Importing libraries

In [34]:
```
import re
import string
import scipy
import pickle
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import *
from sklearn.preprocessing import LabelEncoder
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import BernoulliNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import LinearSVC
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt
from prettytable import PrettyTable
from joblib import dump, load
from astropy.table import Table, Column
```

Data Understanding

In [35]:
```
df=pd.read_excel("ds assign sp17-bcs-c.xlsx")
```

In [36]:
```
df.head()
```

Out[36]:

	Serial	News	Link	Category
0	1	...لاہور: اعتماد سے مالا مال پاکستان دیارغیرمیں	https://www.express.pk/story/1045388	sports
1	2	...کرائسٹ چرچ: کینگروز کے بعد پاکستانی جونیئر کرک	https://www.express.pk/story/1045385	sports
2	3	...مانچسٹر: انگلش پریمیئر فٹبال لیگ میں مانچسٹر ی	https://www.express.pk/story/1043046	sports
3	4	...لاہور: پاکستان کرکٹ ٹیم کے نیوزی لینڈ پہنچتے ہ	https://www.express.pk/story/1042194	sports
4	5	...میلبورن: آسٹریلوی کپتان اسٹیون اسمتھ کے بیٹ س	https://www.express.pk/story/1042193	sports

In [37]:
```
df.Category.value_counts()
```

Out[37]:
```
science     849
health      849
sports      849
business    849
Name: Category, dtype: int64
```

In [38]:
```
train_Sports=df[df['Category']=='sports']
train_Health=df[df['Category']=='health']
train_Business=df[df['Category']=='business']
train_Science=df[df['Category']=='science']
```

Data Preprocessing

In [39]:
```
train_Sports_pre=pd.Series(' '.join(train_Sports['News'].astype(str)).lower().split(" ")).value_counts()
# printing words and their count used by 'male'
print("\n\nWords used by 'Sports' Category in train data:")
print("=======================================\n")
print("Counts    Words\n")
print(train_Sports_pre)
```

```
Words used by 'Sports' Category in train data:
=======================================

Counts    Words

8461                              میں
8314                              کے
5989                              کی
4627                              سے
4371                              نے
3949                              کا
3766                              اور
3520                              کو
3151                              پر
2533                              کہ
2312                              بھی
1613                              اس
1387                              ٹیم
1374                              نے
1124                         پاکستان
1042                             نہیں
1007                              کر
991                               ایک
889                               کیا
867                               ہے،
861                               رنز
851                               پی
832                               بعد
825                               ہی
821                              کرکٹ
819                              میچ
809                               ان
801                              کرنے
796                              ہوئے
```

```
782                                کیلیے
                                   ...
3481                               رنز
1                                  طیارہ
1                                 اسکلیٹون
1                                 اورامید
1                                 گئے۔مگر
1                                 بدنصیبی
1                                  محتاج
1                                 چھڑکاؤ
1                                 پلیئرز،
1                                 اتھلیٹک
down?watch                            1
،1                                 ملحقہ
1                                 بھربھرے
1                                  جیہان
1                                 پڑگئی
1                                  مری،
pic.twitter.com/deuuc81ot5—           1
1                                  لہجے
maroof's                              1
1                                  فاخر
(1                                 ڈائی)
1                                 ہے۔جو
1                                 ہے۔‘
1                                 utterتھا۔
1                                 گے۔گرومنگ
1                                کیا،پاکستان
161                                وکٹیں
1\n\nبیں۔                          بریسٹ
’’1                                کسی
1                                  مرحلے
Length: 19482, dtype: int64
```

In [40]:
```
train_Business_pre=pd.Series(' '.join(train_Business['News'].astype(str)).lower().split(" ")).value_counts()
# printing words and their count used by 'male'
print("\n\nWords used by 'Business' Category in train data:")
print("========================================\n")
print("Counts    Words\n")
print(train_Business_pre)
```

```
Words used by 'Business' Category in train data:
========================================

Counts    Words

11430              کے
10229              کی
8003               میں
5585               سے
4613               اور
3283               کا
3265               کو
3015               نے
2802               پر
2456               کہ
2444               ہے
1757               اس
1753           پاکستان
1615               بھی
1476              روپے
1378              لیے
1195              ڈالر
1143              کرنے
1040               کیا
1031               کر
953               سال
894               ہے۔
894               جس
847               جانب
833               ٹیکس
827              حکومت
812              فیصد
791                تک
767                ایک
```

```
755          ہیں
               ...
1          لاپٹخا
1       پاسکوانتظامیہ
1          اگرجاپان
1          کے840روپے
1          برونائی،
1          ہوگیا،ڈسٹڑی
1          ہیں۔جرمنی
1          جسٹرڈ
1          شکوہ
1          جنرزکے
1          تاجردشمن
1          غداری
1          مداحوں
31.241     فیصد
1          کمی  نیپرا
1          چہل
1          ریگولرائز
1          گا۔ایشیائی
1          کرجاتی
1          جاسکیں۔روس
1          پرزوردیا۔
1          ریفنڈزدے
1          گی۔100
1          افئیر
1          عبدالحمید
1          ہٹی
1          تعمیرکئے
1          ملزبھارت
1          کراچی:  وفاق
1          پڑھی
Length: 19003, dtype: int64
```

In [41]:
```
train_Science_pre=pd.Series(' '.join(train_Science['News'].astype(str)).lower().split(" ")).value_counts()
# printing words and their count used by 'male'
print("\n\nWords used by 'Science' Category in train data:")
print("=======================================\n")
print("Counts    Words\n")
print(train_Sports_pre)
```

```
Words used by 'Science' Category in train data:
=======================================

Counts    Words

8461                                 میں
8314                                 کے
5989                                 کی
4627                                 سے
4371                                 نے
3949                                 کا
3766                                 اور
3520                                 کو
3151                                 پر
2533                                 کہ
2312                                 بھی
1613                                 اس
1387                                 ٹیم
1374                                 ہے
1124                            پاکستان
1042                                 نہیں
1007                                 کر
991                                  ایک
889                                  کیا
867                                  ہے،
861                                  رنز
851                                  پی
832                                  بعد
825                                  ہی
821                                 کرکٹ
819                                  میچ
809                                  ان
801                                 کرنے
796                                 ہوئے
```

```
782                                کیلیے
                                   ...
3481                               رنز
1                                  طیارہ
1                                 اسکلیٹون
1                                 اورامید
1                                 گئے۔مگر
1                                 بدنصیبی
1                                  محتاج
1                                 چھڑکاؤ
1                                 پلیئرز،
1                                 اتھلیٹک
down?watch                            1
،1                                 ملحقہ
1                                 بھربھرے
1                                  جیہان
1                                 پڑگئی
1                                  مری،
pic.twitter.com/deuuc81ot5—           1
1                                  لہجے
maroof's                              1
1                                  فاخر
(1                                 ڈائی)
1                                 ہے۔جو
1                                 ہے۔‘
utterتھا۔                             1
1                                گے۔گرومنگ
1                                کیا،پاکستان
161                                وکٹیں
ہیں۔\n\n1                          بریسٹ
’’1                                کسی
1                                  مرحلے
Length: 19482, dtype: int64
```

In [42]:
```
train_Health_pre=pd.Series(' '.join(train_Health['News'].astype(str)).lower().split(" ")).value_counts()
# printing words and their count used by 'male'
print("\n\nWords used by 'Health' Category in train data:")
print("=======================================\n")
print("Counts    Words\n")
print(train_Health_pre)
```

```
Words used by 'Health' Category in train data:
=======================================

Counts    Words

14654                          کے
13022                          میں
10694                          کی
9431                           سے
8660                           اور
6930                           کا
6239                           نے
5296                           کو
5012                           اس
4341                           کہ
3967                           بھی
3697                           پر
```

In [ ]:

In [43]:
```
STOP_WORDS = frozenset("""
آ آئی آئیں آئے آتا آتی آتے آداب آدھ آدھا آدھی آدھے آس
آمدید آنا آنسہ آنی آنے آپ آگے آہ آہا آیا اب ابھی ابے
اتوار ارب اربویں ارے اس اسکا اسکی اسکے اسی اسے اف افوہ الاول البتہ
الثانی الحرام السلام الف المکرم ان اندر انکا انکی انکے انہوں انہی انہیں
اوئے اور اوپر اوہو اپ اپنا اپنوں اپنی اپنے اپنےآپ اکبر اکثر اگر اگرچہ
اگست اہاہا ایسا ایسی ایسے ایک بائیں بار بارے بالکل باوجود باہر بج بجے
بخیر برسات بشرطیکہ بعض بغیر بلکہ بن بنا بناؤ بند بڑی بھر بھریں
بھی بہار بہت بہتر بیگم تاکہ تاہم تب تجھ تجھی تجھے ترا تری
تلک تم تمام تمہارا تمہاروں تمہاری تمہارے تمہیں تو تک تھا تھی تھیں تھے
تہائی تیرا تیری تیرے تین جا جاؤ جائیں جائے جاتا جاتی جاتے جانی جانے
جب جبکہ جدھر جس جسے جن جناب جنہوں جنہیں جو جہاں جی جیسا
جیسوں جیسی جیسے جیٹھ حالانکہ حالاں حصہ حضرت خاطر خالی خدا خزاں خواہ خوب
خود دائیں درمیان دریں دو دوران دوسرا دوسروں دوسری دوشنبہ دوں دکھائیں دگنا دی
دیئے دیا دیتا دیتی دیتے دیر دینا دینی دینے دیکھو دیں دیے دے ذریعے
رکھا رکھتا رکھتی رکھتے رکھنا رکھنی رکھنے رکھو رکھی رکھے رہ رہا رہتا
رہتی رہتے رہنا رہنی رہنے رہو رہی رہیں رہے ساتھ سامنے ساڑھے سب سبھی
سراسر سلام سمیت سوا سوائے سکا سکتا سکتے سہ سہی سی سے شام شاید
شکریہ صاحب صاحبہ صرف ضرور طرح طرف طور علاوہ عین فروری فقط فلاں
فی قبل قطا لائی لائے لاتا لاتی لاتے لانا لانی لایا لو لوجی لوگوں
لگ لگا لگتا لگتی لگی لگیں لگے لہذا لی لیا لیتا لیتی لیتے لیکن
لیں لیے لے ماسوا مت مجھ مجھی مجھے محترم محترمہ محترمی محض مرا مرحبا
مری مرے مزید مس مسز مسٹر مطابق مطلق مل منٹ منٹوں مکرمی مگر
مگھر مہربانی میرا میروں میری میرے میں نا نزدیک نما نو نومبر نہ نہیں
نیز نیچے نے و وار واسطے واقعی والا والوں والی والے واہ وجہ ورنہ
وعلیکم وغیرہ ولے وگرنہ وہ وہاں وہی وہیں ویسا ویسے ویں پاس
پایا پر پس پلیز پون پونا پونی پونے پھاگن پھر پہ پہر پہلا پہلی
پہلے پیر پیچھے چاہئے چاہتے چاہیئے چاہے چلا چلو چلیں چلے چناچہ چند چونکہ
چوگنی چکی چکیں چکے چہارشنبہ چیت ڈالنی ڈالنے ڈالے کئے کا کاتک کاش کب
کبھی کدھر کر کرتا کرتی کرتے کرم کرنا کرنے کرو کریں کرے کس
کسی کسے کل کم کن کنہیں کو کوئی کون کونسا کونسے کچھ کہ کہا
کہاں کہہ کہی کہیں کہے کی کیا کیسا کیسے کیونکر کیونکہ کیوں کیے کے
گئی گئے گا گرما گرمی گنا گو گویا گھنٹا گھنٹوں گھنٹے گی گیا
ہائیں ہائے ہاڑ ہاں ہر ہرچند ہرگز ہزار ہفتہ ہم ہمارا ہماری ہمارے ہمی
ہمیں ہو ہوئی ہوئیں ہوئے ہوا ہوبہو ہوتا ہوتی ہوتیں ہوتے ہونا ہونگے ہونی
ہونے ہوں ہی ہیلو ہیں ہے یا یات یعنی یک یہ یہاں یہی یہیں
""".split())
```

In [56]:
```
vector = CountVectorizer(
                        strip_accents='unicode',
                        analyzer='word',
                        token_pattern=r'\w{1,}',
                        stop_words=STOP_WORDS,
                        ngram_range=(1, 1),
                        max_features=500)
#concatenate train and test data
allData = df
print(allData)
# fit_transform vector on all Dataset (comments column)
all_data_features = vector.fit_transform(allData["News"])
```

```
      Serial                                                 News  \
0          1   اعتماد سے مالا مال پاکستان دیارغیرمیں  :لاہور...
1          2  کرائسٹ چرچ: کینگروز کے بعد پاکستانی جونیئر کرک...
2          3  مانچسٹر: انگلش پریمیئر فٹبال لیگ میں مانچسٹر ی...
3          4  لاہور: پاکستان کرکٹ ٹیم کے نیوزی لینڈ پہنچتے ہ...
4          5  میلبورن: آسٹریلوی کپتان اسٹیون اسمتھ کے بیٹ س...
5          6  کولمبو: سری لنکن پیسر لسیتھ مالنگا کی جانب سے ...
6          7  کراچی: پاکستانی اوپنر محمد حفیظ نے نیوزی لینڈ ...
7          8  کیپ ٹاؤن: بھارتی کپتان ویرات کوہلی نے جنوبی اف...
8          9  لاہور: 2017 کا سال قومی کھیل ہاکی سمیت دوسرے گ...
9         10  پاکستان میں امن و امان کی بہتر ہوتی صورتحال کے...
10        11  ء میں پاکستانی ٹیم نیوزی لینڈ کے دورے پ1978-79...
11        12  کیپ ٹاؤن: بھارتی اوپنر شیکھر دھون طویل مسافت ک...
12        13  میلبورن: انگلش کوچ ٹریور بیلس نے میلبورن ٹیسٹ ...
13        14  ابو ظہبی: سابق ورلڈ نمبر ون نووک جوکووک نے کہن...
14        15  نیلسن: سیریز کے پہلے ٹوئنٹی 20 انٹرنیشنل میں ...
15        16  جوہانسبرگ: 5 جنوری سے کیپ ٹاؤن میں شیڈول پہلے ...
16        17  میلبورن: آسٹریلیا کیخلاف انڈر19 کرکٹ سیریز می...
17        18  کراچی: پی ایچ ایف کے سیکریٹری اولمپئن شہباز...
```

In [57]:
```
all_data_feature_names = vector.get_feature_names()

#print feature names
print("Feature Name (Bag of words) : \n", all_data_feature_names)
```

```
Feature Name (Bag of words) :
 ['1', '10', '11', '12', '15', '16', '2', '20', '2017', '3', '30', '4', '5', '50', '6', '7', '8', '9', 'ball',
'com', 'pic', 'twitter', ', ', 'ا', 'اباد', 'ابتدایی', 'اثر', 'اثرات', 'اج', 'اجلاس', 'احمد', 'اخری', 'ادا', 'اداروں', 'ادارے', 'ار',
'اسان', 'استعمال', 'اسلام', 'اسمارٹ', 'اسٹریلیا', 'اسٹیٹ', 'اسٹیڈیم', 'اشیا', 'اضافہ', 'اضافے', 'اعلان', 'اغاز', 'افراد', 'اقدامات', 'اقسا
م', 'الاقوامی', 'امراض', 'امریکا', 'امریکی', 'انتہایی', 'انسان', 'انسانی', 'انداز', 'اننگز', 'انٹرنیشنل', 'انکشاف', 'انہوں', 'انہیں', 'انے',
'او', 'اوٹ', 'اینده', 'ایی', 'ایے', 'اچھی', 'اگلے', 'اگے', 'اہم', 'ای', 'ایس', 'ایسوسی', 'ایشن', 'ایف', 'ایل', 'ایم', 'این', 'اینٹ
ی', 'اینڈ', 'ایونٹ', 'ایپ', 'ایپل', 'ایچ', 'ایکسپریس', 'اے', 'بات', 'باعث', 'بال', 'بالوں', 'بتایا', 'بجایے', 'بجلی', 'بجٹ', 'برامدا
ت', 'برس', 'بعد', 'بلڈ', 'بنانے', 'بنایے', 'بنایا', 'بورڈ', 'بچوں', 'بچے', 'بڑا', 'بڑھ', 'بڑے', 'بک', 'بھارت', 'بھارتی', 'بھرپو
ر', 'بہترین', 'بی', 'بیماری', 'بیماریوں', 'بین', 'بینک', 'بے', 'تبدیل', 'تبدیلی', 'تجارت', 'تجارتی', 'تحت', 'تحقیق', 'تر', 'ترقی',
'ترین', 'تصاویر', 'تصویر', 'تعداد', 'تعلق', 'تقریبا', 'توانایی', 'توجہ', 'تھری', 'تیار', 'تیز', 'تیزی', 'تیل', 'ثابت', 'جارہا', 'جاری',
'جاسکتا', 'جاسکتی', 'جاسکے', 'جان', 'جانب', 'جایزہ', 'جاییں', 'جایے', 'جدید', 'جسم', 'جسمانی', 'جلد', 'جمع', 'جنوبی', 'جگہ',
'حاصل', 'حال', 'حامل', 'حد', 'حرارت', 'حسن', 'حل', 'حوالے', 'حکام', 'حکومت', 'خاص', 'خان', 'خبر', 'ختم', 'خصوصی', 'خط
ره', 'خلاف', 'خواتین', 'خوراک', 'خون', 'خیال', 'دار', 'درامد', 'درجہ', 'درد', 'درست', 'دریافت', 'دس', 'دستیاب', 'دل', 'دماغ',
'دماغی', 'دن', 'دنیا', 'دوا', 'دوباره', 'دودھ', 'دور', 'دوسرے', 'دونوں', 'دکھایی', 'دیکھ', 'دیکھا', 'دیکھنے', 'دیگر', 'ذرایع', 'ذیابی
طس', 'رات', 'رفتار', 'رقم', 'رنز', 'رواں', 'روبوٹ', 'روز', 'روزانہ', 'روشنی', 'روپے', 'رپورٹ', 'رکھ', 'ریڈیو', 'ریکارڈ', 'زاي
د', 'زمین', 'زندگی', 'زیاده', 'سابق', 'سال', 'سالانہ', 'سامنا', 'سائنس', 'سایٹ', 'سبب', 'سخت', 'سر', 'سرمایہ', 'سسٹم', 'سطح',
'سفر', 'سندھ', 'سوشل', 'سونے', 'سپر', 'سکتی', 'سکیں', 'سکے', 'سہولت', 'سیریز', 'سیکیورٹی', 'شامل', 'شایع', 'شخص', 'شده',
'شدید', 'شرح', 'شروع', 'شعبے', 'شکار', 'شکست', 'شکل', 'شہر', 'صارفین', 'صاف', 'صحت', 'صدر', 'صلاحیت', 'صورت', 'ضرو
رت', 'ضروری', 'طبی', 'طریقہ', 'طویل', 'ظاہر', 'عالمی', 'عام', 'عرصے', 'علاج', 'علی', 'عمر', 'عمل', 'غذا', 'غیر', 'فایده',
'فاینل', 'فراہم', 'فروخت', 'فروغ', 'فوری', 'فون', 'فیس', 'فیصد', 'فیصلہ', 'قابل', 'قایم', 'قدرتی', 'قرار', 'قریب', 'قسم', 'قومی',
'قیمت', 'قیمتوں', 'لاین', 'لاکھ', 'لاہور', 'لندن', 'لوگ', 'لیے', 'لینڈ', 'لینے', 'لیگ', 'ماحول', 'مارکیٹ', 'مالی', 'ماه', 'ماہرین',
'مبتلا', 'متاثر', 'متعارف', 'متعلق', 'مجموعی', 'محدود', 'محسوس', 'محفوظ', 'محمد', 'مختلف', 'مدد', 'مرتبہ', 'مرحلے', 'مرض', 'مری
ض', 'مریضوں', 'مسایل', 'مسلسل', 'مشتمل', 'مشکل', 'مشین', 'مصنوعات', 'مضبوط', 'معلوم', 'معلومات', 'مفید', 'مقابلے', 'مقامی', 'مق
ار', 'مقصد', 'ملتان', 'ملک', 'ملکی', 'ملین', 'ملے', 'ممالک', 'ممکن', 'مند', 'منصوبہ', 'منصوبے', 'موبایل', 'موجود', 'موسم', 'موق
ع', 'مکمل', 'میچ', 'میچز', 'میڈیا', 'نام', 'نتایج', 'نتیجے', 'نظام', 'نظر', 'نقصان', 'نمبر', 'نوٹ', 'نیی', 'نیے', 'نیا', 'نیشنل',
'واضح', 'واقع', 'واپس', 'ورلڈ', 'وزارت', 'وزن', 'وزیر', 'وفاقی', 'وقت', 'ون', 'وٹامن', 'وکٹ', 'وی', 'ویب', 'ویسٹ', 'ویڈیو',
'ٹوینٹی', 'ٹی', 'ٹیسٹ', 'ٹیم', 'ٹیکس', 'ٹیکنالوجی', 'پالیسی', 'پانچ', 'پانی', 'پاکستان', 'پاکستانی', 'پروفیسر', 'پروگرام', 'پشاور', 'پنجا
ب', 'پوائنٹس', 'پورا', 'پوری', 'پورے', 'پہنچ', 'پی', 'پیدا', 'پیداوار', 'پیش', 'پیک', 'چار', 'چاے', 'چاہیے', 'چھوٹے', 'چہرے',
'چین', 'چینی', 'چیرمین', 'ڈال', 'ڈالر', 'ڈاکٹر', 'ڈی', 'ڈیٹا', 'کار', 'کارکردگی', 'کاری', 'کافی', 'کام', 'کامیاب', 'کامیابی', 'کراچ
ی', 'کردار', 'کرده', 'کردی', 'کردیا', 'کررہے', 'کرسکتے', 'کرلیا', 'کروڑ', 'کرکٹ', 'کرکٹرز', 'کرکے', 'کمپنی', 'کمپنیوں', 'کمی',
'کمیٹی', 'کنٹرول', 'کوشش', 'کویٹہ', 'کویی', 'کیی', 'کپ', 'کپتان', 'کھانے', 'کھلاڑی', 'کھلاڑیوں', 'کھیل', 'کہاکہ', 'کہتے', 'کہنا',
'کیلیے', 'کینسر', 'کےلیے', 'گرام', 'گرم', 'گروپ', 'گزشتہ', 'گوشت', 'گوگل', 'گیی', 'گیے', 'گھر', 'گیمز', 'گے', 'ہاتھ', 'ہاتھو
ں', 'ہاکی', 'ہفتے', 'ہوجاتا', 'ہوجاتی', 'ہوجاتے', 'ہوجایے', 'ہوسکتا', 'ہوسکتی', 'ہویی', 'ہویے', 'ہوگا', 'ہوگیی', 'ہوگیے', 'ہوگی',
'ہ', 'وگیا', 'یاد', 'یو', 'یونیورسٹی', 'یوں']
```

In [58]:
```
#allData.to_excel('news file.xlsx')
```

In [59]:
```
all_data_features
```

Out[59]:
```
<3396x500 sparse matrix of type '<class 'numpy.int64'>'
        with 148824 stored elements in Compressed Sparse Row format>
```

In [60]:
```
train_data=pd.DataFrame(all_data_features.toarray(),columns=all_data_feature_names)
```

In [61]:
```
train_data=train_data.assign(category=df['Category'])
```

In [62]:
```
from sklearn.preprocessing import LabelEncoder
cat=LabelEncoder()
cat=cat.fit_transform(train_data['category'])
train_data['category']=cat
```

In [63]:
```
train_data
```

Out[63]:

	1	10	11	12	15	16	2	20	2017	3	...	ہوگا	ہوگیی	ہوگیے	ہوگی	ہوگیا	یاد	یو	یونیورسٹی	یوں	category
0	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	3
1	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	1	0	0	0	3
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	3
3	0	0	0	0	0	0	0	1	0	0	...	4	0	0	0	0	0	0	1	0	3
4	0	0	0	0	2	0	0	0	0	3	...	0	0	0	0	0	0	0	0	0	3
5	0	2	0	2	0	0	0	2	0	0	...	0	0	0	0	0	2	0	0	0	3
6	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	3
7	0	0	0	0	1	0	0	0	0	0	...	2	0	0	0	0	0	0	0	0	3
8	0	1	0	0	0	0	3	0	2	0	...	0	0	0	0	0	0	0	0	0	3
9	0	0	0	0	0	0	2	1	0	1	...	0	0	0	0	0	0	0	1	0	3
10	0	0	0	0	1	0	1	1	0	2	...	2	0	0	1	0	0	0	0	0	3
11	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	3
12	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	3
13	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	3
14	1	1	0	1	1	1	3	5	0	2	...	0	0	0	0	0	0	0	0	0	3
15	0	0	0	0	2	0	0	0	0	1	...	0	1	0	0	0	0	0	0	0	3
16	4	0	0	1	0	0	3	0	0	1	...	0	1	2	0	1	0	0	0	0	3
17	0	0	0	0	0	0	0	0	0	0	...	2	0	0	0	0	0	0	0	0	3
18	0	0	0	0	0	1	0	1	0	1	...	0	0	0	1	0	0	0	0	0	3
19	0	1	0	1	0	0	1	0	1	0	...	0	0	0	0	0	0	1	0	0	3
20	0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	3
21	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	3
22	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	3
23	0	1	0	0	0	0	1	1	0	0	...	0	0	1	0	0	0	0	0	0	3
24	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1	0	0	0	3

	1	10	11	12	15	16	2	20	2017	3	...	ہوگا	ہوگیی	ہوگیے	ہوگی	ہوگیا	یاد	یو	یونیورسٹی	یوں	category
25	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	1	1
26	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
27	0	1	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	2	0	1
28	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
29	1	0	0	0	0	0	2	0	0	0	...	0	1	0	0	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3366	0	1	0	0	0	0	2	0	0	0	...	3	0	0	4	0	0	0	0	0	2
3367	0	0	0	0	2	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	2
3368	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	2
3369	0	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0	1	0	0	2
3370	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	2
3371	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	2
3372	1	0	0	1	0	0	0	2	0	0	...	0	0	0	1	0	0	0	0	0	0
3373	0	0	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	0
3374	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	0
3375	0	1	0	0	0	1	1	0	2	1	...	0	0	0	1	0	0	0	0	0	0
3376	0	0	0	0	0	0	0	0	0	0	...	0	2	0	0	0	0	0	0	0	0
3377	0	0	0	0	0	0	1	0	0	1	...	0	0	0	0	0	0	0	0	0	0
3378	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	2	0	0	0
3379	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
3380	1	1	0	0	1	0	0	0	2	0	...	0	0	0	0	0	0	0	0	0	0
3381	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3382	0	0	0	0	0	0	0	0	2	0	...	0	0	0	0	0	0	0	0	0	0
3383	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	0
3384	1	0	0	0	0	1	1	1	1	1	...	0	0	0	0	0	0	0	0	0	0
3385	1	2	1	0	0	0	1	0	1	0	...	0	0	0	0	0	0	0	0	0	0
3386	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	0

	1	10	11	12	15	16	2	20	2017	3	...	ہوگا	ہوگی	ہوگیے	ہوگی	ہوگیا	یاد	یو	یونیورسٹی	یوں	category
3387	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3388	0	1	0	1	1	0	2	0	0	2	...	9	0	0	7	0	0	0	0	1	0
3389	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	0	0
3390	0	0	0	0	0	0	0	0	0	1	...	1	0	0	0	0	0	0	0	0	0
3391	0	0	0	0	0	1	0	1	0	0	...	0	0	0	0	2	0	0	0	0	0
3392	0	3	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	0
3393	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3394	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3395	0	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0	0	0

3396 rows × 501 columns

In [64]:
```
train_data.to_excel('News Final File.xlsx')
```

ML Algo Training Phase Using Train Data

In [65]:
```
from sklearn.tree import DecisionTreeClassifier
tree=DecisionTreeClassifier()
X=train_data[train_data.columns]
X=X.drop('category',axis=1)
y=train_data['category']
#print(X)
X_train,X_test,y_train,y_test=train_test_split(X,y,random_state=1,test_size=0.2)
tree.fit(X_train,y_train)
```

Out[65]:
```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                       max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort=False,
                       random_state=None, splitter='best')
```

In [66]:
```
#decision tree classifier evaluation using test data
y_pred=tree.predict(X_test)
print("accuracy of decision tree classifier is "+str(accuracy_score(y_pred,y_test)*100))
```

```
accuracy of decision tree classifier is 80.88235294117648
```

Train And Evaluate Training Data using KNN Algo

In [67]:
```
from sklearn.neighbors import KNeighborsClassifier

model = KNeighborsClassifier(n_neighbors=5)

# Train the model using the training sets
model.fit(X_train,y_train)

#Predict Output
predicted= model.predict(X_test)
#print(predicted)
print("accuracy is "+str(accuracy_score(predicted,y_test)*100))
```

```
accuracy is 77.94117647058823
```

In [ ]: