Disease-condition-detection-drug-reviews

January 21, 2024

0.0.1 Importing libraries

[2]:
```
import pandas as pd # data preprocessing
import itertools # confusion matrix
import string
import numpy as np
import seaborn as sns
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import PassiveAggressiveClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn import metrics
import matplotlib.pyplot as plt
from wordcloud import WordCloud
%matplotlib inline
# To show all the rows of pandas dataframe
pd.set_option('display.max_rows', None)
```

[3]:
```
# https://archive.ics.uci.edu/datasets
df=pd.read_csv('drugsComTrain_raw.tsv', sep='\t')
```

[4]:
```
df.head()
```

[4]:

	Unnamed: 0	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati…	9.0		
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of …	8.0		
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh…	5.0		
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth…	8.0		
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around…	9.0		

```
0       May 20, 2012           27
1     April 27, 2010          192
2  December 14, 2009           17
3   November 3, 2015           10
4  November 27, 2016           37
```

[ ]:

0.1 EDA

[54]:
```
# Print the top 10 conditions and their percentage distribution
df['condition'].value_counts(normalize=False).head(10)
```

[54]:
```
Birth Control      28788
Depression          9069
Pain                6145
Anxiety             5904
Acne                5588
Bipolar Disorde     4224
Insomnia            3673
Weight Loss         3609
Obesity             3568
ADHD                3383
Name: condition, dtype: int64
```

[7]:
```
# select 4 conditions
df_train = df[(df['condition']=='Birth Control') |␣
 ↪(df['condition']=='Depression') | (df['condition']=='High Blood␣
 ↪Pressure')|(df['condition']=='Diabetes, Type 2')]
```

[8]:
```
#shape of the original dataset
df.shape
```

[8]:
```
(161297, 7)
```

[9]:
```
# Shape of new df_train
df_train.shape
```

[9]:
```
(42732, 7)
```

[10]:
```
# Drop junk columns
X = df_train.drop(['Unnamed:␣
 ↪0','drugName','rating','date','usefulCount'],axis=1)
```

[11]:
```
X.condition.value_counts()
```

[11]:
```
Birth Control          28788
Depression              9069
Diabetes, Type 2        2554
High Blood Pressure     2321
Name: condition, dtype: int64
```

[12]:
```
X.head()
```

[12]:

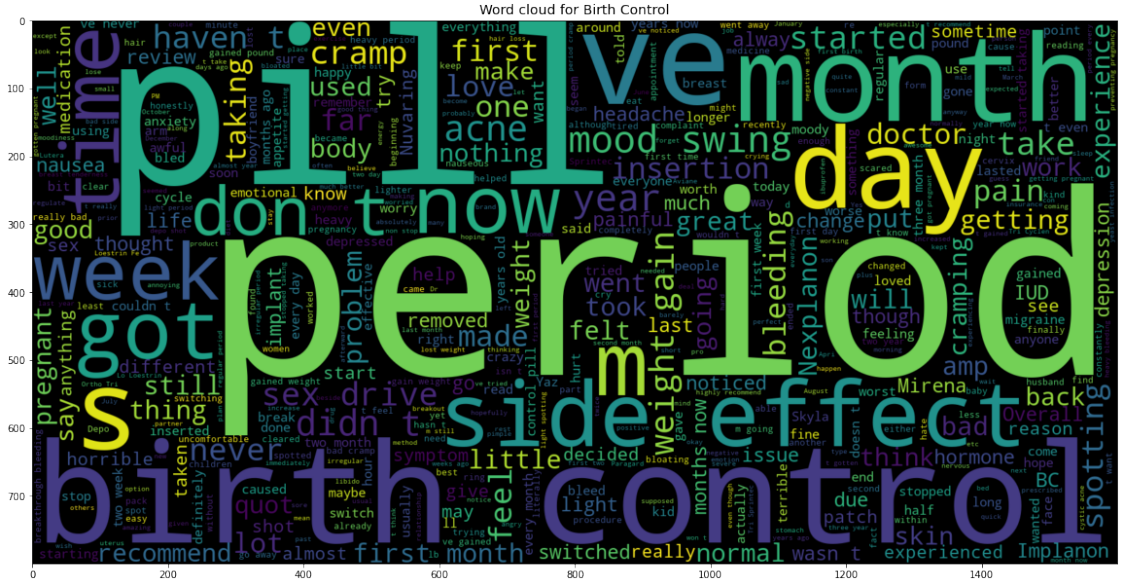
	condition	review
2	Birth Control	"I used to take another oral contraceptive, wh…
3	Birth Control	"This is my first time using any form of birth…
9	Birth Control	"I had been on the pill for many years. When m…
11	Depression	"I have taken anti-depressants for years, with…
14	Birth Control	"Started Nexplanon 2 months ago because I have…

[13]:
```
# segregating dataframe for analyzing individual condition
X_birth=X[(X['condition']=='Birth Control')]
X_dep=X[(X['condition']=='Depression')]
X_bp=X[(X['condition']=='High Blood Pressure')]
X_diab=X[(X['condition']=='Diabetes, Type 2')]
```

[ ]:
```

```

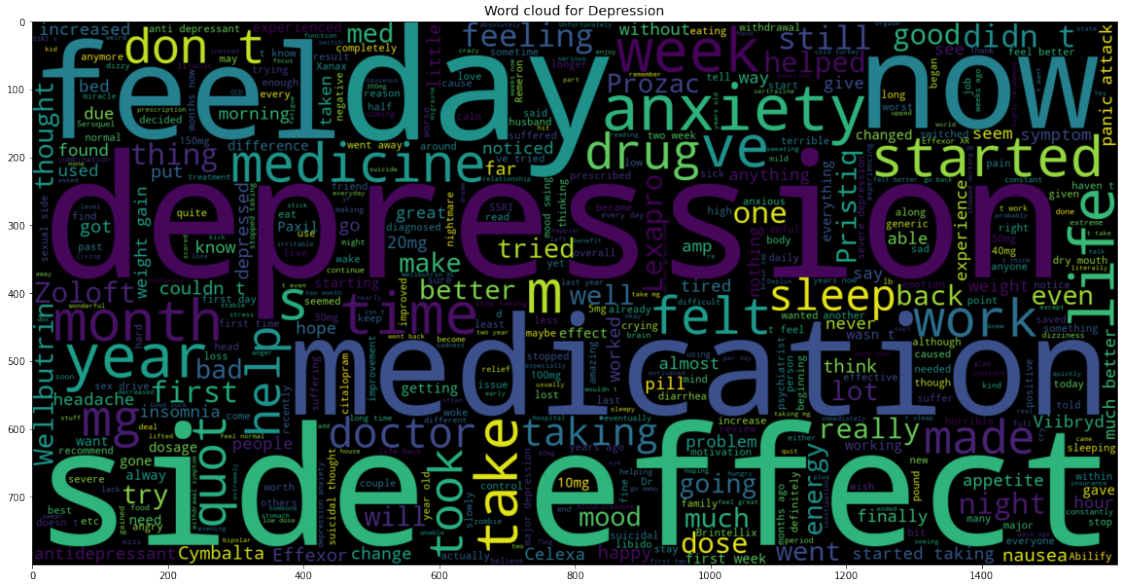
[14]:
```
plt.figure(figsize = (20,20)) # Text that is Fake News Headlines
wc = WordCloud(max_words = 500 , width = 1600 , height = 800).generate(" ".
 ↪join(X_birth.review))
plt.imshow(wc , interpolation = 'bilinear')
plt.title('Word cloud for Birth Control',fontsize=14)
```

[14]:
```
Text(0.5, 1.0, 'Word cloud for Birth Control')
```



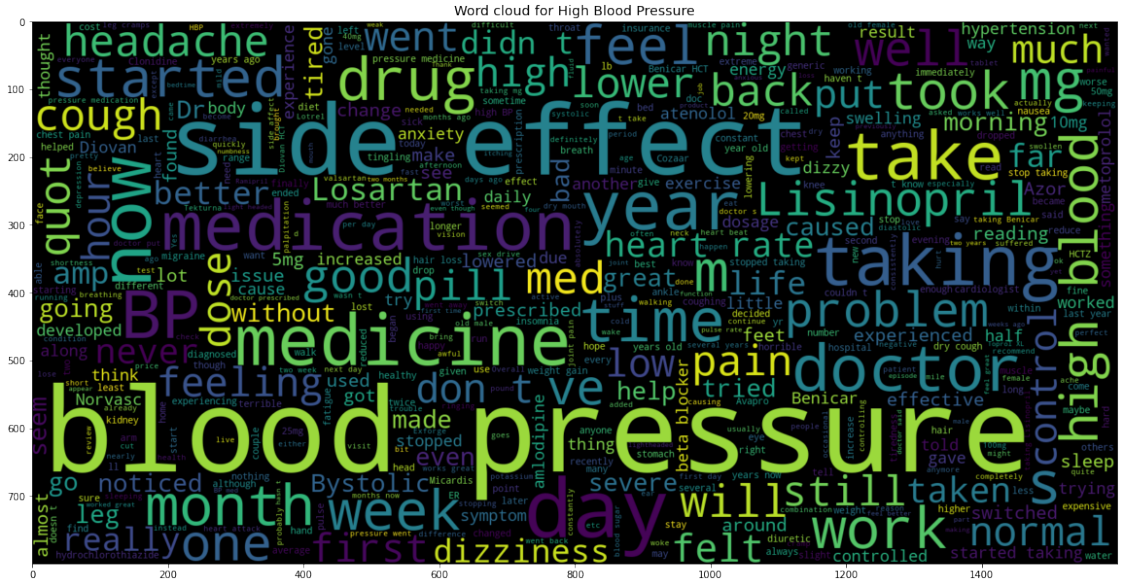
[15]:
```
plt.figure(figsize = (20,20)) # Text that is Fake News Headlines
wc = WordCloud(max_words = 500 , width = 1600 , height = 800).generate(" ".
↪join(X_dep.review))
plt.imshow(wc , interpolation = 'bilinear')
plt.title('Word cloud for Depression',fontsize=14)
```

[15]: Text(0.5, 1.0, 'Word cloud for Depression')



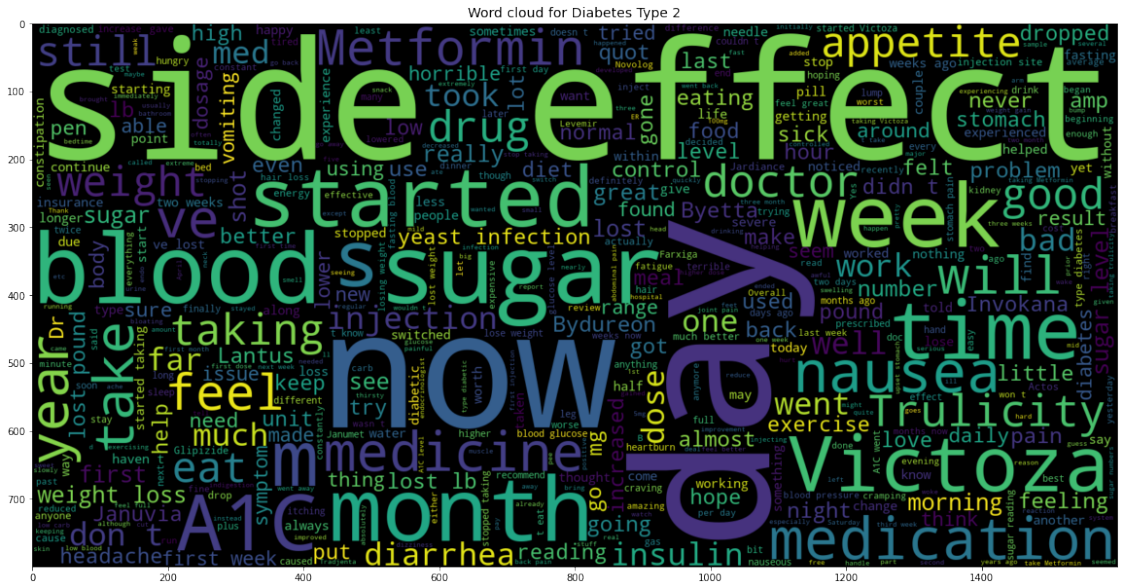
[16]:
```
plt.figure(figsize = (20,20)) # Text that is Fake News Headlines
wc = WordCloud(max_words = 500 , width = 1600 , height = 800).generate(" ".
↪join(X_bp.review))
plt.imshow(wc , interpolation = 'bilinear')
plt.title('Word cloud for High Blood Pressure',fontsize=14)
```

[16]: Text(0.5, 1.0, 'Word cloud for High Blood Pressure')



[17]:
```
plt.figure(figsize = (20,20)) # Text that is Fake News Headlines
wc = WordCloud(max_words = 500 , width = 1600 , height = 800).generate(" ".
↪join(X_diab.review))
plt.imshow(wc , interpolation = 'bilinear')
plt.title('Word cloud for Diabetes Type 2',fontsize=14)
```

[17]: Text(0.5, 1.0, 'Word cloud for Diabetes Type 2')



[ ]:

0.2 data preprocessing

[18]:
```
X['review'][2]
```

[18]:
```
'"I used to take another oral contraceptive, which had 21 pill cycle, and was
very happy- very light periods, max 5 days, no other side effects. But it
contained hormone gestodene, which is not available in US, so I switched to
Lybrel, because the ingredients are similar. When my other pills ended, I
started Lybrel immediately, on my first day of period, as the instructions said.
And the period lasted for two weeks. When taking the second pack- same two
weeks. And now, with third pack things got even worse- my third period lasted
for two weeks and now it&#039;s the end of the third week- I still have daily
brown discharge.\r\nThe positive side is that I didn&#039;t have any other side
effects. The idea of being period free was so tempting… Alas."'
```

[19]:
```
X['review'][11]
```

[19]:
```
'"I have taken anti-depressants for years, with some improvement but mostly
moderate to severe side affects, which makes me go off them.\r\n\r\nI only take
Cymbalta now mostly for pain.\r\n\r\nWhen I began Deplin, I noticed a major
improvement overnight. More energy, better disposition, and no sinking to the
low lows of major depression. I have been taking it for about 3 months now and
feel like a normal person for the first time ever. Best thing, no side
effects."'
```

[20]:
```
# Remove double quotes
for i, col in enumerate(X.columns):
    X.iloc[:, i] = X.iloc[:, i].str.replace('"', '')
```

[21]:
```
# To set the width of the column to maximum
pd.set_option('max_colwidth', -1)
```

```
C:\Users\christopher.wachira_\AppData\Local\Temp\ipykernel_16832\999061969.py:2:
FutureWarning: Passing a negative integer is deprecated in version 1.0 and will
not be supported in future version. Instead, use None to not limit the column
width.
  pd.set_option('max_colwidth', -1)
```

[22]:
```
X.head()
```

[22]:
```
         condition  \
2   Birth Control
3   Birth Control
9   Birth Control
11  Depression
```

14 Birth Control

review
2 I used to take another oral contraceptive, which had 21 pill cycle, and was very happy- very light periods, max 5 days, no other side effects. But it contained hormone gestodene, which is not available in US, so I switched to Lybrel, because the ingredients are similar. When my other pills ended, I started Lybrel immediately, on my first day of period, as the instructions said. And the period lasted for two weeks. When taking the second pack- same two weeks. And now, with third pack things got even worse- my third period lasted for two weeks and now it's the end of the third week- I still have daily brown discharge.\r\nThe positive side is that I didn't have any other side effects. The idea of being period free was so tempting… Alas.
3 This is my first time using any form of birth control. I'm glad I went with the patch, I have been on it for 8 months. At first It decreased my libido but that subsided. The only downside is that it made my periods longer (5-6 days to be exact) I used to only have periods for 3-4 days max also made my cramps intense for the first two days of my period, I never had cramps before using birth control. Other than that in happy with the patch
9 I had been on the pill for many years. When my doctor changed my RX to chateal, it was as effective. It really did help me by completely clearing my acne, this takes about 6 months though. I did not gain extra weight, or develop any emotional health issues. I stopped taking it bc I started using a more natural method of birth control, but started to take it bc I hate that my acne came back at age 28. I really hope symptoms like depression, or weight gain do not begin to affect me as I am older now. I'm also naturally moody, so this may worsen things. I was in a negative mental rut today. Also I hope this doesn't push me over the edge, as I believe I am depressed. Hopefully it'll be just like when I was younger.
11 I have taken anti-depressants for years, with some improvement but mostly moderate to severe side affects, which makes me go off them.\r\n\r\nI only take Cymbalta now mostly for pain.\r\n\r\nWhen I began Deplin, I noticed a major improvement overnight. More energy, better disposition, and no sinking to the low lows of major depression. I have been taking it for about 3 months now and feel like a normal person for the first time ever. Best thing, no side effects.
14 Started Nexplanon 2 months ago because I have a minimal amount of contraception's I can take due to my inability to take the hormone that is used in most birth controls. I'm trying to give it time because it is one of my only options right now. But honestly if I had options I'd get it removed.\r\nI've never had acne problems in my life, and immediately broke out after getting it implanted. Sex drive is completely gone, and I used to have sex with my boyfriend a few days a week, now its completely forced and not even fun for me anymore. I mean I'm on birth control because I like having sex but don't want to get pregnant, why take a birth control that takes away sex? Very unhappy and hope that I get it back with time or I'm getting it removed.

0.2.1 What are stopwords ?

Stopwords are the most common words in any natural language. For the purpose of building NLP models, these stopwords might not add much value to the meaning of the document.

The most common words used in a text are **“the”, “is”, “in”, “for”, “where”, “when”, “to”, “at”** etc.

[23]:
```
# Remove stopwords
from nltk.corpus import stopwords

stop = stopwords.words('english')
```

[24]:
```
stop
```

[24]:
```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
 'him',
 'his',
 'himself',
 'she',
 "she's",
 'her',
 'hers',
 'herself',
 'it',
 "it's",
 'its',
 'itself',
 'they',
 'them',
 'their',
 'theirs',
```

```
'themselves',
'what',
'which',
'who',
'whom',
'this',
'that',
"that'll",
'these',
'those',
'am',
'is',
'are',
'was',
'were',
'be',
'been',
'being',
'have',
'has',
'had',
'having',
'do',
'does',
'did',
'doing',
'a',
'an',
'the',
'and',
'but',
'if',
'or',
'because',
'as',
'until',
'while',
'of',
'at',
'by',
'for',
'with',
'about',
'against',
'between',
'into',
'through',
```

```
'during',
'before',
'after',
'above',
'below',
'to',
'from',
'up',
'down',
'in',
'out',
'on',
'off',
'over',
'under',
'again',
'further',
'then',
'once',
'here',
'there',
'when',
'where',
'why',
'how',
'all',
'any',
'both',
'each',
'few',
'more',
'most',
'other',
'some',
'such',
'no',
'nor',
'not',
'only',
'own',
'same',
'so',
'than',
'too',
'very',
's',
't',
```

```
'can',
'will',
'just',
'don',
"don't",
'should',
"should've",
'now',
'd',
'll',
'm',
'o',
're',
've',
'y',
'ain',
'aren',
"aren't",
'couldn',
"couldn't",
'didn',
"didn't",
'doesn',
"doesn't",
'hadn',
"hadn't",
'hasn',
"hasn't",
'haven',
"haven't",
'isn',
"isn't",
'ma',
'mightn',
"mightn't",
'mustn',
"mustn't",
'needn',
"needn't",
'shan',
"shan't",
'shouldn',
"shouldn't",
'wasn',
"wasn't",
'weren',
"weren't",
```

```
 'won',
 "won't",
 'wouldn',
 "wouldn't"]
```

0.3 Lemmitization and Stemming

Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.

Stemming is a technique used to reduce words to their base or root form. The goal of stemming is to simplify words by removing suffixes or prefixes, resulting in a common base form that represents multiple variations of a word.

For example, consider the words: "run," "running," and "runner." Through stemming, these words would be reduced to the common base form "run."

[25]:
```
from nltk.stem import WordNetLemmatizer
from nltk.stem import PorterStemmer

porter = PorterStemmer()

lemmatizer = WordNetLemmatizer()
```

Stemming vs lemmatization

[26]:
```
print(porter.stem("sportingly"))
print(porter.stem("very"))
print(porter.stem("troubled"))
```

```
sportingli
veri
troubl
```

[27]:
```
print(lemmatizer.lemmatize("sportingly"))
print(lemmatizer.lemmatize("very"))
print(lemmatizer.lemmatize("troubled"))
```

```
sportingly
very
troubled
```

Data Cleaning Function below performs a series of text cleaning operations, including HTML tag removal, character filtering, lowercase conversion, stopwords removal, and lemmatization, to prepare the text data for further analysis or machine learning tasks.

[28]:
```
from bs4 import BeautifulSoup
import re
```

[29]:
```
def review_to_words(raw_review):
    # 1. Delete HTML
    review_text = BeautifulSoup(raw_review, 'html.parser').get_text()
    # 2. Make a space
    letters_only = re.sub('[^a-zA-Z]', ' ', review_text)
    # 3. lower letters
    words = letters_only.lower().split()
    # 5. Stopwords
    meaningful_words = [w for w in words if not w in stop]
    # 6. lemmitization
    lemmitize_words = [lemmatizer.lemmatize(w) for w in meaningful_words]
    # 7. space join words
    return( ' '.join(lemmitize_words))
```

[30]:
```
X['review_clean'] = X['review'].apply(review_to_words)
```

```
C:\Users\christopher.wachira_\anaconda3\lib\site-packages\bs4\__init__.py:435:
MarkupResemblesLocatorWarning: The input looks more like a filename than markup.
You may want to open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
```

[31]:
```
X.head()
```

[31]:

	condition	review
2	Birth Control	I used to take another oral contraceptive, which had 21 pill cycle, and was very happy- very light periods, max 5 days, no other side effects. But it contained hormone gestodene, which is not available in US, so I switched to Lybrel, because the ingredients are similar. When my other pills ended, I started Lybrel immediately, on my first day of period, as the instructions said. And the period lasted for two weeks. When taking the second pack- same two weeks. And now, with third pack things got even worse- my third period lasted for two weeks and now it's the end of the third week- I still have daily brown discharge.\r\nThe positive side is that I didn't have any other side effects. The idea of being period free was so tempting… Alas.
3	Birth Control	This is my first time using any form of birth control. I'm glad I went with the patch, I have been on it for 8 months. At first It decreased my libido but that subsided. The only downside is that it made my periods longer (5-6 days to be exact) I used to only have periods for 3-4 days max also made my cramps intense for the first two days of my period, I never had cramps before using birth control. Other than that in happy with the patch
9	Birth Control	
11	Depression	
14	Birth Control	

9 I had been on the pill for many years. When my doctor changed my RX to chateal, it was as effective. It really did help me by completely clearing my acne, this takes about 6 months though. I did not gain extra weight, or develop any emotional health issues. I stopped taking it bc I started using a more natural method of birth control, but started to take it bc I hate that my acne came back at age 28. I really hope symptoms like depression, or weight gain do not begin to affect me as I am older now. I'm also naturally moody, so this may worsen things. I was in a negative mental rut today. Also I hope this doesn't push me over the edge, as I believe I am depressed. Hopefully it'll be just like when I was younger.
11 I have taken anti-depressants for years, with some improvement but mostly moderate to severe side affects, which makes me go off them.\r\n\r\nI only take Cymbalta now mostly for pain.\r\n\r\nWhen I began Deplin, I noticed a major improvement overnight. More energy, better disposition, and no sinking to the low lows of major depression. I have been taking it for about 3 months now and feel like a normal person for the first time ever. Best thing, no side effects.
14 Started Nexplanon 2 months ago because I have a minimal amount of contraception's I can take due to my inability to take the hormone that is used in most birth controls. I'm trying to give it time because it is one of my only options right now. But honestly if I had options I'd get it removed.\r\nI've never had acne problems in my life, and immediately broke out after getting it implanted. Sex drive is completely gone, and I used to have sex with my boyfriend a few days a week, now its completely forced and not even fun for me anymore. I mean I'm on birth control because I like having sex but don't want to get pregnant, why take a birth control that takes away sex? Very unhappy and hope that I get it back with time or I'm getting it removed.

review_clean
2 used take another oral contraceptive pill cycle happy light period max day side effect contained hormone gestodene available u switched lybrel ingredient similar pill ended started lybrel immediately first day period instruction said period lasted two week taking second pack two week third pack thing got even worse third period lasted two week end third week still daily brown discharge positive side side effect idea period free tempting ala
3 first time using form birth control glad went patch month first decreased libido subsided downside made period longer day exact used period day max also made cramp intense first two day period never cramp using birth control happy patch
9 pill many year doctor changed rx chateal effective really help completely clearing acne take month though gain extra weight develop emotional health issue stopped taking bc started using natural method birth control started take bc hate acne came back age really hope symptom like depression weight gain begin affect older also naturally moody may worsen thing negative mental rut today also hope push edge believe depressed hopefully like younger
11 taken anti depressant year improvement mostly moderate severe side affect make go take cymbalta mostly pain began deplin noticed major improvement

```
overnight energy better disposition sinking low low major depression taking
month feel like normal person first time ever best thing side effect
14  started nexplanon month ago minimal amount contraception take due inability
take hormone used birth control trying give time one option right honestly
option get removed never acne problem life immediately broke getting implanted
sex drive completely gone used sex boyfriend day week completely forced even fun
anymore mean birth control like sex want get pregnant take birth control take
away sex unhappy hope get back time getting removed
```

0.4 Creating features and Target Variable

[32]:
```
X_feat=X['review_clean']
y=X['condition']
```

[33]:
```
X_train, X_test, y_train, y_test = train_test_split(X_feat,␣
↪y,stratify=y,test_size=0.2, random_state=0)
```

[34]:
```
def plot_confusion_matrix(cm, classes,
                          normalize=False,
                          title='Confusion matrix',
                          cmap=plt.cm.Blues):
    """
    See full source and example:
    http://scikit-learn.org/stable/auto_examples/model_selection/
↪plot_confusion_matrix.html

    This function prints and plots the confusion matrix.
    Normalization can be applied by setting `normalize=True`.
    """
    plt.imshow(cm, interpolation='nearest', cmap=cmap)
    plt.title(title)
    plt.colorbar()
    tick_marks = np.arange(len(classes))
    plt.xticks(tick_marks, classes, rotation=45)
    plt.yticks(tick_marks, classes)

    if normalize:
        cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
        print("Normalized confusion matrix")
    else:
        print('Confusion matrix, without normalization')

    thresh = cm.max() / 2.
    for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
        plt.text(j, i, cm[i, j],
                 horizontalalignment="center",
```

```
                color="white" if cm[i, j] > thresh else "black")

    plt.tight_layout()
    plt.ylabel('True label')
    plt.xlabel('Predicted label')
```

0.4.1 Bag of Words

Bag of Words model is employed to convert the raw text data into a numerical format, creating a Document-Term Matrix (DTM) where each document is represented by the frequency of words. This DTM is then used as input to train and evaluate a machine learning classifier. Below, Bag of Words (BoW) model is applied to transform the textual drug reviews into a numerical representation suitable for machine learning models.

[35]:
```
# In train set, fit_transform
count_vectorizer = CountVectorizer(stop_words='english')

count_train = count_vectorizer.fit_transform(X_train)

count_test = count_vectorizer.transform(X_test)
```

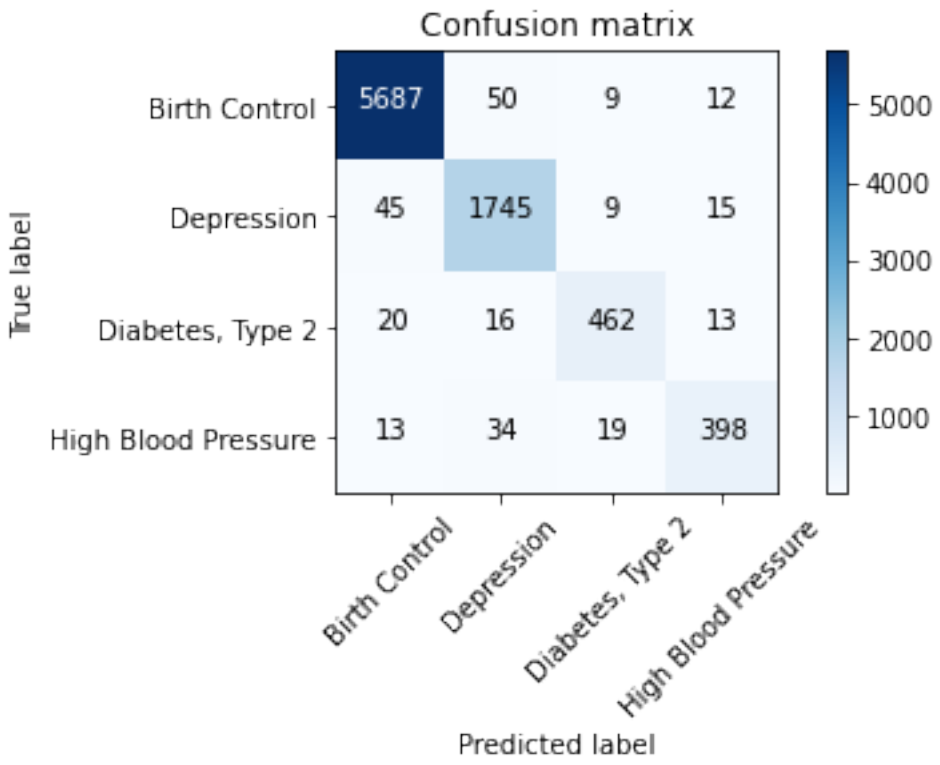
[36]:
```
count_train
```

[36]:
```
<34185x15995 sparse matrix of type '<class 'numpy.int64'>'
        with 1092752 stored elements in Compressed Sparse Row format>
```

0.5 Machine Learning Model : Naive Bayes

[37]:
```
# Train a Naive Bayes classifier using the Bag of Words (BoW) representation.
# Calculate accuracy on the test set to evaluate the model's performance.

mnb = MultinomialNB()
mnb.fit(count_train, y_train)
pred = mnb.predict(count_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)

cm = metrics.confusion_matrix(y_test, pred, labels=['Birth Control',␣
 ↪'Depression','Diabetes, Type 2','High Blood Pressure'])
plot_confusion_matrix(cm, classes=['Birth Control', 'Depression','Diabetes,␣
 ↪Type 2','High Blood Pressure'])
```

```
accuracy:   0.970
Confusion matrix, without normalization
```

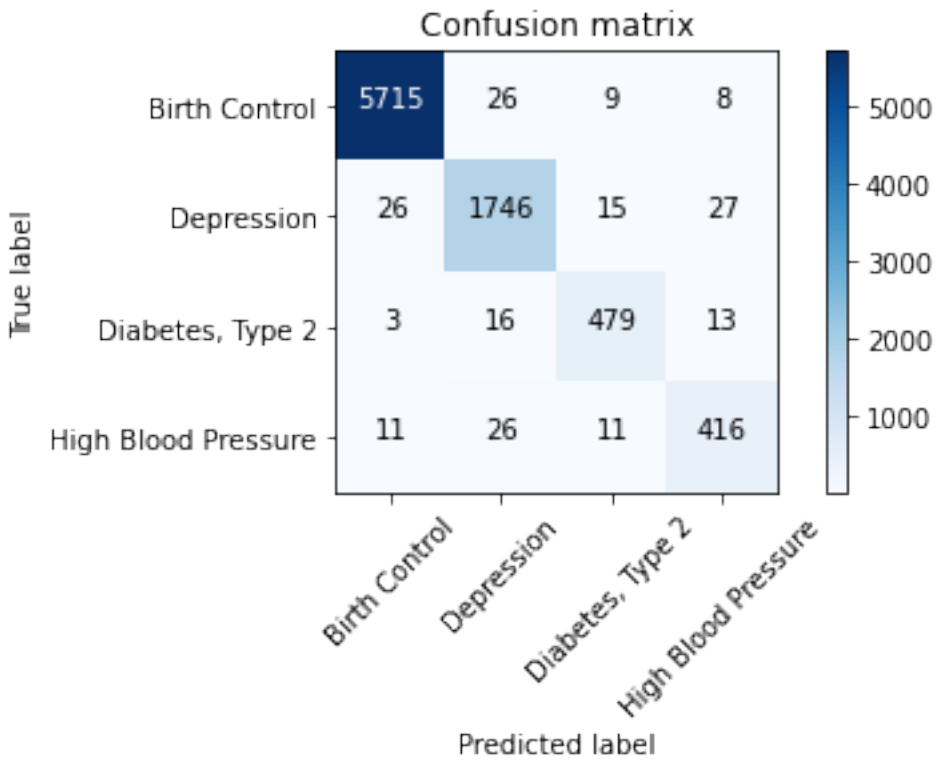


[ ]:

0.6 Machine Learning Model : Passive Aggressive Classifier

[38]:
```
from sklearn.linear_model import PassiveAggressiveClassifier,LogisticRegression

passive = PassiveAggressiveClassifier()
passive.fit(count_train, y_train)
pred = passive.predict(count_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred, labels=['Birth Control',␣
 ↪'Depression','Diabetes, Type 2','High Blood Pressure'])
plot_confusion_matrix(cm, classes=['Birth Control', 'Depression','Diabetes,␣
 ↪Type 2','High Blood Pressure'])
```

```
accuracy:   0.978
Confusion matrix, without normalization
```



0.6.1 TFIDF Vectorizer

The TF-IDF (Term Frequency-Inverse Document Frequency) Vectorizer is another text representation technique commonly used in natural language processing (NLP) and information retrieval. It addresses some limitations of the Bag of Words (BoW) model by taking into account the importance of words in a document relative to their importance in the entire corpus.

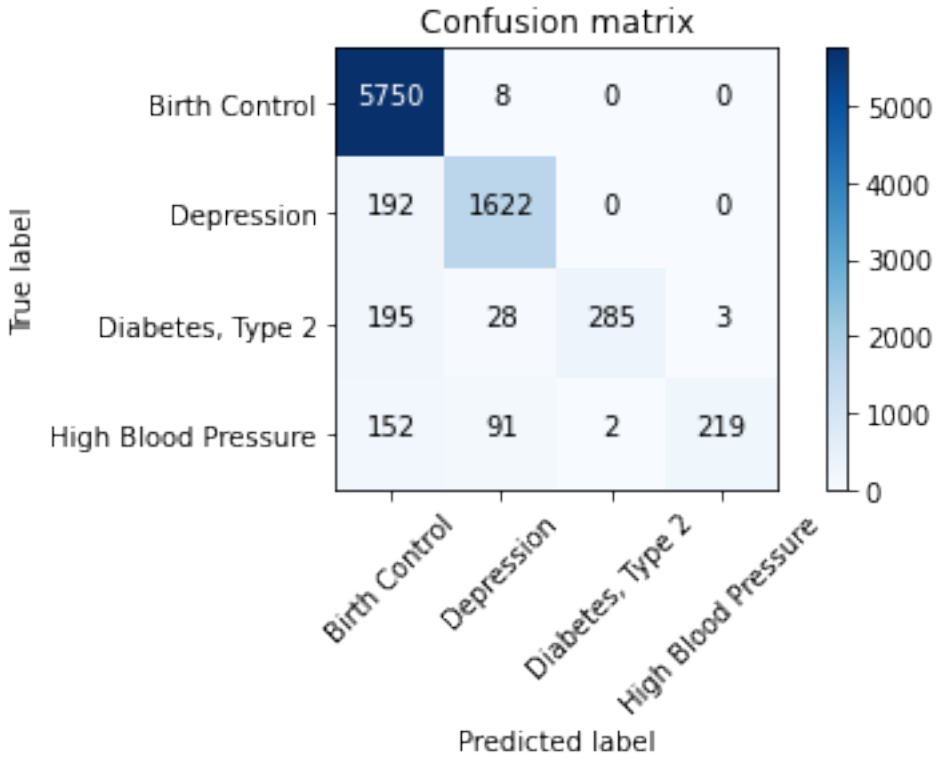
[39]:
```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf_vectorizer = TfidfVectorizer(stop_words='english', max_df=0.8)
tfidf_train_2 = tfidf_vectorizer.fit_transform(X_train)
tfidf_test_2 = tfidf_vectorizer.transform(X_test)
```

0.6.2 Naive Bayes

[40]:
```
mnb_tf = MultinomialNB()
mnb_tf.fit(tfidf_train_2, y_train)
pred = mnb_tf.predict(tfidf_test_2)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred, labels=['Birth Control',␣
↪'Depression','Diabetes, Type 2','High Blood Pressure'])
```

```
plot_confusion_matrix(cm, classes=['Birth Control', 'Depression','Diabetes,␣
 ↪Type 2','High Blood Pressure'])
```

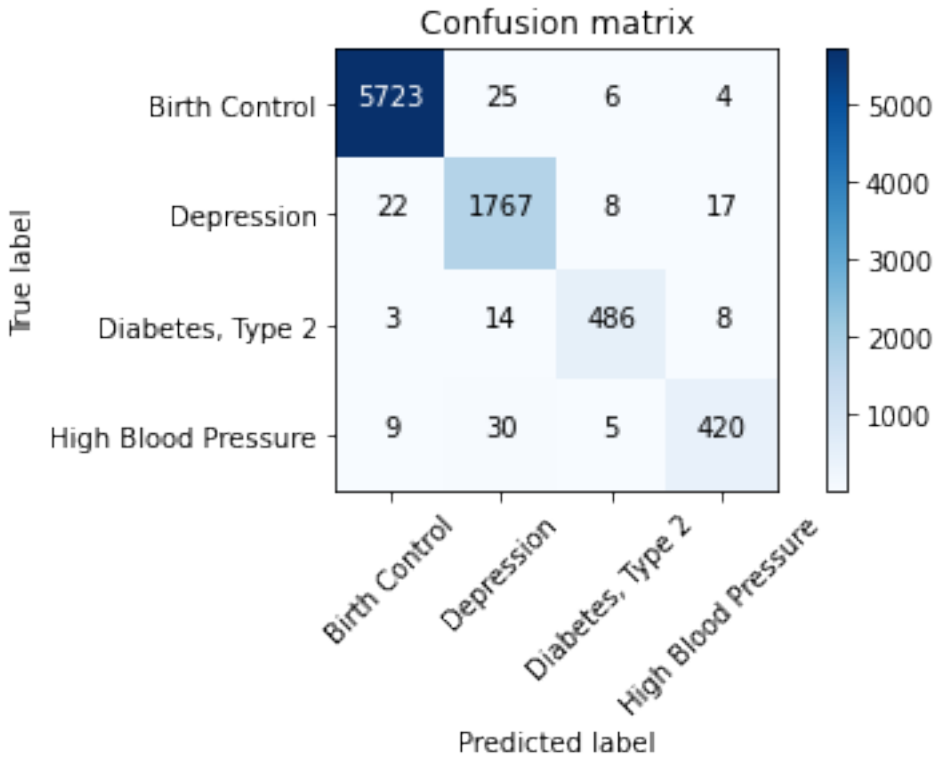
```
accuracy:   0.921
Confusion matrix, without normalization
```



0.6.3 Passive Agressive Classifier

[41]:
```
tfidf_vectorizer = TfidfVectorizer(stop_words='english', max_df=0.8)
tfidf_train = tfidf_vectorizer.fit_transform(X_train)
tfidf_test = tfidf_vectorizer.transform(X_test)

pass_tf = PassiveAggressiveClassifier()
pass_tf.fit(tfidf_train, y_train)
pred = pass_tf.predict(tfidf_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred, labels=['Birth Control',␣
 ↪'Depression','Diabetes, Type 2','High Blood Pressure'])
plot_confusion_matrix(cm, classes=['Birth Control', 'Depression','Diabetes,␣
 ↪Type 2','High Blood Pressure'])
```

```
accuracy:   0.982
Confusion matrix, without normalization
```



Passive agressive classifier seems to respond well to TFIDF

0.7 TFIDF: Bigrams

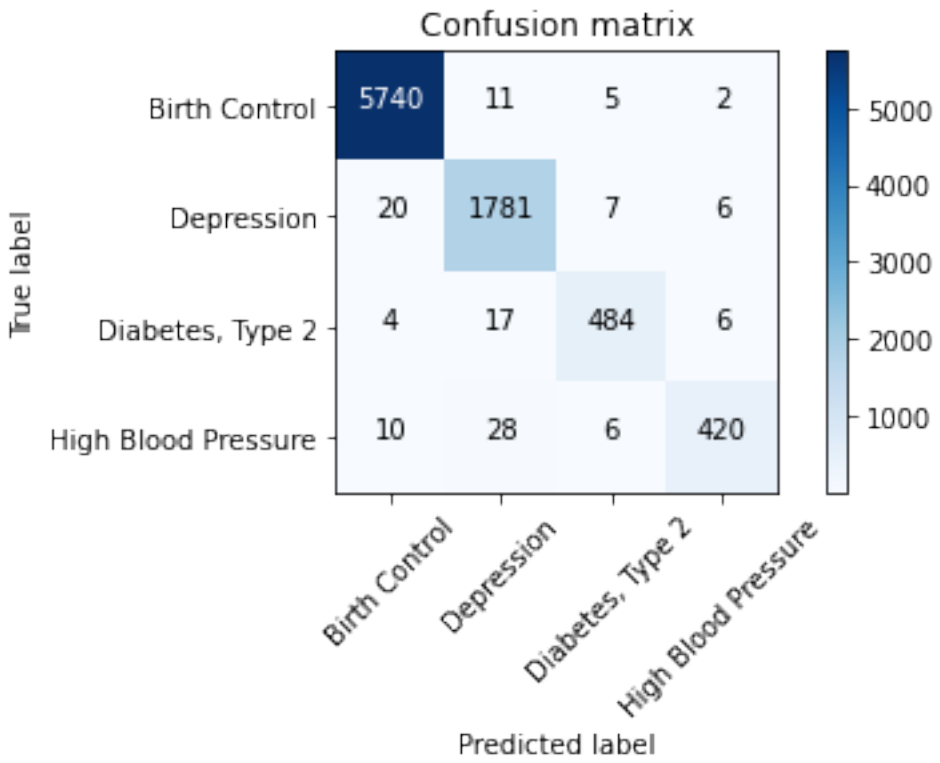
TF-IDF with Bigrams: In the context of bigrams, the TF-IDF vectorizer considers pairs of consecutive words in addition to single words. The process is similar to the standard TF-IDF, but the feature space includes both individual words and pairs of adjacent words.

[42]:
```
tfidf_vectorizer2 = TfidfVectorizer(stop_words='english', max_df=0.8,␣
↪ngram_range=(1,2))
tfidf_train_2 = tfidf_vectorizer2.fit_transform(X_train)
tfidf_test_2 = tfidf_vectorizer2.transform(X_test)
```

[43]:
```
pass_tf = PassiveAggressiveClassifier()
pass_tf.fit(tfidf_train_2, y_train)
pred = pass_tf.predict(tfidf_test_2)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred, labels=['Birth Control',␣
↪'Depression','Diabetes, Type 2','High Blood Pressure'])
```

```
plot_confusion_matrix(cm, classes=['Birth Control', 'Depression','Diabetes,␣
 ↪Type 2','High Blood Pressure'])
```

```
accuracy:   0.986
Confusion matrix, without normalization
```



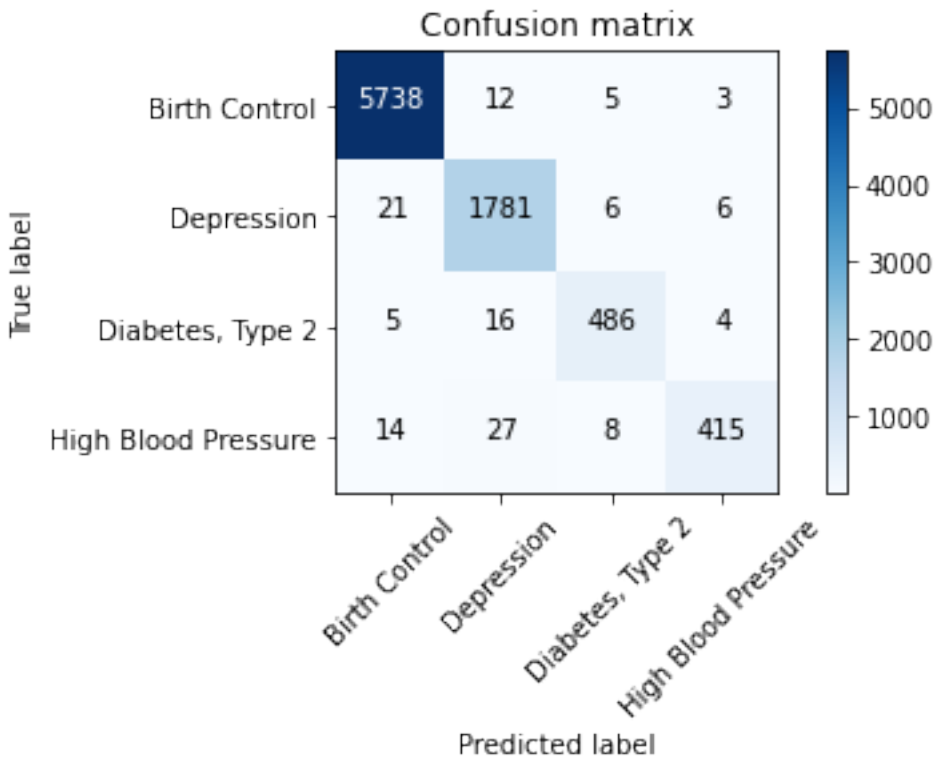
0.8 TFIDF : Trigrams

TF-IDF with Trigrams: Similarly, for trigrams, the TF-IDF vectorizer considers sequences of three consecutive words.

[44]:
```
tfidf_vectorizer3 = TfidfVectorizer(stop_words='english', max_df=0.8,␣
 ↪ngram_range=(1,3))
tfidf_train_3 = tfidf_vectorizer3.fit_transform(X_train)
tfidf_test_3 = tfidf_vectorizer3.transform(X_test)

pass_tf = PassiveAggressiveClassifier()
pass_tf.fit(tfidf_train_3, y_train)
pred = pass_tf.predict(tfidf_test_3)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred, labels=['Birth Control',␣
 ↪'Depression','Diabetes, Type 2','High Blood Pressure'])
```

```
plot_confusion_matrix(cm, classes=['Birth Control', 'Depression','Diabetes,␣
↪Type 2','High Blood Pressure'])
```

```
accuracy:   0.985
Confusion matrix, without normalization
```



0.9 Most important Features

[45]:
```
def most_informative_feature_for_class(vectorizer, classifier, classlabel,␣
↪n=10):
    labelid = list(classifier.classes_).index(classlabel)
    feature_names = vectorizer.get_feature_names()
    topn = sorted(zip(classifier.coef_[labelid], feature_names))[-n:]

    for coef, feat in topn:
        print (classlabel, feat, coef)



most_informative_feature_for_class(tfidf_vectorizer, pass_tf, 'Birth Control')
```

```
Birth Control catatonic 0.4797683621064484
```

```
Birth Control correct 0.48524472559349296
Birth Control tricyclone 0.5187155220021766
Birth Control tricylcen 0.5187155220021766
Birth Control aesthetician 0.5215445503156878
Birth Control packaged 0.5667734234181802
Birth Control addon 0.6201012990223873
Birth Control tiny 0.6571031661989708
Birth Control commit 1.3106793534622592
Birth Control freeway 7.176427164663145
```

```
C:\Users\christopher.wachira_\anaconda3\lib\site-
packages\sklearn\utils\deprecation.py:87: FutureWarning: Function
get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and will
be removed in 1.2. Please use get_feature_names_out instead.
  warnings.warn(msg, category=FutureWarning)
```

[46]:
```
most_informative_feature_for_class(tfidf_vectorizer, pass_tf, 'Depression')
```

```
Depression ben 0.5244209480256637
Depression sink 0.5359932594668678
Depression sinnce 0.6152450244611816
Depression develope 0.6360594169258027
Depression strung 0.677801119594675
Depression assaulted 0.99124930203493
Depression subjective 1.0475649283369308
Depression significantly 1.0835464952185736
Depression apnea 1.3432100395208288
Depression aliveness 3.787458957550651
```

[47]:
```
most_informative_feature_for_class(tfidf_vectorizer, pass_tf, 'High Blood␣
↪Pressure')
```

```
High Blood Pressure fatter 0.5224380929019873
High Blood Pressure fattest 0.5224380929019873
High Blood Pressure folic 0.5861203233211871
High Blood Pressure barley 0.6038451305591331
High Blood Pressure fluctuation 0.6332191507858369
High Blood Pressure encouraged 0.6796165173491691
High Blood Pressure end 0.6837131409535077
High Blood Pressure shine 0.7473789247996497
High Blood Pressure enforcement 0.7508385812540955
High Blood Pressure folk 1.0340138944338548
```

[48]:
```
most_informative_feature_for_class(tfidf_vectorizer, pass_tf, 'Diabetes, Type␣
↪2')
```

```
Diabetes, Type 2 acitve 0.3772992329555981
Diabetes, Type 2 folic 0.37991830625220335
Diabetes, Type 2 prom 0.4565725378239645
```

```
Diabetes, Type 2 fluctuation 0.48858593857652216
Diabetes, Type 2 absorbs 0.4896910849862655
Diabetes, Type 2 base 0.49415284426712386
Diabetes, Type 2 orthotricyclenlo 0.6904144083973198
Diabetes, Type 2 proliferating 0.7595604123219974
Diabetes, Type 2 fot 0.7703627961625229
Diabetes, Type 2 problem 2.74432103603248
```

[ ]:

0.10 Sample Predictions using TFIDF Passive Agressive Classifier Model

Making sample predictions using the TF-IDF model with the Passive Aggressive Classifier.

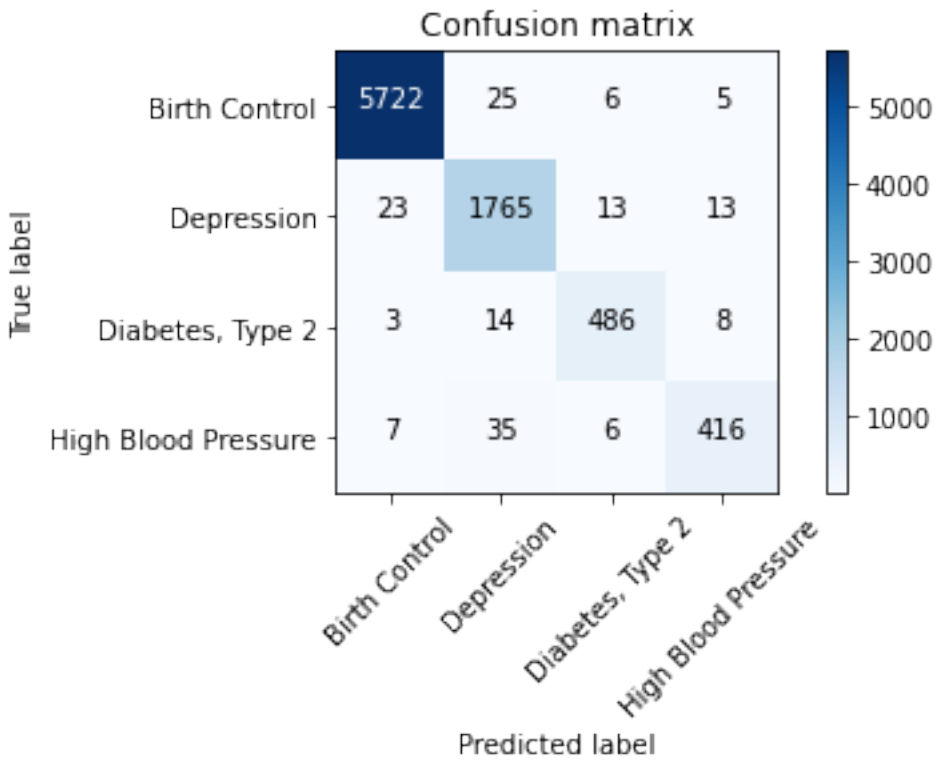
[49]:
```
# Print some test data
X.tail()
```

[49]:
```
                  condition  \
161273  Birth Control
161278  Diabetes, Type 2
161286  Depression
161290  High Blood Pressure
161291  Birth Control

                                                   review  \
161273  I have had the Nexplanon since Dec. 27, 2016 \r\r\nI got my first period
at the end of January and it lasted about a month and a half. In March of 2017 I
didn&#039;t bleed for close to three weeks and then started bleeding again March
28th and have been bleeding every since. I have gained about 13 lbs so far since
getting the birth control. Although for now the weight gain isn&#039;t a deal
breaker for me but the bleeding is..  I am trying to be very patient to see how
my body adjusts to the implant. It has been three months so far and I have my
fingers crossed that my cycle will go away for awhile.
161278  I just got diagnosed with type 2. My doctor prescribed Invokana and
metformin from the beginning. My sugars went down to normal by the second week.
I  am losing so much weight. No side effects yet. Miracle medicine for me
161286  This is the third med I&#039;ve tried for anxiety and mild depression.
Been on it for a week and I hate it so much. I am so dizzy, I have major
diarrhea and feel worse than I started. Contacting my doc in the am  and
changing asap.
161290  I have only been on Tekturna for 9 days. The effect was immediate. I am
also on a calcium channel blocker (Tiazac) and hydrochlorothiazide. I was put on
Tekturna because of palpitations experienced with Diovan (ugly drug in my
opinion, same company produces both however). The palpitations were pretty bad
on Diovan, 24 hour monitor by EKG etc. After a few days of substituting Tekturna
for Diovan, there are no more palpitations.
161291  This would be my second month on Junel. I&#039;ve been on Birth Control
for about 10 years now. I changed due to spotting and increased mood swings with
```

```
my previous birth control. Since the switch I have had shorter periods about 2-3
days, but I have gained major weight and increased appetite. I switched up my
regular exercise routine and still have not managed to drop the extra 7 lbs ;(

review_clean
161273  nexplanon since dec got first period end january lasted month half march
bleed close three week started bleeding march th bleeding every since gained lb
far since getting birth control although weight gain deal breaker bleeding
trying patient see body adjusts implant three month far finger crossed cycle go
away awhile
161278  got diagnosed type doctor prescribed invokana metformin beginning sugar
went normal second week losing much weight side effect yet miracle medicine
161286  third med tried anxiety mild depression week hate much dizzy major
diarrhea feel worse started contacting doc changing asap
161290  tekturna day effect immediate also calcium channel blocker tiazac
hydrochlorothiazide put tekturna palpitation experienced diovan ugly drug
opinion company produce however palpitation pretty bad diovan hour monitor ekg
etc day substituting tekturna diovan palpitation
161291  would second month junel birth control year changed due spotting
increased mood swing previous birth control since switch shorter period day
gained major weight increased appetite switched regular exercise routine still
managed drop extra lb
```

[50]:
```
tfidf_vectorizer = TfidfVectorizer(stop_words='english', max_df=0.8)
tfidf_train = tfidf_vectorizer.fit_transform(X_train)
tfidf_test = tfidf_vectorizer.transform(X_test)

pass_tf = PassiveAggressiveClassifier()
pass_tf.fit(tfidf_train, y_train)
pred = pass_tf.predict(tfidf_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred, labels=['Birth Control',␣
 ↪'Depression','Diabetes, Type 2','High Blood Pressure'])
plot_confusion_matrix(cm, classes=['Birth Control', 'Depression','Diabetes,␣
 ↪Type 2','High Blood Pressure'])
```

```
accuracy:   0.982
Confusion matrix, without normalization
```



[51]:
```
text =["I have only been on Tekturna for 9 days. The effect was immediate. I am␣
 ↪also on a calcium channel blocker (Tiazac) and hydrochlorothiazide. I was␣
 ↪put on Tekturna because of palpitations experienced with Diovan (ugly drug␣
 ↪in my opinion, same company produces both however). The palpitations were␣
 ↪pretty bad on Diovan, 24 hour monitor by EKG etc. After a few days of␣
 ↪substituting Tekturna for Diovan, there are no more palpitations."]
test = tfidf_vectorizer.transform(text)
pred1=pass_tf.predict(test)[0]
pred1
```

[51]:
```
'High Blood Pressure'
```

[52]:
```
text =["This is the third med I&#039;ve tried for anxiety and mild depression.␣
 ↪Been on it for a week and I hate it so much. I am so dizzy, I have major␣
 ↪diarrhea and feel worse than I started. Contacting my doc in the am and␣
 ↪changing asap."]
test = tfidf_vectorizer.transform(text)
pred1=pass_tf.predict(test)[0]
pred1
```

[52]:
```
'Depression'
```

[53]:
```
text =["I just got diagnosed with type 2. My doctor prescribed Invokana and␣
 ↪metformin from the beginning. My sugars went down to normal by the second␣
 ↪week. I am losing so much weight. No side effects yet. Miracle medicine for␣
 ↪me"]
test = tfidf_vectorizer.transform(text)
pred1=pass_tf.predict(test)[0]
pred1
```

[53]:
```
'Diabetes, Type 2'
```

The model predicts the corresponding medical condition for each sample text based on the TF-IDF features.

[ ]:
```

```