Fake or Phishing website detection !!!

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:
```
data = pd.read_csv(r"C:\Users\ASUS\Downloads\archive (4)\phishing_site_urls.csv")
```

In [3]:
```
data.head()
```

Out[3]:

	URL	Label
0	nobell.it/70ffb52d079109dca5664cce6f317373782/...	bad
1	www.dghjdgf.com/paypal.co.uk/cycgi-bin/webscrc...	bad
2	serviciosbys.com/paypal.cgi.bin.get-into.herf....	bad
3	mail.printakid.com/www.online.americanexpress....	bad
4	thewhiskeydregs.com/wp-content/themes/widescre...	bad

In [4]:
```
data.tail()
```

Out[4]:

	URL	Label
549341	23.227.196.215/	bad
549342	apple-checker.org/	bad
549343	apple-iclods.org/	bad
549344	apple-uptoday.org/	bad
549345	apple-search.info	bad

In [5]:
```
data.sample(10)
```

Out[5]:

	URL	Label
143528	anzmilan-cantonese.blogspot.com/	good
518501	0yuq.padudopacyjpps.com/jhufqjew2h\nwww.doblat...	bad
288172	baseball-almanac.com/teamstats/roster.php?y=19...	good
426043	rpmwin.com/user/padagge/obituary_of_horace.htm	good
374966	linkedin.com/directory/people/kroeger.html	good
278532	americanillustrators.com/artist.php?id=8153	good
426372	rwbjv.org/	good
240792	soulmusicstore.com/	good
321836	encyclopedia.com/topic/Montreal.aspx	good
461781	washingtonprintclub.org/Drawn_to_Washington.htm	good

In [6]:
```
data.shape
```

Out[6]:
```
(549346, 2)
```

In [7]:
```
data.size
```

Out[7]:
```
1098692
```

In [10]:
```
data.describe()
```

Out[10]:

	URL	Label
count	549346	549346
unique	507195	2
top	jhomitevd2abj3fk.tor2web.org/	good
freq	52	392924

In [11]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 549346 entries, 0 to 549345
Data columns (total 2 columns):
 #   Column  Non-Null Count   Dtype 
---  ------  --------------   ----- 
 0   URL     549346 non-null  object
 1   Label   549346 non-null  object
dtypes: object(2)
memory usage: 8.4+ MB
```

In [8]:
```
data.URL
```

Out[8]:
```
0         nobell.it/70ffb52d079109dca5664cce6f317373782/...
1         www.dghjdgf.com/paypal.co.uk/cycgi-bin/webscrc...
2         serviciosbys.com/paypal.cgi.bin.get-into.herf....
3         mail.printakid.com/www.online.americanexpress....
4         thewhiskeydregs.com/wp-content/themes/widescre...
                                ...
549341                                     23.227.196.215/
549342                                  apple-checker.org/
549343                                   apple-iclods.org/
549344                                  apple-uptoday.org/
549345                                   apple-search.info
Name: URL, Length: 549346, dtype: object
```

In [9]:
```
data.Label
```

Out[9]:
```
0         bad
1         bad
2         bad
3         bad
4         bad
         ...
549341    bad
549342    bad
549343    bad
549344    bad
549345    bad
Name: Label, Length: 549346, dtype: object
```

In [12]:
```
data.isnull().sum()
```

Out[12]:
```
URL      0
Label    0
dtype: int64
```

In [13]:
```
data.Label.value_counts()
```

Out[13]:
```
Label
good    392924
bad     156422
Name: count, dtype: int64
```

In [14]:
```
good = data[data.Label == "good"]
bad = data[data.Label == "bad"]
```

In [15]:
```
good.shape
```

Out[15]:
```
(392924, 2)
```

In [16]:
```
good.head()
```

Out[16]:

	URL	Label
18231	esxcc.com/js/index.htm?us.battle.net/noghn/en/...	good
18232	wwweira¯&nvinip¿ncH¯wVö%ÆåyDaHðû/ÏyEùuË\nÓ6...	good
18233	'www.institutocgr.coo/web/media/syqvem/dk-óij...	good
18234	YìêkoãÕ»Î§DéÎl½ñ¡ââqtò¸/à; Í	good
18236	ruta89fm.com/images/AS@Vies/1i75cf7b16vc<Fd16...	good

In [17]:
```
good.sample(10)
```

Out[17]:

	URL	Label
271901	allkpop.com/2009/06/big_bang_gara_gara_go	good
350024	havingxxx.com/escorts/montreal/incalls-gfe-pfe...	good
225480	paulboylan.wordpress.com/category/morbidly-obe...	good
221455	newsroom.lds.org/leader-biographies/elder-davi...	good
248109	timmusgrove.com/	good
474223	youtube.com/watch?v=RT8pnPsa7xw	good
345644	gifts.com/ssp?slkw=winter%20rose%20poinsettia&...	good
50812	www.cens.com/signtek	good
244581	supermarchbyblos.foodpages.ca/	good
53590	www.peak.org/~moco/	good

In [18]:
```
bad.shape
```

Out[18]:
```
(156422, 2)
```

In [19]:
```
bad.head()
```

Out[19]:

	URL	Label
0	nobell.it/70ffb52d079109dca5664cce6f317373782/...	bad
1	www.dghjdgf.com/paypal.co.uk/cycgi-bin/webscrc...	bad
2	serviciosbys.com/paypal.cgi.bin.get-into.herf....	bad
3	mail.printakid.com/www.online.americanexpress....	bad
4	thewhiskeydregs.com/wp-content/themes/widescre...	bad

In [20]:
```
bad.sample(10)
```

Out[20]:

	URL	Label
500003	www.mbeccarini.com/xkzd7c	bad
98331	103.234.36.75/rd927.exe	bad
498686	bgnakano.web.fc2.com/581n98q	bad
484231	available.pearlstorehouse.net/upvw4wok53\nkooo...	bad
507725	95.213.139.104/upd/61	bad
8123	www.explosaodepremios.net/cadastro/promocao/in...	bad
125807	hallmarkteam.com/uuu/Ed/Ed/	bad
514879	123.249.34.199:9655/17733	bad
507942	dinttobogo.com/zapoy/gate.php/	bad
41502	'9d345009-a-62cb3a1a-s-sites.googlegroups.com/...	bad

In [22]:
```
good = good.sample(n=156422)
good.shape
```

Out[22]:
```
(156422, 2)
```

In [23]:
```
bad.shape
```

Out[23]:
```
(156422, 2)
```

In [24]:
```
data = pd.concat([good, bad], axis=0)
```

In [25]:
```
data.head()
```

Out[25]:

	URL	Label
181925	en.wikipedia.org/wiki/Vernon_Vanoy	good
418044	publicbackgroundchecks.com/searchresponse.aspx...	good
204394	kansascity.citysearch.com/profile/5832684/kans...	good
181371	en.wikipedia.org/wiki/The_Walton_Experience	good
467238	wunderground.com/US/MI/McMillan/KISQ.html	good

In [26]:
```
data.shape
```

Out[26]:
```
(312844, 2)
```

In [27]:
```
data.Label.value_counts()
```

Out[27]:
```
Label
good    156422
bad     156422
Name: count, dtype: int64
```

In [30]:
```
import time
import nltk
from nltk.corpus import stopwords
import re
from nltk.corpus import stopwords
from nltk.tokenize import RegexpTokenizer
from nltk.stem.snowball import SnowballStemmer
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

In [31]:
```
tokenizer = RegexpTokenizer(r'[A-Za-z]+')
```

In [32]:
```
data.URL[0]
```

Out[32]:
```
'nobell.it/70ffb52d079109dca5664cce6f317373782/login.SkyPe.com/en/cgi-bin/verification/login/70ffb52d079109dca5664cce6f317373/index.php?cmd=_profile-ach&outdated_page_tmpl=p/gen/failed-to-load&nav=0.5.1&login_access=1322408526'
```

In [33]:
```
tokenizer.tokenize(data.URL[0])
```

Out[33]:
```
['nobell',
 'it',
 'ffb',
 'd',
 'dca',
 'cce',
 'f',
 'login',
 'SkyPe',
 'com',
 'en',
 'cgi',
 'bin',
 'verification',
 'login',
 'ffb',
 'd',
 'dca',
 'cce',
 'f',
 'index',
 'php',
 'cmd',
 'profile',
 'ach',
 'outdated',
 'page',
 'tmpl',
 'p',
 'gen',
 'failed',
 'to',
 'load',
 'nav',
 'login',
 'access']
```

In [34]:
```
print('Getting words tokenized ...')
t0= time.perf_counter()
data['text_tokenized'] = data.URL.map(lambda t: tokenizer.tokenize(t)) # doing with all rows
t1 = time.perf_counter() - t0
print('Time taken',t1 ,'sec')
```

```
Getting words tokenized ...
Time taken 0.9497446000004857 sec
```

In [35]:
```
data.sample(5)
```

Out[35]:

	URL	Label	text_tokenized
534189	platforms-root-technologies.com/8fh34f3	bad	[platforms, root, technologies, com, fh, f]
420542	reachingquiet.com/	good	[reachingquiet, com]
413990	pick-rick.com/	good	[pick, rick, com]
523941	militarygradehosting.com/util/cp.php?m=login	bad	[militarygradehosting, com, util, cp, php, m, ...
205046	kdford2006.com/2011/10/31/happy-halloween-ever...	good	[kdford, com, happy, halloween, everyone]

In [36]:
```
stemmer = SnowballStemmer("english")
```

In [37]:
```
print('Getting words stemmed ...')
t0= time.perf_counter()
data['text_stemmed'] = data['text_tokenized'].map(lambda l: [stemmer.stem(word) for word in l])
t1= time.perf_counter() - t0
print('Time taken',t1 ,'sec')
```

```
Getting words stemmed ...
Time taken 16.277618100000836 sec
```

In [38]:
```
data.sample(5)
```

Out[38]:

	URL	Label	text_tokenized	text_stemmed
545847	fajne-rolety.pl/9yZDkdwS4/2cw6IEglu.php	bad	[fajne, rolety, pl, yZDkdwS, cw, IEglu, php]	[fajn, roleti, pl, yzdkdws, cw, ieglu, php]
34406	haryojackson.com/secure/secure-code68/security/	bad	[haryojackson, com, secure, secure, code, secu...	[haryojackson, com, secur, secur, code, secur]
48193	www.jobsin.co.uk/?h=forestry	good	[www, jobsin, co, uk, h, forestry]	[www, jobsin, co, uk, h, forestri]
85823	www.grok2.com/sonyvaio.html	good	[www, grok, com, sonyvaio, html]	[www, grok, com, sonyvaio, html]
131076	royalgateenergy.com/wp-admin/js/images/httpdoc...	bad	[royalgateenergy, com, wp, admin, js, images, ...	[royalgateenergi, com, wp, admin, js, imag, ht...

In [39]:
```
print('Getting joiningwords ...')
t0= time.perf_counter()
data['text_sent'] = data['text_stemmed'].map(lambda l: ' '.join(l))
t1= time.perf_counter() - t0
print('Time taken',t1 ,'sec')
```

```
Getting joiningwords ...
Time taken 0.11203430000023218 sec
```

In [40]:
```
data.sample(5)
```

Out[40]:

	URL	Label	text_tokenized	text_stemmed	text_sent
192980	genforum.genealogy.com/mackenzie/all.html	good	[genforum, genealogy, com, mackenzie, all, html]	[genforum, genealog, com, mackenzi, all, html]	genforum genealog com mackenzi all html
299906	cathedral.vancouver.bc.ca/	good	[cathedral, vancouver, bc, ca]	[cathedr, vancouv, bc, ca]	cathedr vancouv bc ca
451146	trinitywaconia.org/School4Staff.html	good	[trinitywaconia, org, School, Staff, html]	[trinitywaconia, org, school, staff, html]	trinitywaconia org school staff html
530224	aquatixbottle.com/ygyngc	bad	[aquatixbottle, com, ygyngc]	[aquatixbottl, com, ygyngc]	aquatixbottl com ygyngc
512036	speciaaldesign.nl/74t3nf4gv4	bad	[speciaaldesign, nl, t, nf, gv]	[speciaaldesign, nl, t, nf, gv]	speciaaldesign nl t nf gv

In [42]:
```
data = data.drop(['URL', 'text_tokenized', 'text_stemmed'], axis = 1)
data.head()
```

Out[42]:

	Label	text_sent
181925	good	en wikipedia org wiki vernon vanoy
418044	good	publicbackgroundcheck com searchrespons aspx v...
204394	good	kansasc citysearch com profil kansa citi mo st...
181371	good	en wikipedia org wiki the walton experi
467238	good	wunderground com us mi mcmillan kisq html

In [43]:
```
X = data.text_sent.values
```

In [44]:
```
Y = data.Label.values
```

In [45]:
```
X.shape
```

Out[45]: (312844,)

In [46]:
```
Y.shape
```

Out[46]: (312844,)

In [47]:
```
vector = TfidfVectorizer()
vector.fit(X)
X = vector.transform(X)
```

In [48]:
```
X.shape
```

Out[48]: (312844, 245257)

In [49]:
```
print(X)
```

```
  (0, 229443)   0.2752682630400188
  (0, 229405)   0.2606424795193824
  (0, 220861)   0.6031743501680439
  (0, 219700)   0.6368478053856346
  (0, 155535)   0.1937598765051533
  (0, 63356)    0.22216112973354754
  (1, 221868)   0.2010931146393387
  (1, 186947)   0.38203192682241754
  (1, 168851)   0.37773644979310633
  (1, 147666)   0.24479259416832205
  (1, 102637)   0.38868108520751343
  (1, 79599)    0.35048598642080914
  (1, 40544)    0.051994515451006085
  (1, 36705)    0.39252352034498866
  (1, 17269)    0.38203192682241754
  (1, 12675)    0.1853779302477225
  (2, 196841)   0.26226802069195915
  (2, 185817)   0.2528210355190081
  (2, 167558)   0.25734491318082947
  (2, 159267)   0.2850086766294769
  (2, 137116)   0.2899264009432102
  (2, 111584)   0.3167338567139021
  (2, 111577)   0.25263314049275604
  (2, 95088)    0.11498934358996307
  (2, 64308)    0.4353089015074692
  :     :
  (312837, 45582)       0.11238764606703207
  (312837, 33739)       0.1629748029465446
  (312837, 26051)       0.07957046606200853
  (312837, 22429)       0.14327730711879016
  (312837, 18533)       0.1945360704443338
  (312837, 8249)        0.39820003490665684
  (312838, 137590)      0.5255986348663262
  (312838, 131962)      0.29460096745367464
  (312838, 103441)      0.5379246497187811
  (312838, 92707)       0.3926758102020993
  (312838, 54667)       0.2448195878761175
  (312838, 46335)       0.36018196029208166
  (312838, 40544)       0.06108558898288707
  (312840, 155535)      0.2711535459166016
  (312840, 35658)       0.7731869522682461
  (312840, 10246)       0.5732867444647524
  (312841, 155535)      0.23640406299440517
  (312841, 97106)       0.8332439508762426
  (312841, 10246)       0.4998176040596082
  (312842, 217286)      0.8332439508762426
  (312842, 155535)      0.23640406299440517
  (312842, 10246)       0.4998176040596082
  (312843, 186863)      0.5612225704339859
  (312843, 100285)      0.45137873858161687
  (312843, 10246)       0.6937481248925563
```

In [50]:
```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, stratify = Y,random_state = 2)
```

In [51]:
```
print(X_train)
```

```
  (0, 233166)   0.19154839625880943
  (0, 225604)   0.7757394620377778
  (0, 99795)    0.22209277498275254
  (0, 95088)    0.17796219474970676
  (0, 40544)    0.08533538909472256
  (0, 15798)    0.5227424096401568
  (1, 233166)   0.10548461508962889
  (1, 231324)   0.4271953206662644
  (1, 149967)   0.8543906413325288
  (1, 100285)   0.16151062973777905
  (1, 99795)    0.12230523116251762
  (1, 12608)    0.18801471457207766
  (2, 135574)   0.6907115694690159
  (2, 114659)   0.6907115694690159
  (2, 95081)    0.20015446520773353
  (2, 40544)    0.07598187718597012
  (3, 229443)   0.4229800694454751
  (3, 229405)   0.4005059387159407
  (3, 155535)   0.2977334368836465
  (3, 99598)    0.41669828758431743
  (3, 71792)    0.5309079341323443
  (3, 63356)    0.3413750973141416
  (4, 201555)   0.47035403171808676
  (4, 182166)   0.21133717108513658
  (4, 168809)   0.38509436990663876
  :     :
  (250271, 185104)      0.47645706271382915
  (250271, 155329)      0.2864803578706816
  (250271, 152241)      0.19589561301188577
  (250271, 95088)       0.10930389484352732
  (250271, 40544)       0.05241276333526681
  (250271, 30254)       0.3244369022682307
  (250271, 23550)       0.20075751923923207
  (250272, 201339)      0.5082337144103014
  (250272, 177419)      0.36131942186564264
  (250272, 162398)      0.13991836658093124
  (250272, 158199)      0.453946461503629
  (250272, 66173)       0.5657151908080638
  (250272, 40544)       0.06696610846162637
  (250272, 3038)        0.24694957158809983
  (250273, 237161)      0.291898486226686
  (250273, 219275)      0.47320838195368653
  (250273, 211067)      0.48620036317439613
  (250273, 211038)      0.3946931633126311
  (250273, 201963)      0.5415380324591459
  (250273, 40544)       0.07370328399043179
  (250274, 162398)      0.1493042347893807
  (250274, 83542)       0.6495897643257039
  (250274, 66321)       0.3587035518969018
  (250274, 40544)       0.07145826402213316
  (250274, 13366)       0.6495897643257039
```

In [52]:
```
Y_train
```

Out[52]:
```
array(['good', 'good', 'good', ..., 'bad', 'good', 'good'], dtype=object)
```

In [53]:
```
print(Y_train)
```

```
['good' 'good' 'good' ... 'bad' 'good' 'good']
```

In [54]:
```
X_test
```

Out[54]:
```
<62569x245257 sparse matrix of type '<class 'numpy.float64'>'
        with 443132 stored elements in Compressed Sparse Row format>
```

In [58]:
```
print(X_test)
```

```
  (0, 197011)   0.6352132508838236
  (0, 191195)   0.6635173919825793
  (0, 42958)    0.3447458971113051
  (0, 40544)    0.045665788022767474
  (0, 23461)    0.18791939406943026
  (1, 227387)   0.11638371304640055
  (1, 187207)   0.1169845749923703
  (1, 182431)   0.13470707819551564
  (1, 162398)   0.06957159295692066
  (1, 159556)   0.10371632125561032
  (1, 141876)   0.26756262853207874
  (1, 124407)   0.08760242153415994
  (1, 70752)    0.14311647441069916
  (1, 70027)    0.29861668225426596
  (1, 67747)    0.455026370794394
  (1, 53410)    0.21352724512307986
  (1, 50054)    0.25011129419130584
  (1, 50041)    0.2162460310107553
  (1, 48799)    0.11717037700668372
  (1, 40544)    0.06659510046676335
  (1, 39140)    0.11245994499901243
  (1, 34439)    0.16408006024922442
  (1, 34385)    0.3110263509392469
  (1, 32528)    0.1314112361700905
  (1, 18417)    0.30378311943002884
  :     :
  (62565, 97275)        0.28936441220105635
  (62565, 95081)        0.2348139560572062
  (62565, 40544)        0.0891391813426357
  (62565, 38976)        0.4648498926257003
  (62566, 231819)       0.20847620247930077
  (62566, 181269)       0.48377444121309077
  (62566, 109629)       0.24562569008622553
  (62566, 99045)        0.23134246174964349
  (62566, 95174)        0.4809300870908935
  (62566, 60163)        0.3154078528326994
  (62566, 40544)        0.0709467285369086
  (62566, 33230)        0.32278170777745163
  (62566, 3858) 0.3165464826474948
  (62566, 3038) 0.2616288241955472
  (62567, 170491)       0.2832689361066491
  (62567, 157775)       0.21116740893992927
  (62567, 97275)        0.1865383709430024
  (62567, 63356)        0.3292239125437421
  (62567, 63289)        0.46174784159905685
  (62567, 29581)        0.18459288630671175
  (62567, 12608)        0.22990253892606724
  (62567, 3356) 0.5223702254477449
  (62567, 3329) 0.39871124147497394
  (62568, 202217)       0.9940037921405545
  (62568, 40544)        0.10934560443930737
```

In [56]:
```
Y_test
```

Out[56]:
```
array(['good', 'bad', 'good', ..., 'bad', 'good', 'good'], dtype=object)
```

In [57]:
```
print(Y_test)
```

```
['good' 'bad' 'good' ... 'bad' 'good' 'good']
```

In [59]:
```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
```

In [60]:
```
Y_train = le.fit_transform(Y_train)
Y_train
```

Out[60]:
```
array([1, 1, 1, ..., 0, 1, 1])
```

In [61]:
```
Y_test = le.fit_transform(Y_test)
Y_test
```

Out[61]:
```
array([1, 0, 1, ..., 0, 1, 1])
```

In [62]:
```
model = LogisticRegression()
model.fit(X_train, Y_train)
```

```
C:\Users\ASUS\AppData\Roaming\Python\Python311\site-packages\sklearn\linear_model\_logistic.py:460: ConvergenceW
arning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
```

Out[62]:

▾ LogisticRegression

LogisticRegression()

In [63]:
```
model.predict(X_test)
```

Out[63]:
```
array([1, 0, 1, ..., 0, 1, 1])
```

In [64]:
```
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

In [65]:
```
training_data_accuracy
```

Out[65]:
```
0.9721506343022676
```

In [66]:
```
X_test_prediction = model.predict(X_test)
testing_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

In [67]:
```
testing_data_accuracy
```

Out[67]:
```
0.953235627866835
```

In [68]:
```
data.text_sent[0]
```

Out[68]:
```
'nobel it ffb d dca cce f login skype com en cgi bin verif login ffb d dca cce f index php cmd profil ach outda
t page tmpl p gen fail to load nav login access'
```

In [69]:
```
testing = data.text_sent[0]
testing = [testing]
testing = vector.transform(testing)
model.predict(testing)
```

Out[69]:
```
array([0])
```

In [72]:
```
import pickle
with open('model.pickle', 'wb') as file:
     pickle.dump(model, file)
```

In [73]:
```
with open('model.pickle', 'rb') as file:
     model = pickle.load(file)
```

In [ ]:
```

```