1         0  So how are you really. What are you up to. How…
 2         0                     I see a cup of coffee animation
 3         0      This pain couldn't have come at a worse time.
 4         0                            Also where's the piece
 …       …                                                 …
 4689      1  our mobile number has won â£5000, to claim cal…
 4690      1  5 Free Top Polyphonic Tones call 087018728737,…
 4691      0  Beautiful Truth against Gravity.. Read careful…
 4692      0  'An Amazing Quote'' - \Sometimes in life its d…
```

```
  4693      0  Wishing you and your family Merry \X\" mas and…

  [4694 rows x 2 columns],
        label                                                     text
  0         0  Also sir, i sent you an email about how to log…
  1         0                    Are you free now?can i call now?
  2         0  My supervisor find 4 me one lor i thk his stud…
  3         0  She.s good. She was wondering if you wont say …
  4         1  Sorry I missed your call let's talk when you h…
  …     …                                                  …
  1170      0  Yup i shd haf ard 10 pages if i add figures…
  1171      1  500 New Mobiles from 2004, MUST GO! Txt: NOKIA…
  1172      0  \Petey boy whereare you me and all your friend…
  1173      0                        I havent add Ì_ yet right..
  1174      0  Well done, blimey, exercise, yeah, i kinda rem…

  [1175 rows x 2 columns])
```

[330]:
```
# Download latest version
path = kagglehub.dataset_download("ozlerhakan/spam-or-not-spam-dataset")

files = os.listdir(path)
print("Files in dataset directory:", files)

csv_file_path = os.path.join(path, "spam_or_not_spam.csv")
df = pd.read_csv(csv_file_path)

print(df.head())
```

```
Files in dataset directory: ['spam_or_not_spam.csv']
                                             email  label
0    date wed NUMBER aug NUMBER NUMBER NUMBER NUMB…      0
1   martin a posted tassos papadopoulos the greek …      0
2   man threatens explosion in moscow thursday aug…      0
3   klez the virus that won t die already the most…      0
4    in adding cream to spaghetti carbonara which …      0
```

[331]:
```
df = prepare_dataset(df, "label", "email")
```

```
   label                                               text
0      0   date wed NUMBER aug NUMBER NUMBER NUMBER NUMB…
1      0  martin a posted tassos papadopoulos the greek …
2      0  man threatens explosion in moscow thursday aug…
3      0  klez the virus that won t die already the most…
4      0   in adding cream to spaghetti carbonara which …
label    0
text     0
dtype: int64
```

```
[0 1]
(2872, 2)
```

```
[332]: df[df["label"] == 1]["text"].iloc[0]
```

```
[332]: ' save up to NUMBER on life insurance why spend more than you have to life quote
       savings ensuring your family s financial security is very important life quote
       savings makes buying life insurance simple and affordable we provide free access
       to the very best companies and the lowest rates life quote savings is fast easy
       and saves you money let us help you get started with the best values in the
       country on new coverage you can save hundreds or even thousands of dollars by
       requesting a free quote from lifequote savings our service will take you less
       than NUMBER minutes to complete shop and compare save up to NUMBER on all types
       of life insurance hyperlink click here for your free quote protecting your
       family is the best investment you ll ever make if you are in receipt of this
       email in error and or wish to be removed from our list hyperlink please click
       here and type remove if you reside in any state which prohibits e mail
       solicitations for insurance please disregard this email '
```

```
[333]: concat_with(df)
```

```
Train shape: (6991, 2)
Test shape: (1750, 2)
```

```
[333]: (      label                                               text
        0         0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
        1         0  So how are you really. What are you up to. How…
        2         0                    I see a cup of coffee animation
        3         0      This pain couldn't have come at a worse time.
        4         0                             Also where's the piece
        …       …                                                …
        6986      0  on NUMBER sep NUMBER at NUMBER NUMBER guido va…
        6987      0  on wed feb NUMBER NUMBER at NUMBER NUMBER NUMB…
        6988      0   it seems that something changed during the la…
        6989      0  on mon NUMBER oct NUMBER jesse keating wrote o…
        6990      0   over on arstechnica www arstechnica com i saw…

        [6991 rows x 2 columns],
              label                                               text
        0         0  Also sir, i sent you an email about how to log…
        1         0                   Are you free now?can i call now?
        2         0  My supervisor find 4 me one lor i thk his stud…
        3         0  She.s good. She was wondering if you wont say …
        4         1  Sorry I missed your call let's talk when you h…
        …       …                                                …
        1745      0  url URL date not supplied ben hammersley posts…
        1746      0  rohit khare wrote why am i so passionate about…
        1747      0  url URL date NUMBER NUMBER NUMBERtNUMBER NUMBE…
```

```
  1748      0  hello again i tried all the suggestions for th…
  1749      1   write down hello it is time to refinance your…

  [1750 rows x 2 columns])
```

[334]:

```
# Download latest version
path = kagglehub.dataset_download("venky73/spam-mails-dataset")

files = os.listdir(path)
print("Files in dataset directory:", files)

csv_file_path = os.path.join(path, "spam_ham_dataset.csv")
df = pd.read_csv(csv_file_path)

print(df.head())
```

```
Files in dataset directory: ['spam_ham_dataset.csv']
   Unnamed: 0 label                                               text  \
0         605   ham  Subject: enron methanol ; meter # : 988291\r\n…
1        2349   ham  Subject: hpl nom for january 9 , 2001\r\n( see…
2        3624   ham  Subject: neon retreat\r\nho ho ho , we ' re ar…
3        4685  spam  Subject: photoshop , windows , office . cheap …
4        2030   ham  Subject: re : indian springs\r\nthis deal is t…

   label_num
0          0
1          0
2          0
3          1
4          0
```

[335]:

```
df = prepare_dataset(df, "label", "text")
```

```
   label                                               text
0      0  Subject: enron methanol ; meter # : 988291\r\n…
1      0  Subject: hpl nom for january 9 , 2001\r\n( see…
2      0  Subject: neon retreat\r\nho ho ho , we ' re ar…
3      1  Subject: photoshop , windows , office . cheap …
4      0  Subject: re : indian springs\r\nthis deal is t…
label    0
text     0
dtype: int64
[0 1]
(5171, 2)
```

[336]:

```
df['text'] = df['text'].str.replace(r'Subject:\s*', '', regex=True)
print(df.head())
```

```
   label                                               text
```

```
0       0  enron methanol ; meter # : 988291\r\nthis is a…
1       0  hpl nom for january 9 , 2001\r\n( see attached…
2       0  neon retreat\r\nho ho ho , we ' re around to t…
3       1  photoshop , windows , office . cheap . main tr…
4       0  re : indian springs\r\nthis deal is to book th…
```

```
[337]: df[df["label"] == 1]["text"].iloc[0]
```

```
[337]: 'photoshop , windows , office . cheap . main trending\r\nabasements darer
       prudently fortuitous undergone\r\nlighthearted charm orinoco taster\r\nrailroad
       affluent pornographic cuvier\r\nirvin parkhouse blameworthy chlorophyll\r\nrobed
       diagrammatic fogarty clears bayda\r\ninconveniencing managing represented
       smartness hashish\r\nacademies shareholders unload badness\r\ndanielson pure
       caffein\r\nspaniard chargeable levin\r\n'
```

```
[338]: concat_with(df)
```

```
Train shape: (10985, 2)
Test shape: (2749, 2)
```

```
[338]: (       label                                               text
        0          0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
        1          0  So how are you really. What are you up to. How…
        2          0                         I see a cup of coffee animation
        3          0      This pain couldn't have come at a worse time.
        4          0                                  Also where's the piece
        …        …                                                  …
        10980      1  get your hand clock repliacs todday carson\r\n…
        10981      1  a huge load inside her pussy .\r\nembattle sla…
        10982      1                            best software prices .\r\n
        10983      0  natural gas nomination for december 2000 - - r…
        10984      0  defs may 2001\r\ndaren :\r\nplease enter a dem…

        [10985 rows x 2 columns],
              label                                               text
        0         0  Also sir, i sent you an email about how to log…
        1         0                         Are you free now?can i call now?
        2         0  My supervisor find 4 me one lor i thk his stud…
        3         0  She.s good. She was wondering if you wont say …
        4         1  Sorry I missed your call let's talk when you h…
        …       …                                                 …
        2744      0  re : copano line gain contract needed at meter…
        2745      1  discount meds right from home\r\nvalium , xana…
        2746      0  mobil beaumont\r\nbrian nichols of exxonmobil …
        2747      0  re : occidental battleground meter 98 - 1485 o…
        2748      1     growth is good\r\nclick here to be removed\r\n

        [2749 rows x 2 columns])
```

[339]:

```
# Download latest version
path = kagglehub.dataset_download("mfaisalqureshi/spam-email")

files = os.listdir(path)
print("Files in dataset directory:", files)

csv_file_path = os.path.join(path, "spam.csv")
df = pd.read_csv(csv_file_path)

print(df.head())
```

```
Files in dataset directory: ['spam.csv']
  Category                                            Message
0      ham  Go until jurong point, crazy.. Available only …
1      ham                      Ok lar… Joking wif u oni…
2     spam  Free entry in 2 a wkly comp to win FA Cup fina…
3      ham  U dun say so early hor… U c already then say…
4      ham  Nah I don't think he goes to usf, he lives aro…
```

[340]:

```
df = prepare_dataset(df, "Category", "Message")
```

```
   label                                               text
0      0  Go until jurong point, crazy.. Available only …
1      0                      Ok lar… Joking wif u oni…
2      1  Free entry in 2 a wkly comp to win FA Cup fina…
3      0  U dun say so early hor… U c already then say…
4      0  Nah I don't think he goes to usf, he lives aro…
label    0
text     0
dtype: int64
[0 1]
(5157, 2)
```

[341]:

```
concat_with(df)
```

[341]:

```
(       label                                                text
 0          0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
 1          0  So how are you really. What are you up to. How…
 2          0                    I see a cup of coffee animation
 3          0      This pain couldn't have come at a worse time.
 4          0                             Also where's the piece
 …      …                                                    …
 10980      1  get your hand clock repliacs todday carson\r\n…
 10981      1  a huge load inside her pussy .\r\nembattle sla…
 10982      1                       best software prices .\r\n
 10983      0  natural gas nomination for december 2000 - - r…
 10984      0  defs may 2001\r\ndaren :\r\nplease enter a dem…
 [10985 rows x 2 columns],
       label                                               text
 0         0  Also sir, i sent you an email about how to log…
 1         0                   Are you free now?can i call now?
 2         0  My supervisor find 4 me one lor i thk his stud…
 3         0  She.s good. She was wondering if you wont say …
 4         1  Sorry I missed your call let's talk when you h…
 …       …                                                  …
 2744      0  re : copano line gain contract needed at meter…
 2745      1  discount meds right from home\r\nvalium , xana…
 2746      0  mobil beaumont\r\nbrian nichols of exxonmobil …
 2747      0  re : occidental battleground meter 98 - 1485 o…
 2748      1     growth is good\r\nclick here to be removed\r\n

 [2749 rows x 2 columns])
```

[342]:

```
# Download latest version
path = kagglehub.dataset_download("purusinghvi/
 ↪email-spam-classification-dataset")

files = os.listdir(path)
print("Files in dataset directory:", files)

csv_file_path = os.path.join(path, "combined_data.csv")
df = pd.read_csv(csv_file_path)

print(df.head())
```

```
Files in dataset directory: ['combined_data.csv']
   label                                               text
0      1  ounce feather bowl hummingbird opec moment ala…
1      1  wulvob get your medircations online qnb ikud v…
2      0   computer connection from cnn com wednesday es…
3      1  university degree obtain a prosperous future m…
4      0  thanks for all your answers guys i know i shou…
```

[343]:

```
df = prepare_dataset(df, "label", "text")
```

```
   label                                               text
0      1  ounce feather bowl hummingbird opec moment ala…
1      1  wulvob get your medircations online qnb ikud v…
2      0   computer connection from cnn com wednesday es…
3      1  university degree obtain a prosperous future m…
4      0  thanks for all your answers guys i know i shou…
label    0
text     0
dtype: int64
[1 0]
```

```
(83448, 2)
```

[344]:

```
concat_with(df)
```

```
Train shape: (77743, 2)
Test shape: (19439, 2)
```

[344]:

```
(       label                                               text
 0          0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
 1          0  So how are you really. What are you up to. How…
 2          0                      I see a cup of coffee animation
 3          0      This pain couldn't have come at a worse time.
 4          0                               Also where's the piece
 …        …                                                  …
 77738      0  anita . from our conversation today with daren…
 77739      0  business highlights\nenron freight markets\nen…
 77740      0  i am rebuilding r in a mandriva linux environm…
 77741      0  alternative medicine database over escapenumbe…
 77742      0   escapenumberfxml version escapenumberd escape…

 [77743 rows x 2 columns],
        label                                               text
 0          0  Also sir, i sent you an email about how to log…
 1          0                     Are you free now?can i call now?
 2          0  My supervisor find 4 me one lor i thk his stud…
 3          0  She.s good. She was wondering if you wont say …
 4          1  Sorry I missed your call let's talk when you h…
 …        …                                                  …
 19434      1  ci - ialis softabs is better than pfizer viiag…
 19435      1  atasket autumn yorkbe clown begin beguine mir …
 19436      1  curve energy loses paper glasses goes digital …
 19437      0  alert name bush wife of evangelist billy graha…
 19438      1  jovilabe hydromorphous amygdaloid offlet diver…

 [19439 rows x 2 columns])
```

[345]:

```
# Download latest version
path = kagglehub.dataset_download("team-ai/spam-text-message-classification")

files = os.listdir(path)
print("Files in dataset directory:", files)

csv_file_path = os.path.join(path, "SPAM text message 20170820 - Data.csv")
df = pd.read_csv(csv_file_path)

print(df.head())
```

```
Files in dataset directory: ['SPAM text message 20170820 - Data.csv']
  Category                                            Message
```

```
0      ham  Go until jurong point, crazy.. Available only …
1      ham                      Ok lar… Joking wif u oni…
2     spam  Free entry in 2 a wkly comp to win FA Cup fina…
3      ham  U dun say so early hor… U c already then say…
4      ham  Nah I don't think he goes to usf, he lives aro…
```

[346]:
```
df = prepare_dataset(df, "Category", "Message")
```

```
   label                                               text
0      0  Go until jurong point, crazy.. Available only …
1      0                      Ok lar… Joking wif u oni…
2      1  Free entry in 2 a wkly comp to win FA Cup fina…
3      0  U dun say so early hor… U c already then say…
4      0  Nah I don't think he goes to usf, he lives aro…
label    0
text     0
dtype: int64
[0 1]
(5157, 2)
```

[347]:
```
concat_with(df)
```

[347]:
```
(       label                                               text
 0          0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
 1          0  So how are you really. What are you up to. How…
 2          0                      I see a cup of coffee animation
 3          0      This pain couldn't have come at a worse time.
 4          0                                Also where's the piece
 …        …                                                  …
 77738      0  anita . from our conversation today with daren…
 77739      0  business highlights\nenron freight markets\nen…
 77740      0  i am rebuilding r in a mandriva linux environm…
 77741      0  alternative medicine database over escapenumbe…
 77742      0   escapenumberfxml version escapenumberd escape…

 [77743 rows x 2 columns],
        label                                               text
 0          0  Also sir, i sent you an email about how to log…
 1          0                      Are you free now?can i call now?
 2          0  My supervisor find 4 me one lor i thk his stud…
 3          0  She.s good. She was wondering if you wont say …
 4          1  Sorry I missed your call let's talk when you h…
 …        …                                                  …
 19434      1  ci - ialis softabs is better than pfizer viiag…
 19435      1  atasket autumn yorkbe clown begin beguine mir …
 19436      1  curve energy loses paper glasses goes digital …
 19437      0  alert name bush wife of evangelist billy graha…
 19438      1  jovilabe hydromorphous amygdaloid offlet diver…
 [19439 rows x 2 columns])
```

[348]:

```
# Download latest version
path = kagglehub.dataset_download("jackksoncsie/spam-email-dataset")

files = os.listdir(path)
print("Files in dataset directory:", files)

csv_file_path = os.path.join(path, "emails.csv")
df = pd.read_csv(csv_file_path)

print(df.head())
```

```
Files in dataset directory: ['emails.csv']
                                                text  spam
0  Subject: naturally irresistible your corporate…     1
1  Subject: the stock trading gunslinger  fanny i…     1
2  Subject: unbelievable new homes made easy  im …     1
3  Subject: 4 color printing special  request add…     1
4  Subject: do not have money , get software cds …     1
```

[349]:

```
df = prepare_dataset(df, "spam", "text")
```

```
   label                                               text
0      1  Subject: naturally irresistible your corporate…
1      1  Subject: the stock trading gunslinger  fanny i…
2      1  Subject: unbelievable new homes made easy  im …
3      1  Subject: 4 color printing special  request add…
4      1  Subject: do not have money , get software cds …
label    0
text     0
dtype: int64
[1 0]
(5695, 2)
```

[350]:

```
df['text'] = df['text'].str.replace(r'Subject:\s*', '', regex=True)
```

[351]:

```
concat_with(df)
```

```
Train shape: (82299, 2)
Test shape: (20578, 2)
```

[351]:

```
(       label                                               text
 0          0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
 1          0  So how are you really. What are you up to. How…
 2          0                        I see a cup of coffee animation
 3          0      This pain couldn't have come at a worse time.
 4          0                                Also where's the piece
 …      …                                                 …
 82294      0  thomas knudsen  hi vince  i met with thomas th…
 82295      0  sevil yamin  vince ,  do you want me to do thi…
 82296      0  re : a request  zimin ,  i also enjoyed our ta…
 82297      0  6 / 30 aga forecast at 66  mike ,  my number f…
 82298      1  perfect visual solution for your business now …

 [82299 rows x 2 columns],
        label                                               text
 0          0  Also sir, i sent you an email about how to log…
 1          0                    Are you free now?can i call now?
 2          0  My supervisor find 4 me one lor i thk his stud…
 3          0  She.s good. She was wondering if you wont say …
 4          1  Sorry I missed your call let's talk when you h…
 …      …                                                 …
 20573      0  enroncredit . com report for 12 . 10  fyi `  -…
 20574      1  save your money by getting an oem software !  …
 20575      0  re : 12 / 17 churn - - eb 29 to ebl 9  job don…
 20576      1  you may want to look into funding from grants …
 20577      1  you want to submit your website to search engi…

 [20578 rows x 2 columns])
```

[352]:

```python
# Download latest version
path = kagglehub.dataset_download("nitishabharathi/email-spam-dataset")

files = os.listdir(path)
print("Files in dataset directory:", files)

csv_file_path = os.path.join(path, "completeSpamAssassin.csv")
df = pd.read_csv(csv_file_path)

print(df.head())
```

```
Files in dataset directory: ['enronSpamSubset.csv', 'lingSpam.csv',
'completeSpamAssassin.csv']
   Unnamed: 0                                                Body  Label
0           0  \nSave up to 70% on Life Insurance.\nWhy Spend…       1
1           1  1) Fight The Risk of Cancer!\nhttp://www.adcli…       1
2           2  1) Fight The Risk of Cancer!\nhttp://www.adcli…       1
3           3  ##############################################…       1
4           4  I thought you might like these:\n1) Slim Down …       1
```

[353]:

```python
df = prepare_dataset(df, "Label", "Body")
```

```
   label                                                text
0      1  \nSave up to 70% on Life Insurance.\nWhy Spend…
1      1  1) Fight The Risk of Cancer!\nhttp://www.adcli…
```

```
2      1  1) Fight The Risk of Cancer!\nhttp://www.adcli…
3      1  ##############################################…
4      1  I thought you might like these:\n1) Slim Down …
label    0
text     0
dtype: int64
[1 0]
(6045, 2)
```

[354]:
```python
concat_with(df)
```

```
Train shape: (86533, 2)
Test shape: (21637, 2)
```

[354]:
```
(       label                                                text
 0          0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
 1          0  So how are you really. What are you up to. How…
 2          0                    I see a cup of coffee animation
 3          0      This pain couldn't have come at a worse time.
 4          0                             Also where's the piece
 …      …                                                 …
 86528      0  \nForwarded-by: Chris Wedgwood \nFrom: Bert01…
 86529      0  On Wed, Aug 21, 2002 at 02:52:11PM +0800, al@m…
 86530      0  \nZDNet AnchorDesk NewsletterTHURSDAY, JULY 18…
 86531      0  SEARCHSECURITY | Security and Industry News\nJ…
 86532      1  \nBODY {font-family="Arial"}\nTT {font-family=…

 [86533 rows x 2 columns],
        label                                                text
 0          0  Also sir, i sent you an email about how to log…
 1          0                   Are you free now?can i call now?
 2          0  My supervisor find 4 me one lor i thk his stud…
 3          0  She.s good. She was wondering if you wont say …
 4          1  Sorry I missed your call let's talk when you h…
 …      …                                                 …
 21632      1  Dear Sirs,\nWe know your esteemed company in b…
 21633      0  use Perl Daily Headline MailerDamian Conway Pu…
 21634      0  \nPolitical mail (the snail kind) doesn't both…
 21635      0  So, a new family moved in down the street, wit…
 21636      1  \nUntitled Document\n Â Â Â Â Â WE CAN HELP YO…

 [21637 rows x 2 columns])
```

[355]:
```python
# Download latest version
path = kagglehub.dataset_download("nitishabharathi/email-spam-dataset")

files = os.listdir(path)
print("Files in dataset directory:", files)
```

```
csv_file_path = os.path.join(path, "lingSpam.csv")
df = pd.read_csv(csv_file_path)

print(df.head())
```

```
Files in dataset directory: ['enronSpamSubset.csv', 'lingSpam.csv',
'completeSpamAssassin.csv']
   Unnamed: 0                                               Body  Label
0           0  Subject: great part-time or summer job !\n \n …      1
1           1  Subject: auto insurance rates too high ?\n \n …      1
2           2  Subject: do want the best and economical hunti…      1
3           3  Subject: email 57 million people for $ 99\n \n…      1
4           4  Subject: do n't miss these !\n \n attention ! …      1
```

[356]:
```
df = prepare_dataset(df, "Label", "Body")
```

```
   label                                               text
0      1  Subject: great part-time or summer job !\n \n …
1      1  Subject: auto insurance rates too high ?\n \n …
2      1  Subject: do want the best and economical hunti…
3      1  Subject: email 57 million people for $ 99\n \n…
4      1  Subject: do n't miss these !\n \n attention ! …
label    0
text     0
dtype: int64
[1 0]
(2605, 2)
```

[357]:
```
df['text'] = df['text'].str.replace(r'Subject:\s*', '', regex=True)
```

[358]:
```
concat_with(df)
```

```
Train shape: (88605, 2)
Test shape: (22156, 2)
```

[358]:
```
(       label                                               text
 0          0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
 1          0  So how are you really. What are you up to. How…
 2          0                       I see a cup of coffee animation
 3          0          This pain couldn't have come at a worse time.
 4          0                                    Also where's the piece
 …        …                                                 …
 88600      0  jireem @ utxvms . cc . utexas . edu\n \n does …
 88601      0  references on non-human language\n \n content …
 88602      0  call for papers : linguistics session of the m…
 88603      0  inquiry re : slang and rock music\n \n i am wo…
 88604      0  are most people bilingual ? - - summary\n \n a…
 [88605 rows x 2 columns],
        label                                               text
 0          0  Also sir, i sent you an email about how to log…
 1          0                  Are you free now?can i call now?
 2          0  My supervisor find 4 me one lor i thk his stud…
 3          0  She.s good. She was wondering if you wont say …
 4          1  Sorry I missed your call let's talk when you h…
 …        …                                                …
 22151      0  sum : quantification\n \n about four weeks ago…
 22152      0  re : comparative linguistics\n \n ) from : amr…
 22153      1  make unlimited income !\n \n make unlimited in…
 22154      0  chechen\n \n who are the chechen ? johanna nic…
 22155      0  re : 6 . 249 dick armey 's slip and correction…

 [22156 rows x 2 columns])
```

[359]:

```
# Download latest version
path = kagglehub.dataset_download("nitishabharathi/email-spam-dataset")

files = os.listdir(path)
print("Files in dataset directory:", files)

csv_file_path = os.path.join(path, "enronSpamSubset.csv")
df = pd.read_csv(csv_file_path)

print(df.head())
```

```
Files in dataset directory: ['enronSpamSubset.csv', 'lingSpam.csv',
'completeSpamAssassin.csv']
   Unnamed: 0.1  Unnamed: 0  \
0          2469        2469
1          5063        5063
2         12564       12564
3          2796        2796
4          1468        1468

                                                Body  Label
0  Subject: stock promo mover : cwtd\n * * * urge…      1
1  Subject: are you listed in major search engine…      1
2  Subject: important information thu , 30 jun 20…      1
3  Subject: = ? utf - 8 ? q ? bask your life with…      1
4  Subject: " bidstogo " is places to go , things…      1
```

[360]:

```
df = prepare_dataset(df, "Label", "Body")
```

```
   label                                               text
0      1  Subject: stock promo mover : cwtd\n * * * urge…
1      1  Subject: are you listed in major search engine…
```

```
2      1  Subject: important information thu , 30 jun 20…
3      1  Subject: = ? utf - 8 ? q ? bask your life with…
4      1  Subject: " bidstogo " is places to go , things…
label    0
text     0
dtype: int64
[1 0]
(10000, 2)
```

[361]:
```
df['text'] = df['text'].str.replace(r'Subject:\s*', '', regex=True)
print(df.head())
```

```
   label                                               text
0      1  stock promo mover : cwtd\n * * * urgent invest…
1      1  are you listed in major search engines ?\n sub…
2      1  important information thu , 30 jun 2005 .\n su…
3      1  = ? utf - 8 ? q ? bask your life with ? =\n = …
4      1  " bidstogo " is places to go , things to do\n …
```

[362]:
```
concat_with(df)
```

```
Train shape: (96349, 2)
Test shape: (24093, 2)
```

[362]:
```
(       label                                               text
 0          0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
 1          0  So how are you really. What are you up to. How…
 2          0                  I see a cup of coffee animation
 3          0    This pain couldn't have come at a worse time.
 4          0                           Also where's the piece
 …        …                                                …
 96344      0  re : enron / stanford program\n nick ,\n i spo…
 96345      0  re : tenaska iv 10 / 00\n i don ' t see anythi…
 96346      0  fw : attorney client privledge - important ! n…
 96347      1  fw : neevr seen prono flash animation\n buenos…
 96348      0  january production estimate\n daren / carlos :…

 [96349 rows x 2 columns],
        label                                               text
 0          0  Also sir, i sent you an email about how to log…
 1          0                 Are you free now?can i call now?
 2          0  My supervisor find 4 me one lor i thk his stud…
 3          0  She.s good. She was wondering if you wont say …
 4          1  Sorry I missed your call let's talk when you h…
 …        …                                                …
 24088      1  all graphics software available , cheap oem ve…
 24089      1  male muscle boosting system\n i ' ve been usin…
 24090      0  japanese power market\n another article i thou…
 24091      1  calling all small stock players\n ames is fasc…
 24092      0  mg memo\n i am sending you an updated version …

 [24093 rows x 2 columns])
```

[363]:

```python
# Download latest version
path = kagglehub.dataset_download("abdmental01/email-spam-dedection")

files = os.listdir(path)
print("Files in dataset directory:", files)

csv_file_path = os.path.join(path, "mail_data.csv")
df = pd.read_csv(csv_file_path)

print(df.head())
```

```
Files in dataset directory: ['mail_data.csv']
  Category                                            Message
0      ham  Go until jurong point, crazy.. Available only …
1      ham                      Ok lar… Joking wif u oni…
2     spam  Free entry in 2 a wkly comp to win FA Cup fina…
3      ham  U dun say so early hor… U c already then say…
4      ham  Nah I don't think he goes to usf, he lives aro…
```

[364]:

```python
df = prepare_dataset(df, "Category", "Message")
```

```
   label                                               text
0      0  Go until jurong point, crazy.. Available only …
1      0                      Ok lar… Joking wif u oni…
2      1  Free entry in 2 a wkly comp to win FA Cup fina…
3      0  U dun say so early hor… U c already then say…
4      0  Nah I don't think he goes to usf, he lives aro…
label    0
text     0
dtype: int64
[0 1]
(5157, 2)
```

[365]:

```python
concat_with(df)
```

[365]:

```
(       label                                               text
 0          0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
 1          0  So how are you really. What are you up to. How…
 2          0                   I see a cup of coffee animation
 3          0      This pain couldn't have come at a worse time.
 4          0                              Also where's the piece
 …        …                                                …
 96344      0  re : enron / stanford program\n nick ,\n i spo…
 96345      0  re : tenaska iv 10 / 00\n i don ' t see anythi…
 96346      0  fw : attorney client privledge - important ! n…
 96347      1  fw : neevr seen prono flash animation\n buenos…
 96348      0  january production estimate\n daren / carlos :…

 [96349 rows x 2 columns],
        label                                               text
 0          0  Also sir, i sent you an email about how to log…
 1          0                 Are you free now?can i call now?
 2          0  My supervisor find 4 me one lor i thk his stud…
 3          0  She.s good. She was wondering if you wont say …
 4          1  Sorry I missed your call let's talk when you h…
 …      …                                                …
 24088      1  all graphics software available , cheap oem ve…
 24089      1  male muscle boosting system\n i ' ve been usin…
 24090      0  japanese power market\n another article i thou…
 24091      1  calling all small stock players\n ames is fasc…
 24092      0  mg memo\n i am sending you an updated version …

 [24093 rows x 2 columns])
```

[366]:

```python
# Download latest version
path = kagglehub.dataset_download("noeyislearning/spam-emails")

files = os.listdir(path)
print("Files in dataset directory:", files)

csv_file_path = os.path.join(path, "emails.csv")
df = pd.read_csv(csv_file_path)

print(df.head())
```

```
Files in dataset directory: ['emails.csv']
                                                text  spam
0  Subject: naturally irresistible your corporate…     1
1  Subject: the stock trading gunslinger  fanny i…     1
2  Subject: unbelievable new homes made easy  im …     1
3  Subject: 4 color printing special  request add…     1
4  Subject: do not have money , get software cds …     1
```

[367]:

```python
df = prepare_dataset(df, "spam", "text")
```

```
   label                                               text
0      1  Subject: naturally irresistible your corporate…
1      1  Subject: the stock trading gunslinger  fanny i…
2      1  Subject: unbelievable new homes made easy  im …
3      1  Subject: 4 color printing special  request add…
4      1  Subject: do not have money , get software cds …
```

```
label    0
text     0
dtype: int64
[1 0]
(5695, 2)
```

[368]:

```
df['text'] = df['text'].str.replace(r'Subject:\s*', '', regex=True)
```

[369]:

```
concat_with(df)
```

[369]:

```
(       label                                               text
 0          0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
 1          0  So how are you really. What are you up to. How…
 2          0                    I see a cup of coffee animation
 3          0      This pain couldn't have come at a worse time.
 4          0                              Also where's the piece
 …        …                                                  …
 96344      0  re : enron / stanford program\n nick ,\n i spo…
 96345      0  re : tenaska iv 10 / 00\n i don ' t see anythi…
 96346      0  fw : attorney client privledge - important ! n…
 96347      1  fw : neevr seen prono flash animation\n buenos…
 96348      0  january production estimate\n daren / carlos :…

 [96349 rows x 2 columns],
        label                                               text
 0          0  Also sir, i sent you an email about how to log…
 1          0                  Are you free now?can i call now?
 2          0  My supervisor find 4 me one lor i thk his stud…
 3          0  She.s good. She was wondering if you wont say …
 4          1  Sorry I missed your call let's talk when you h…
 …        …                                                  …
 24088      1  all graphics software available , cheap oem ve…
 24089      1  male muscle boosting system\n i ' ve been usin…
 24090      0  japanese power market\n another article i thou…
 24091      1  calling all small stock players\n ames is fasc…
 24092      0  mg memo\n i am sending you an updated version …

 [24093 rows x 2 columns])
```

[370]:

```
# Download latest version
path = kagglehub.dataset_download("ahsenwaheed/youtube-comments-spam-dataset")

files = os.listdir(path)
print("Files in dataset directory:", files)

csv_file_path = os.path.join(path, "Youtube-Spam-Dataset.csv")
df = pd.read_csv(csv_file_path)
```

```
print(df.head())
print(df.isnull().sum())
print(df['CLASS'].unique())
print(df.shape)
```

```
Files in dataset directory: ['Youtube-Spam-Dataset.csv']
                                    COMMENT_ID            AUTHOR  \
0  LZQPQhLyRh80UYxNuaDWhIGQYNQ96IuCg-AYWqNPjpU         Julius NM
1  LZQPQhLyRh_C2cTtd9MvFRJedxydaVW-2sNg5Diuo4A       adam riyati
2  LZQPQhLyRh9MSZYnf8djyk0gEF9BHDPYrrK-qCczIY8  Evgeny Murashkin
3          z13jhp0bxqncu512g22wvzkasxmvvzjaz04   ElNino Melendez
4          z13fwbwp1oujthgqj04chlngpvzmtt3r3dw            GsMega

                  DATE                                            CONTENT  \
0  2013-11-07T06:20:48  Huh, anyway check out this you[tube] channel: …
1  2013-11-07T12:37:15  Hey guys check out my new channel and our firs…
2  2013-11-08T17:34:21             just for test I have to say murdev.com
3  2013-11-09T08:28:43   me shaking my sexy ass on my channel enjoy ^_^
4  2013-11-10T16:05:38            watch?v=vtaRGgvGtWQ   Check this out .

                       VIDEO_NAME  CLASS
0  PSY - GANGNAM STYLE(?????) M/V      1
1  PSY - GANGNAM STYLE(?????) M/V      1
2  PSY - GANGNAM STYLE(?????) M/V      1
3  PSY - GANGNAM STYLE(?????) M/V      1
4  PSY - GANGNAM STYLE(?????) M/V      1
COMMENT_ID      0
AUTHOR          0
DATE          245
CONTENT         0
VIDEO_NAME      0
CLASS           0
dtype: int64
[1 0]
(1956, 6)
```

[371]:
```
df = prepare_dataset(df, "CLASS", "CONTENT")
```

```
   label                                               text
0      1  Huh, anyway check out this you[tube] channel: …
1      1  Hey guys check out my new channel and our firs…
2      1             just for test I have to say murdev.com
3      1   me shaking my sexy ass on my channel enjoy ^_^
4      1            watch?v=vtaRGgvGtWQ   Check this out .
label    0
text     0
dtype: int64
[1 0]
```

```
(1710, 2)
```

[372]:

```
concat_with(df)
```

```
Train shape: (97595, 2)
Test shape: (24405, 2)
```

[372]:

```
(       label                                               text
 0          0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
 1          0  So how are you really. What are you up to. How…
 2          0                    I see a cup of coffee animation
 3          0      This pain couldn't have come at a worse time.
 4          0                             Also where's the piece
 …        …                                                  …
 97590      0  Rihanna is so beautiful and amazing    love …
 97591      0                                          waka waka
 97592      0           I hate it when Laura Bennett comes in
 97593      1  Hey Music Fans I really appreciate any of you …
 97594      0                      I could hear this for years ;3

 [97595 rows x 2 columns],
        label                                               text
 0          0  Also sir, i sent you an email about how to log…
 1          0                   Are you free now?can i call now?
 2          0  My supervisor find 4 me one lor i thk his stud…
 3          0  She.s good. She was wondering if you wont say …
 4          1  Sorry I missed your call let's talk when you h…
 …        …                                                  …
 24400      1  I really ask nicely to view my vids:) I subscr…
 24401      1  http://tankionline.com#friend=cd92db3f4 great …
 24402      1        Subscribe me, I will? subscribe you back!!!
 24403      0                                            NICE :3
 24404      0                                      Eminem rocks!

 [24405 rows x 2 columns])
```

[373]:

```
# Download latest version
path = kagglehub.dataset_download("shantanudhakadd/
 ↪email-spam-detection-dataset-classification")

files = os.listdir(path)
print("Files in dataset directory:", files)

csv_file_path = os.path.join(path, "spam.csv")
df = pd.read_csv(csv_file_path, encoding='latin1')

df.drop(columns = ['Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4'], inplace = True)
```

```
print(df.head())
```

```
Files in dataset directory: ['spam.csv']
     v1                                                 v2
0   ham  Go until jurong point, crazy.. Available only …
1   ham                      Ok lar… Joking wif u oni…
2  spam  Free entry in 2 a wkly comp to win FA Cup fina…
3   ham  U dun say so early hor… U c already then say…
4   ham  Nah I don't think he goes to usf, he lives aro…
```

[374]:

```
df = prepare_dataset(df, "v1", "v2")
```

```
   label                                               text
0      0  Go until jurong point, crazy.. Available only …
1      0                      Ok lar… Joking wif u oni…
2      1  Free entry in 2 a wkly comp to win FA Cup fina…
3      0  U dun say so early hor… U c already then say…
4      0  Nah I don't think he goes to usf, he lives aro…
label    0
text     0
dtype: int64
[0 1]
(5169, 2)
```

[375]:

```
concat_with(df)
```

[375]:

```
(       label                                               text
 0          0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
 1          0  So how are you really. What are you up to. How…
 2          0                    I see a cup of coffee animation
 3          0      This pain couldn't have come at a worse time.
 4          0                             Also where's the piece
 …       …                                                …
 97590      0  Rihanna is so beautiful and amazing    love …
 97591      0                                       waka waka
 97592      0             I hate it when Laura Bennett comes in
 97593      1  Hey Music Fans I really appreciate any of you …
 97594      0                    I could hear this for years ;3

 [97595 rows x 2 columns],
        label                                               text
 0          0  Also sir, i sent you an email about how to log…
 1          0                   Are you free now?can i call now?
 2          0  My supervisor find 4 me one lor i thk his stud…
 3          0  She.s good. She was wondering if you wont say …
 4          1  Sorry I missed your call let's talk when you h…
 …       …                                                …
 24400      1  I really ask nicely to view my vids:) I subscr…
```

```
24401      1  http://tankionline.com#friend=cd92db3f4 great …
24402      1        Subscribe me, I will? subscribe you back!!!
24403      0                                          NICE :3
24404      0                                    Eminem rocks!

[24405 rows x 2 columns])
```

[376]:

```python
# Download latest version
path = kagglehub.dataset_download("karthickveerakumar/spam-filter")

files = os.listdir(path)
print("Files in dataset directory:", files)

csv_file_path = os.path.join(path, "emails.csv")
df = pd.read_csv(csv_file_path)

print(df.head())
```

```
Files in dataset directory: ['emails.csv']
                                                text  spam
0  Subject: naturally irresistible your corporate…     1
1  Subject: the stock trading gunslinger  fanny i…     1
2  Subject: unbelievable new homes made easy  im …     1
3  Subject: 4 color printing special  request add…     1
4  Subject: do not have money , get software cds …     1
```

[377]:

```python
df = prepare_dataset(df, "spam", "text")
```

```
   label                                               text
0      1  Subject: naturally irresistible your corporate…
1      1  Subject: the stock trading gunslinger  fanny i…
2      1  Subject: unbelievable new homes made easy  im …
3      1  Subject: 4 color printing special  request add…
4      1  Subject: do not have money , get software cds …
label    0
text     0
dtype: int64
[1 0]
(5695, 2)
```

[378]:

```python
df['text'] = df['text'].str.replace(r'Subject:\s*', '', regex=True)
```

[379]:

```python
df = prepare_dataset(df, "label", "text")
```

```
   label                                               text
0      1  naturally irresistible your corporate identity…
1      1  the stock trading gunslinger  fanny is merrill…
2      1  unbelievable new homes made easy  im wanting t…
3      1  4 color printing special  request additional i…
```

```
4      1  do not have money , get software cds from here…
label    0
text     0
dtype: int64
[1 0]
(5695, 2)
```

[380]:
```
concat_with(df)
```

[380]:
```
(       label                                                text
 0          0  Got fujitsu, ibm, hp, toshiba… Got a lot of …
 1          0  So how are you really. What are you up to. How…
 2          0                     I see a cup of coffee animation
 3          0       This pain couldn't have come at a worse time.
 4          0                              Also where's the piece
 …      …                                                 …
 97590      0  Rihanna is so beautiful and amazing    love …
 97591      0                                          waka waka
 97592      0              I hate it when Laura Bennett comes in
 97593      1  Hey Music Fans I really appreciate any of you …
 97594      0                     I could hear this for years ;3

 [97595 rows x 2 columns],
        label                                                text
 0          0  Also sir, i sent you an email about how to log…
 1          0                    Are you free now?can i call now?
 2          0  My supervisor find 4 me one lor i thk his stud…
 3          0  She.s good. She was wondering if you wont say …
 4          1  Sorry I missed your call let's talk when you h…
 …      …                                                 …
 24400      1  I really ask nicely to view my vids:) I subscr…
 24401      1  http://tankionline.com#friend=cd92db3f4 great …
 24402      1        Subscribe me, I will? subscribe you back!!!
 24403      0                                            NICE :3
 24404      0                                      Eminem rocks!

 [24405 rows x 2 columns])
```

[381]:
```python
import re
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
import nltk

nltk.download('stopwords')
nltk.download('wordnet')

class TextPreprocessor:
    def __init__(self):
        self.stop_words = set(stopwords.words('english'))
        self.lemmatizer = WordNetLemmatizer()

    def clean_text(self, text):
        text = re.sub(r'[^a-zA-Z0-9\s]', '', text)
        return text

    def to_lowercase(self, text):
        return text.lower()

    def remove_stopwords(self, text):
        words = text.split()
        return ' '.join([word for word in words if word not in self.stop_words])

    def lemmatize_text(self, text):
        words = text.split()
        return ' '.join([self.lemmatizer.lemmatize(word) for word in words])

    def preprocess_text(self, text):
        text = self.clean_text(text)
        text = self.to_lowercase(text)
        text = self.remove_stopwords(text)
        text = self.lemmatize_text(text)
        return text

textPreprocessor = TextPreprocessor()
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
```

[382]:
```
main_train_df['text'] = main_train_df['text'].apply(textPreprocessor.
 ↪preprocess_text)
print(main_train_df)
```

```
       label                                               text
0          0       got fujitsu ibm hp toshiba got lot model say
1          0                                 really hows master
2          0                           see cup coffee animation
3          0                       pain couldnt come worse time
4          0                                  also wheres piece
…      …                                                 …
97590      0  rihanna beautiful amazing love much forever ri…
97591      0                                          waka waka
97592      0                            hate laura bennett come
97593      1  hey music fan really appreciate take time read…
97594      0                                  could hear year 3
```

```
[97595 rows x 2 columns]
```

[383]:

```
main_test_df['text'] = main_test_df['text'].apply(textPreprocessor.
 ↪preprocess_text)
print(main_test_df)
```

```
       label                                               text
0          0  also sir sent email log usc payment portal ill…
1          0                                   free nowcan call
2          0  supervisor find 4 one lor thk student havent a…
3          0  shes good wondering wont say hi shes smiling c…
4          1     sorry missed call let talk time im 07090201529
…      …                                                ...
24400      1         really ask nicely view vids subscribe back
24401      1  httptankionlinecomfriendcd92db3f4 great game c…
24402      1                           subscribe subscribe back
24403      0                                             nice 3
24404      0                                        eminem rock

[24405 rows x 2 columns]
```

[384]:

```
from google.colab import files

main_train_df.to_csv('spam_text_train_dataset.csv', index=False)
files.download('spam_text_train_dataset.csv')

main_test_df.to_csv('spam_text_test_dataset.csv', index=False)
files.download('spam_text_test_dataset.csv')
```

```
<IPython.core.display.Javascript object>

<IPython.core.display.Javascript object>

<IPython.core.display.Javascript object>

<IPython.core.display.Javascript object>
```

[384]: