Sentiment Analysis and Drug Review Insights

December 6, 2024

Sentiment Analysis and Drug Review Insights: A Data-Driven Approach to Understanding User Feedback

INTRODUCTION

Welcome to the Sentiment Analysis and Drug Review Insights project! In this project, we explore thousands of user reviews to better understand public sentiment regarding various drugs. By leveraging natural language processing (NLP) techniques and machine learning models, we aim to analyze text data, classify sentiments, and extract key insights from these reviews.

What to Expect in This Project: - Data Exploration & Preprocessing: We begin by cleaning the data, which includes removing unnecessary characters, handling missing values, and preparing the text for analysis (e.g., lemmatization and stopword removal).

- Sentiment Analysis: Using tools like TextBlob, we calculate sentiment scores for each review and categorize them into Positive, Negative, or Neutral sentiments.
- Exploratory Data Analysis (EDA): We visualize patterns within the drug reviews, such as sentiment distribution, top-reviewed drugs, and rating trends, to understand the key factors influencing user opinions.
- Machine Learning Models: Several machine learning models (Logistic Regression, Decision Tree, Random Forest) are built and evaluated to predict the sentiment of a review based on its text.
- Topic Modeling: With Latent Dirichlet Allocation (LDA), we identify the most common topics discussed in the reviews, including side effects, drug effectiveness, and overall user experiences.

Why This Matters: Understanding user feedback on drugs is critical for healthcare providers, pharmaceutical companies, and patients. This project helps uncover common themes and opinions, providing actionable insights into drug performance, potential side effects, and overall satisfaction. The goal is to bridge the gap between patient experiences and the insights that healthcare professionals need to improve drug outcomes.

[4]:

```python
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
%matplotlib inline
import matplotlib
import seaborn as sns
import warnings
```

```python
warnings.filterwarnings("ignore")
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import CountVectorizer
from collections import Counter
from matplotlib import style
style.use('ggplot')
from nltk.corpus import stopwords
import re
from textblob import TextBlob
from nltk.tokenize import word_tokenize
stop_words = set(stopwords.words('english'))
from wordcloud import WordCloud
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold, cross_val_score
import warnings
from sklearn.svm import LinearSVC
from sklearn import metrics
from sklearn.metrics import RocCurveDisplay
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
```

[2]:

```python
from sklearn.metrics import (
    accuracy_score,
    roc_auc_score,
    recall_score,
    precision_score,
    confusion_matrix,
    f1_score,
    precision_recall_curve,
    roc_curve,
    RocCurveDisplay,
    PrecisionRecallDisplay
)
```

Loading the Dataset

[20]:

```python
df = pd.read_csv(r'C:\Users\PC\Desktop\Datasets\drug_review_test.csv')
```

[23]:

```python
#displaying the first five rows of the DataFrame
df.head()
```

[23]:

```
   Unnamed: 0  patient_id      drugName                     condition  \
0           0      163740   Mirtazapine                    depression
1           1      206473    Mesalamine  crohn's disease, maintenance
```

```
2           2       39293         Contrave                   weight loss
3           3       97768  Cyclafem 1 / 35                 birth control
4           4      208087          Zyclara                     keratosis

                                              review  rating  \
0  "i've tried a few antidepressants over the yea…    10.0
1  "my son has crohn's disease and has done very …     8.0
2  "contrave combines drugs that were used for al…     9.0
3  "i have been on this birth control for one cyc…     9.0
4  "4 days in on first 2 weeks.  using on arms an…     4.0

                date  usefulCount  review_length
0  February 28, 2012           22             68
1       May 17, 2009           17             48
2      March 5, 2017           35            143
3   October 22, 2015            4            149
4       July 3, 2014           13             60
```

Exploratory Data Analysis

[50]:

```python
# Counting the frequency of each condition in the 'condition' column
df['condition'].value_counts()
```

[50]:

```
condition
birth control                                                     9257
depression                                                        2761
acne                                                              1724
pain                                                              1590
anxiety                                                           1589
bipolar disorde                                                   1151
weight loss                                                       1139
obesity                                                           1075
insomnia                                                          1004
adhd                                                               978
emergency contraception                                            783
vaginal yeast infection                                            729
diabetes, type 2                                                   687
high blood pressure                                                607
abnormal uterine bleeding                                          607
bowel preparation                                                  603
smoking cessation                                                  594
anxiety and stress                                                 515
migraine                                                           483
ibromyalgia                                                        476
major depressive disorde                                           466
constipation                                                       462
migraine prevention                                                413
panic disorde                                                      406
```

chronic pain	388
urinary tract infection	366
generalized anxiety disorde	353
opiate dependence	340
osteoarthritis	309
erectile dysfunction	298
irritable bowel syndrome	281
muscle spasm	279
allergic rhinitis	264
rheumatoid arthritis	264
bacterial infection	257
sinusitis	239
hyperhidrosis	227
hepatitis c	213
restless legs syndrome	208
nausea/vomiting	198
overactive bladde	195
cough	195
multiple sclerosis	190
gerd	184
endometriosis	181
hypogonadism, male	177
psoriasis	164
hiv infection	161
constipation, chronic	160
obsessive compulsive disorde	153
seizures	144
schizophrenia	144
rosacea	143
high cholesterol	143
bacterial vaginitis	135
narcolepsy	126
social anxiety disorde	124
benign prostatic hyperplasia	124
bronchitis	124
onychomycosis, toenail	122
back pain	122
asthma, maintenance	120
menstrual disorders	118
alcohol dependence	118
herpes simplex	116
not listed / othe	112
headache	110
influenza	107
bladder infection	107
underactive thyroid	103
postmenopausal symptoms	101

post traumatic stress disorde	101
epilepsy	99
schizoaffective disorde	97
cold sores	93
psoriatic arthritis	89
premenstrual dysphoric disorde	89
plaque psoriasis	88
opiate withdrawal	88
inflammatory conditions	86
breast cance	86
chlamydia infection	84
cough and nasal congestion	84
neuropathic pain	78
anesthesia	78
nasal congestion	77
pneumonia	74
helicobacter pylori infection	74
polycystic ovary syndrome	73
constipation, drug induced	70
mance anxiety	68
urticaria	68
crohn's disease, maintenance	66
osteoporosis	66
atrial fibrillation	64
motion sickness	61
ulcerative colitis	60
asthma	58
alcohol withdrawal	56
hot flashes	56
ovarian cysts	55
hashimoto's disease	55
copd, maintenance	55
seizure prevention	55
dry eye disease	52
sedation	52
ankylosing spondylitis	52
atrophic vaginitis	50
peripheral neuropathy	49
hirsutism	47
urinary incontinence	44
gout, acute	43
diabetic peripheral neuropathy	43
atopic dermatitis	43
skin or soft tissue infection	42
bacterial skin infection	42
trigeminal neuralgia	41
prostate cance	41

period pain	41
diarrhea	41
crohn's disease	41
otitis media	40
diverticulitis	40
herpes simplex, suppression	40
borderline personality disorde	39
allergies	39
human papilloma virus	39
ulcerative colitis, active	39
upper respiratory tract infection	39
urinary tract stones	38
nausea/vomiting of pregnancy	37
non-small cell lung cance	37
eczema	37
cluster headaches	37
prevention of thromboembolism in atrial fibrillation	37
copd	37
arrhythmia	37
keratosis	37
postherpetic neuralgia	36
interstitial cystitis	35
obstructive sleep apnea/hypopnea syndrome	34
premature lab	33
strep throat	33
edema	32
diabetes, type 1	31
systemic lupus erythematosus	31
parkinson's disease	31
renal cell carcinoma	29
supraventricular tachycardia	28
head lice	28
binge eating disorde	28
kidney infections	27
postpartum depression	27
alopecia	27
pseudotumor cerebri	27
1</span> users found this comment helpful.	27
hypothyroidism, after thyroid removal	26
melasma	26
anorexia	26
autism	26
benign essential trem	25
moterol)	25
gout	24
sciatica	24
chronic myelogenous leukemia	24

angina	24
chronic fatigue syndrome	24
asthma, acute	24
rhinitis	23
breast cancer, metastatic	23
dermatitis	23
amenorrhea	23
allergic reactions	22
neuralgia	22
0</span> users found this comment helpful.	22
shift work sleep disorde	21
alzheimer's disease	21
dental abscess	21
keratoconjunctivitis sicca	20
dysuria	20
tourette's syndrome	20
hypertriglyceridemia	20
deep vein thrombosis	20
cold symptoms	20
basal cell carcinoma	19
hypersomnia	19
4</span> users found this comment helpful.	18
prostatitis	18
sexual dysfunction, ssri induced	18
cervical dystonia	18
tendonitis	17
hemorrhoids	17
light sedation	17
dietary supplementation	17
opioid-induced constipation	17
tonsillitis/pharyngitis	17
erosive esophagitis	17
diarrhea, chronic	17
sjogren's syndrome	17
pulmonary hypertension	17
menorrhagia	16
emale infertility	16
muscle pain	16
skin and structure infection	16
vertig	16
prevention of bladder infection	16
constipation, acute	16
trichomoniasis	16
heart attack	16
3</span> users found this comment helpful.	16
heart failure	16
androgenetic alopecia	16

conjunctivitis, allergic	16
seasonal allergic conjunctivitis	15
atigue	15
dry skin	15
2</span> users found this comment helpful.	15
psychosis	15
high cholesterol, familial heterozygous	15
vitamin/mineral supplementation during pregnancy/lactation	14
pruritus	14
bulimia	14
pulmonary embolism	13
chronic idiopathic constipation	13
uterine fibroids	13
malaria prevention	13
6</span> users found this comment helpful.	13
barrett's esophagus	13
perimenopausal symptoms	12
min)	12
soft tissue sarcoma	12
sinus symptoms	12
5</span> users found this comment helpful.	12
hyperprolactinemia	12
burns, external	12
gastroparesis	12
breakthrough pain	11
neutropenia associated with chemotherapy	11
iron deficiency anemia	11
mania	11
postoperative pain	11
raynaud's syndrome	11
otitis externa	11
ulcerative colitis, maintenance	11
light anesthesia	11
myasthenia gravis	10
nausea/vomiting, chemotherapy induced	10
organ transplant, rejection prophylaxis	10
10</span> users found this comment helpful.	10
seborrheic dermatitis	10
7</span> users found this comment helpful.	10
indigestion	10
pneumococcal disease prophylaxis	10
glaucoma, open angle	10
lyme disease	9
acute coronary syndrome	9
eye redness	9
eye redness/itching	9
acial wrinkles	9

reflex sympathetic dystrophy syndrome	9
macular degeneration	9
dermatological disorders	9
paranoid disorde	9
methicillin-resistant staphylococcus aureus infection	8
hypoestrogenism	8
conjunctivitis, bacterial	8
mitral valve prolapse	8
temporomandibular joint disorde	8
agitated state	8
prevention of osteoporosis	8
meniere's disease	8
tinea versicol	8
tinea pedis	8
stomach ulce	8
actinic keratosis	8
clostridial infection	8
juvenile rheumatoid arthritis	7
min / sitagliptin)	7
glaucoma	7
diagnosis and investigation	7
surgical prophylaxis	7
periodic limb movement disorde	7
gouty arthritis	7
nocturnal leg cramps	7
ischemic stroke, prophylaxis	7
9</span> users found this comment helpful.	7
primary immunodeficiency syndrome	7
uveitis	7
idiopathic thrombocytopenic purpura	7
inflammatory bowel disease	7
vulvodynia	7
chronic lymphocytic leukemia	7
crohn's disease, acute	7
moterol / mometasone)	6
anemia	6
precocious puberty	6
giardiasis	6
night terrors	6
8</span> users found this comment helpful.	6
human papillomavirus prophylaxis	6
undifferentiated connective tissue disease	6
cutaneous candidiasis	6
vitamin d deficiency	6
gastroenteritis	6
herpes zoste	6
pseudobulbar affect	6

breast cancer, adjuvant	6
juvenile idiopathic arthritis	6
postural orthostatic tachycardia syndrome	6
postoperative ocular inflammation	6
angina pectoris prophylaxis	6
allergic urticaria	6
15</span> users found this comment helpful.	5
dermatologic lesion	5
premenstrual syndrome	5
pancreatic exocrine dysfunction	5
atrial flutte	5
left ventricular dysfunction	5
nausea/vomiting, postoperative	5
bronchospasm prophylaxis	5
swine flu	5
onychomycosis, fingernail	5
12</span> users found this comment helpful.	5
agitation	5
photoaging of the skin	5
intraocular hypertension	5
cystic fibrosis	5
non-hodgkin's lymphoma	5
noninfectious colitis	5
labor pain	5
ovulation induction	5
tinea cruris	5
insulin resistance syndrome	5
13</span> users found this comment helpful.	5
infection prophylaxis	5
17</span> users found this comment helpful.	5
tsh suppression	5
influenza prophylaxis	5
menopausal disorders	4
diabetic kidney disease	4
tinea corporis	4
nasal polyps	4
gingivitis	4
burning mouth syndrome	4
cyclic vomiting syndrome	4
gas	4
vitamin b12 deficiency	4
ge (amlodipine / valsartan)	4
pelvic inflammatory disease	4
auditory processing disorde	4
pediatric growth hormone deficiency	4
myelodysplastic syndrome	4
gastrointestinal decontamination	4

asperger syndrome	4
osteolytic bone lesions of multiple myeloma	4
bullous pemphigoid	4
diabetes insipidus	4
mucositis	4
nightmares	4
opioid overdose	4
pinworm infection (enterobius vermicularis)	4
xerostomia	4
skin disinfection, preoperative	4
hyperlipoproteinemia type iia, elevated ldl	4
breast cancer, prevention	4
14</span> users found this comment helpful.	4
autoimmune hepatitis	4
diaper rash	4
anemia associated with chronic renal failure	4
dyspareunia	4
thyroid cance	4
dystonia	4
spondyloarthritis	4
zen shoulde	3
lyme disease, neurologic	3
primary nocturnal enuresis	3
endometrial hyperplasia	3
sarcoidosis	3
peptic ulce	3
18</span> users found this comment helpful.	3
vitamin/mineral supplementation and deficiency	3
iritis	3
gastritis/duodenitis	3
impetig	3
mycobacterium avium-intracellulare, treatment	3
progesterone insufficiency	3
21</span> users found this comment helpful.	3
tuberculosis, latent	3
primary ovarian failure	3
oophorectomy	3
lipodystrophy	3
lactation augmentation	3
hypocalcemia	3
conjunctivitis	3
von willebrand's disease	3
skin cance	3
copd, acute	3
ventricular tachycardia	3
hyperthyroidism	3
39</span> users found this comment helpful.	3

```
anal itching                                                         3
anal fissure and fistula                                             3
hereditary angioedema                                                3
herbal supplementation                                               3
dercum's disease                                                     3
hyperphosphatemia of renal failure                                   3
bursitis                                                             3
lennox-gastaut syndrome                                              3
dysautonomia                                                         3
melanoma, metastatic                                                 3
reversal of opioid sedation                                          3
mountain sickness / altitude sickness                                3
pe                                                                   3
endometrial cance                                                    3
deep vein thrombosis prophylaxis after hip replacement surgery       3
11</span> users found this comment helpful.                          3
intermittent claudication                                            3
ulcerative proctitis                                                 3
insomnia, stimulant-associated                                       3
anemia, sickle cell                                                  3
oral thrush                                                          3
tinnitus                                                             3
chronic pancreatitis                                                 3
hepatic encephalopathy                                               3
lic acid deficiency                                                  3
malaria                                                              3
deep vein thrombosis, prophylaxis                                    3
benzodiazepine withdrawal                                            3
aphthous ulce                                                        3
multiple myeloma                                                     3
hepatitis b                                                          3
eyelash hypotrichosis                                                3
condylomata acuminata                                                3
perioral dermatitis                                                  3
mild cognitive impairment                                            3
duodenal ulce                                                        3
hypoparathyroidism                                                   2
diabetic macular edema                                               2
bleeding disorde                                                     2
adrenocortical insufficiency                                         2
uterine bleeding                                                     2
hypercalcemia of malignancy                                          2
32</span> users found this comment helpful.                          2
new daily persistent headache                                        2
ophthalmic surgery                                                   2
amebiasis                                                            2
bone infection                                                       2
```

nsaid-induced gastric ulce	2
23</span> users found this comment helpful.	2
periodontitis	2
deep vein thrombosis, recurrent event	2
hypomagnesemia	2
extrapyramidal reaction	2
hyperlipoproteinemia	2
lichen sclerosus	2
atrophic urethritis	2
systemic sclerosis	2
seasonal affective disorde	2
intraabdominal infection	2
occupational exposure	2
gastrointestinal stromal tum	2
deep neck infection	2
cataplexy	2
women (oxybutynin)	2
mulation) (phenylephrine)	2
osteolytic bone metastases of solid tumors	2
hiccups	2
51</span> users found this comment helpful.	2
gallbladder disease	2
neck pain	2
mantle cell lymphoma	2
dandruff	2
deep vein thrombosis, first event	2
systemic mastocytosis	2
prevention of dental caries	2
min / saxagliptin)	2
oligospermia	2
eve	2
breast cancer, palliative	2
hidradenitis suppurativa	2
diarrhea, acute	2
persistent depressive disorde	2
computed tomography	2
pharyngitis	2
pupillary dilation	2
occipital neuralgia	2
abortion	2
traveler's diarrhea	2
schnitzler syndrome	2
lactose intolerance	2
expectoration	2
herpes simplex, mucocutaneous/immunocompetent host	2
hypotension	2
lymphoma	2

```
secondary hyperparathyroidism                                        2
glioblastoma multiforme                                              2
dumping syndrome                                                     2
opiate adjunct                                                       2
platelet aggregation inhibition                                      2
lymphocytic colitis                                                  2
gastrointestinal hemorrhage                                          2
25</span> users found this comment helpful.                          2
corneal refractive surgery                                           2
percutaneous coronary intervention                                   2
acute lymphoblastic leukemia                                         2
ear wax impaction                                                    2
polymyositis/dermatomyositis                                         2
still's disease                                                      2
thromboembolic stroke prophylaxis                                    2
20</span> users found this comment helpful.                          2
skin rash                                                            2
22</span> users found this comment helpful.                          2
m pain disorde                                                       2
premature ventricular depolarizations                                2
ventricular fibrillation                                             2
herpes simplex dendritic keratitis                                   1
27</span> users found this comment helpful.                          1
28</span> users found this comment helpful.                          1
lichen planus                                                        1
amilial cold autoinflammatory syndrome                               1
herpes simplex, mucocutaneous/immunocompromised host                 1
hyperkalemia                                                         1
aids related wasting                                                 1
100</span> users found this comment helpful.                         1
behcet's disease                                                     1
29</span> users found this comment helpful.                          1
strongyloidiasis                                                     1
wegener's granulomatosis                                             1
cal segmental glomerulosclerosis                                     1
glioblastoma multi                                                   1
blepharitis                                                          1
ovarian cance                                                        1
anaplastic astrocytoma                                               1
strabismus                                                           1
ibrocystic breast disease                                            1
premature ejaculation                                                1
epididymitis, sexually transmitted                                   1
squamous cell carcinoma                                              1
macular edema                                                        1
55</span> users found this comment helpful.                          1
salmonella gastroenteritis                                           1
```

acute nonlymphocytic leukemia	1
topical disinfection	1
24</span> users found this comment helpful.	1
adult human growth hormone deficiency	1
105</span> users found this comment helpful.	1
stomach cance	1
cachexia	1
warts	1
local anesthesia	1
yellow fever prophylaxis	1
cerebral edema	1
chronic myofascial pain	1
nonoccupational exposure	1
submental fullness	1
leukemia	1
tic (mycophenolic acid)	1
status epilepticus	1
me	1
primary hyperaldosteronism	1
eosinophilic esophagitis	1
performance anxiety	1
small bowel or pancreatic fistula	1
lyme disease, erythema chronicum migrans	1
steroid responsive inflammatory conditions	1
135</span> users found this comment helpful.	1
head injury	1
urinary retention	1
myelofibrosis	1
cance	1
38</span> users found this comment helpful.	1
renal tubular acidosis	1
pain/feve	1
pertussis prophylaxis	1
head and neck cance	1
uveitis, posteri	1
sunburn	1
nephrotic syndrome	1
ehrlichiosis	1
aplastic anemia	1
34</span> users found this comment helpful.	1
radionuclide myocardial perfusion study	1
paget's disease	1
polycythemia vera	1
myotonia congenita	1
coccidioidomycosis	1
cerebral spasticity	1
zollinger-ellison syndrome	1

liver magnetic resonance imaging	1
stress ulcer prophylaxis	1
subarachnoid hemorrhage	1
37</span> users found this comment helpful.	1
spondylolisthesis	1
hypokalemic periodic paralysis	1
jet lag	1
oral and dental conditions	1
hypoactive sexual desire disorde	1
thrombocythemia	1
dissociative identity disorde	1
somat	1
mixed connective tissue disease	1
transient ischemic attack	1
pancreatic cance	1
ventricular arrhythmia	1
enterocolitis	1
hemorrhoids (pramoxine / zinc oxide)	1
pityriasis rubra pilaris	1
hepatocellular carcinoma	1
varicella-zoste	1
sore throat	1
colorectal cance	1
high cholesterol, familial homozygous	1
autoimmune hemolytic anemia	1
body dysmorphic disorde	1
neutropenia	1
esophageal variceal hemorrhage prophylaxis	1
hyperekplexia	1
post-cholecystectomy diarrhea	1
gastric ulcer maintenance treatment	1
cyclothymic disorde	1
dermatomyositis	1
cmv prophylaxis	1
pulmonary edema	1
ischemic stroke	1
labor induction	1
streptococcal infection	1
cutaneous t-cell lymphoma	1
70</span> users found this comment helpful.	1
acetaminophen overdose	1
mononucleosis	1
dementia	1
prevention of cardiovascular disease	1
peritonitis	1
pertussis	1
upper limb spasticity	1

```
tympanostomy tube placement surgery                                  1
chronic inflammatory demyelinating polyradiculoneuropathy            1
nausea (phosphorated carbohydrate solution)                          1
schilling test                                                       1
31</span> users found this comment helpful.                          1
hypertensive emergency                                               1
trichotillomania                                                     1
prevention of hypokalemia                                            1
pulmonary embolism, first event                                      1
gonococcal infection, uncomplicated                                  1
mycoplasma pneumonia                                                 1
bacterial endocarditis prevention                                    1
dupuytren's contracture                                              1
gender dysphoria                                                     1
dietary fiber supplementation                                        1
83</span> users found this comment helpful.                          1
myeloproliferative disorders                                         1
Name: count, dtype: int64
```

[25]:

```
# Group the data by drugName and condition, and count the number of reviews for␣
  ↪each
pd.set_option('display.max_rows', None)

drug_condition_counts = df.groupby(['drugName', 'condition']).size().
  ↪reset_index(name='review_count')

# Display the first few rows of the grouped data
drug_condition_counts.head(10)
```

[25]:

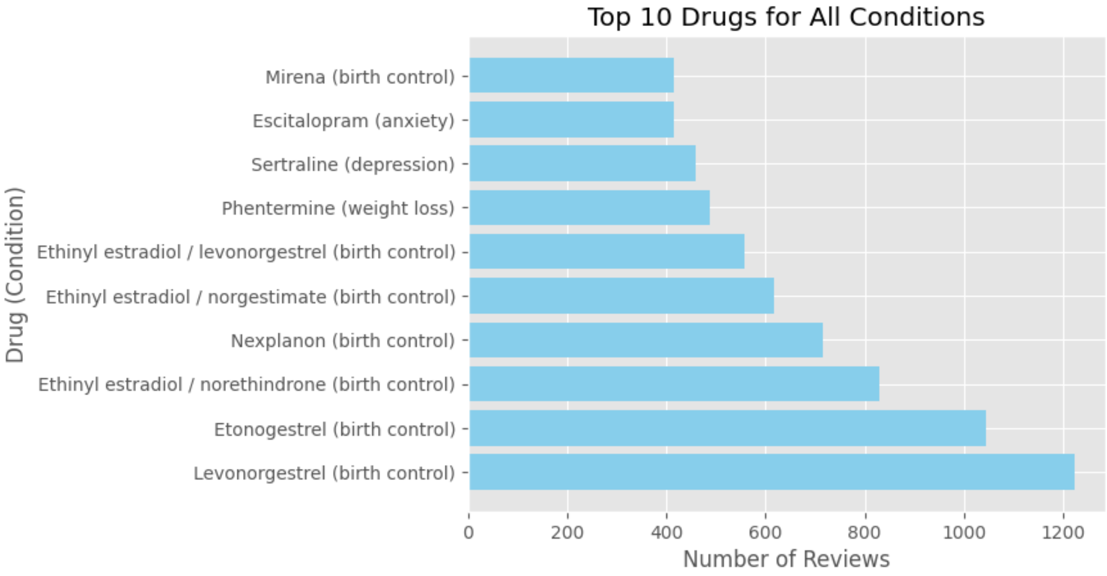
```
                               drugName  \
0                            A / B Otic
1  Abacavir / dolutegravir / lamivudine
2                  Abacavir / lamivudine
3                              Abatacept
4                                Abilify
5                                Abilify
6                                Abilify
7                                Abilify
8                                Abilify
9                                Abilify

                                     condition  review_count
0                                 otitis media             1
1                                hiv infection            16
2                                hiv infection             2
3                         rheumatoid arthritis             4
4  20</span> users found this comment helpful.             1
```

```
5                              agitated state             1
6                                      autism             2
7                            bipolar disorde             39
8                                  depression            39
9                     major depressive disorde           14
```

Top 10 Drugs by Review Count Across Conditions

[26]:

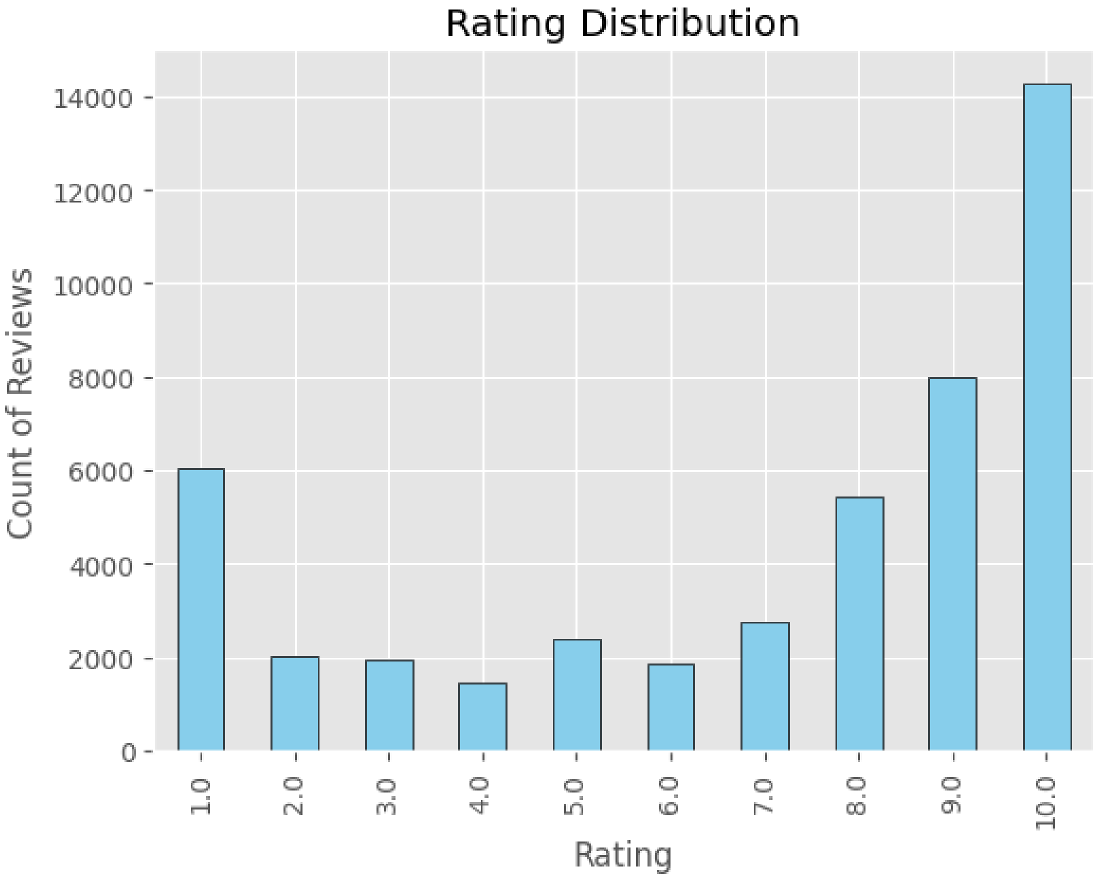
```python
# Group the data by drugName and condition, count the number of reviews, and unstack
drug_condition = df.groupby(['drugName', 'condition']).size().unstack().fillna(0)

# Calculate total number of reviews for each drug (across all conditions)
drug_condition_sums = drug_condition.sum(axis=1)

# Get the top 10 drugs by total number of reviews
top_10_drugs = drug_condition_sums.nlargest(10)

# Create a DataFrame with the drug name and corresponding condition (add the condition back for understanding)
# We take the condition with the maximum reviews for each drug
top_10_drug_conditions = pd.DataFrame({
    'drug': top_10_drugs.index,
    'condition': drug_condition.idxmax(axis=1).loc[top_10_drugs.index],  # Condition with most reviews for each drug
    'review_count': top_10_drugs.values
})

# Create the bar plot with drug names and corresponding conditions
plt.barh(top_10_drug_conditions['drug'] + ' (' + top_10_drug_conditions['condition'] + ')',
         top_10_drug_conditions['review_count'], color='skyblue')

# Set the title and labels
plt.title('Top 10 Drugs for All Conditions')
plt.xlabel('Number of Reviews')
plt.ylabel('Drug (Condition)')
plt.show()
```



Observation - This chart shows that birth control, mental health treatments, and weight loss medications are the most reviewed.

Rating Distribution of Drug Reviews

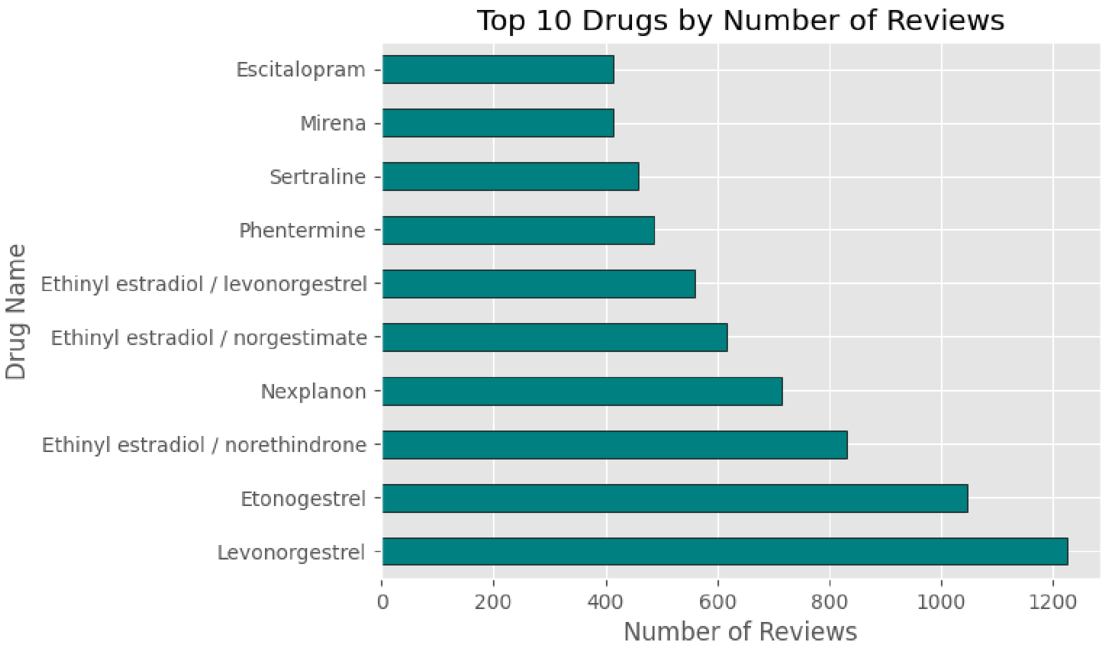
```
[9]: # Bar chart for rating distribution
df['rating'].value_counts().sort_index().plot(kind='bar', color='skyblue',␣
 ↪edgecolor='black')
plt.title('Rating Distribution')
plt.xlabel('Rating')
plt.ylabel('Count of Reviews')
plt.show()
```



Observation - The bar chart shows the number of reviews for different ratings, ranging from 1.0 to 10.0. The highest number of reviews is for the rating of 10.0, with nearly 14,000 reviews, indicating a strong preference for this rating among users. The next highest is for the rating of 9.0, while ratings from 1.0 to 8.0 have significantly fewer reviews, all under approximately half the count of the 9.0 rating or less. This distribution suggests that the majority of users have had very positive experiences, as evidenced by the high number of top ratings.

Top 10 Drugs by Number of Reviews

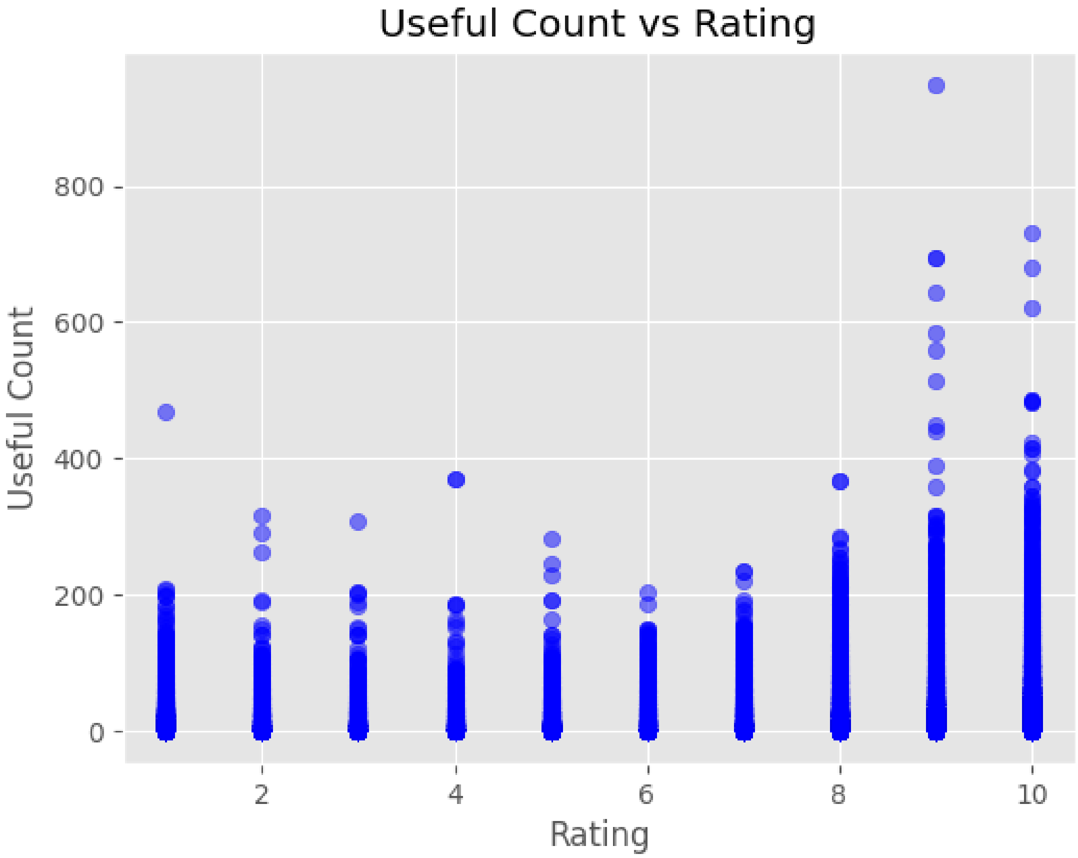
[11]:
```
# Top 10 drugs with the most reviews
top_10_drugs = df['drugName'].value_counts().nlargest(10)

# Plot the horizontal bar chart
top_10_drugs.plot(kind='barh', color='teal', edgecolor='black')
plt.title('Top 10 Drugs by Number of Reviews')
plt.xlabel('Number of Reviews')
plt.ylabel('Drug Name')
plt.show()
```



Observation - The bar chart titled “Top 10 Drugs by Number of Reviews” shows the most reviewed medications, with Escitalopram and Mirena leading the list. The chart highlights the popularity and user engagement with these drugs, indicating their widespread use and the significant number of reviews they have received.

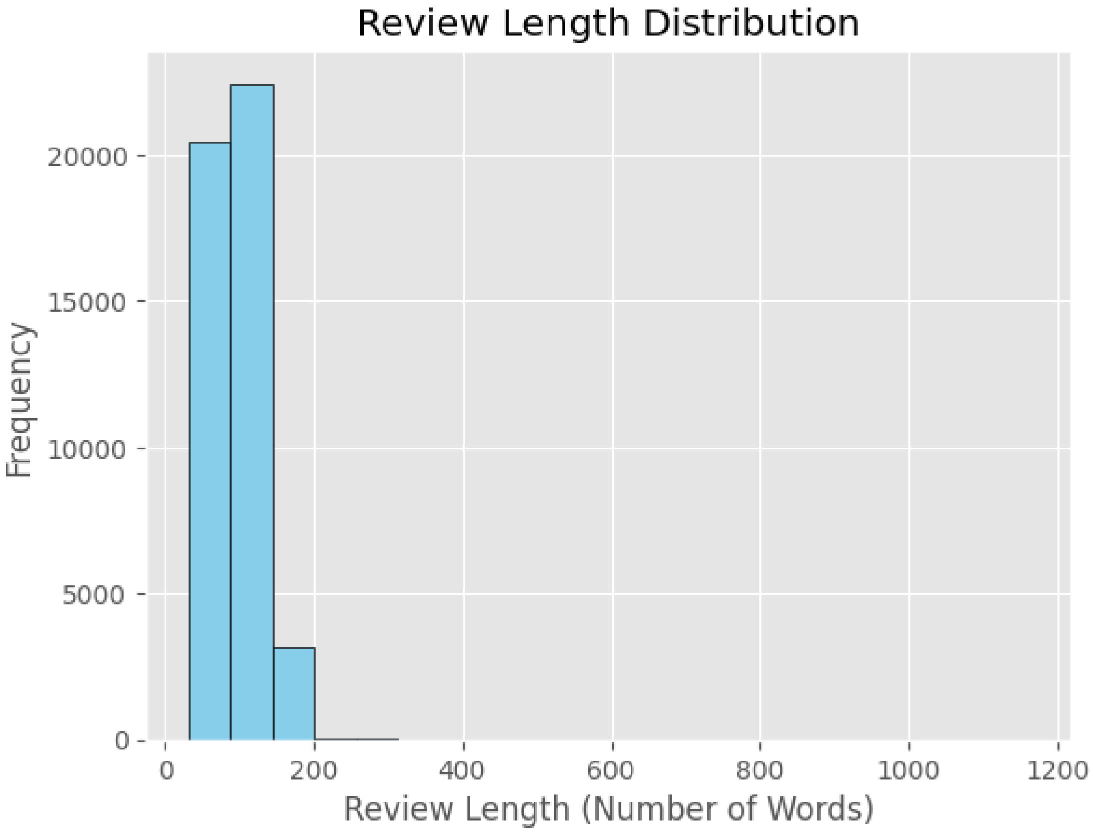
Scatter Plot of Useful Count vs Rating

[13]:
```
# Scatter plot of usefulCount vs rating
plt.scatter(df['rating'], df['usefulCount'], alpha=0.5, color='blue')
plt.title('Useful Count vs Rating')
plt.xlabel('Rating')
plt.ylabel('Useful Count')
plt.grid(True)
plt.show()
```



Review Length Distribution

```
[17]: # Plotting the distribution of review lengths (number of words in each review)
df['review_length'].plot(kind='hist', bins=20, color='skyblue',␣
 ↪edgecolor='black')
plt.title('Review Length Distribution')
plt.xlabel('Review Length (Number of Words)')
plt.ylabel('Frequency')
plt.show()
```



Observation - This histogram shows that most reviews are short, with fewer reviews as the word count increases

Sentiment Analysis of Reviews

[27]:
```
def get_sentiment(text):
    blob = TextBlob(text)
    return blob.sentiment.polarity  # Returns a score between -1 (negative) and␣
 ↪1 (positive)

df['sentiment_score'] = df['review'].apply(get_sentiment)

# Classify into positive, negative, and neutral
df['sentiment'] = df['sentiment_score'].apply(lambda x: 'positive' if x > 0␣
 ↪else ('negative' if x < 0 else 'neutral'))
```

[28]:
```
df.head()
```

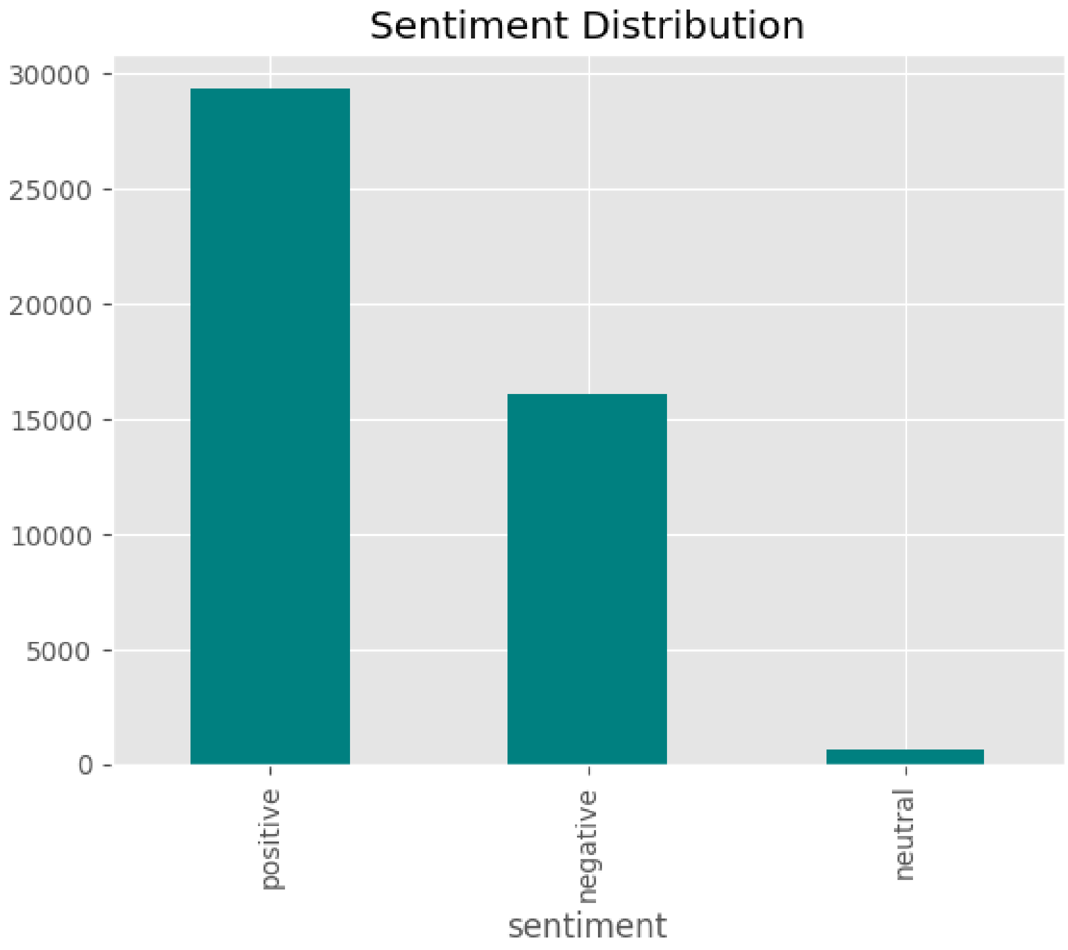
[28]:
```
   Unnamed: 0  patient_id        drugName                     condition  \
0           0      163740      Mirtazapine                    depression
1           1      206473       Mesalamine  crohn's disease, maintenance
```

```
2            2       39293         Contrave                    weight loss
3            3       97768   Cyclafem 1 / 35                birth control
4            4      208087           Zyclara                   keratosis

                                            review  rating  \
0  "i've tried a few antidepressants over the yea…    10.0
1  "my son has crohn's disease and has done very …     8.0
2  "contrave combines drugs that were used for al…     9.0
3  "i have been on this birth control for one cyc…     9.0
4  "4 days in on first 2 weeks.  using on arms an…     4.0

                date  usefulCount  review_length  sentiment_score sentiment
0  February 28, 2012           22             68         0.000000   neutral
1       May 17, 2009           17             48         0.566667  positive
2      March 5, 2017           35            143         0.139063  positive
3   October 22, 2015            4            149         0.260926  positive
4       July 3, 2014           13             60         0.341667  positive
```

Sentiment Distribution of Reviews

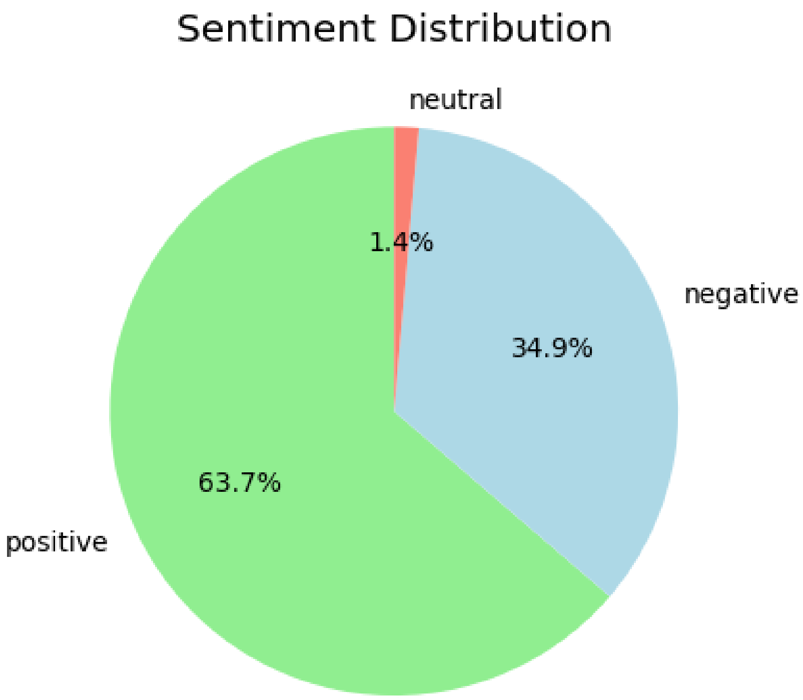
[31]:

```python
import matplotlib.pyplot as plt
df['sentiment'].value_counts().plot(kind='bar', color='teal', title='Sentiment␣
 ↪Distribution')
plt.show()
```



Observation - This bar chart shows that positive sentiments are the most common, followed by negative and neutral sentiments.

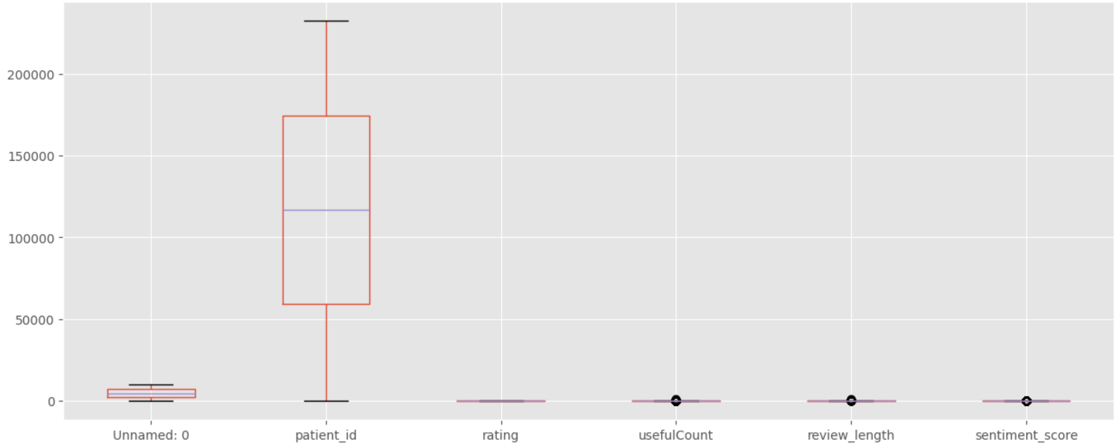
[32]:
```
sentiment_counts = df['sentiment'].value_counts()

# Plot a pie chart
sentiment_counts.plot(kind='pie', autopct='%1.1f%%', startangle=90,␣
 ↪colors=['lightgreen', 'lightblue', 'salmon'])
plt.title('Sentiment Distribution')
plt.ylabel('')
plt.show()
```



Observation - This pie chart shows that 63.7% of sentiments are positive, 34.9% are negative, and 1.4% are neutral.

```
[36]: df.boxplot(figsize=(15,6))
```

```
[36]: <Axes: >
```



Observation - The boxplot shows significant variability in patient_id but limited variation in features like usefulCount, review_length, and sentiment_score, which are heavily skewed towards lower values. This suggests most reviews are short and not frequently marked as useful, with few outliers.

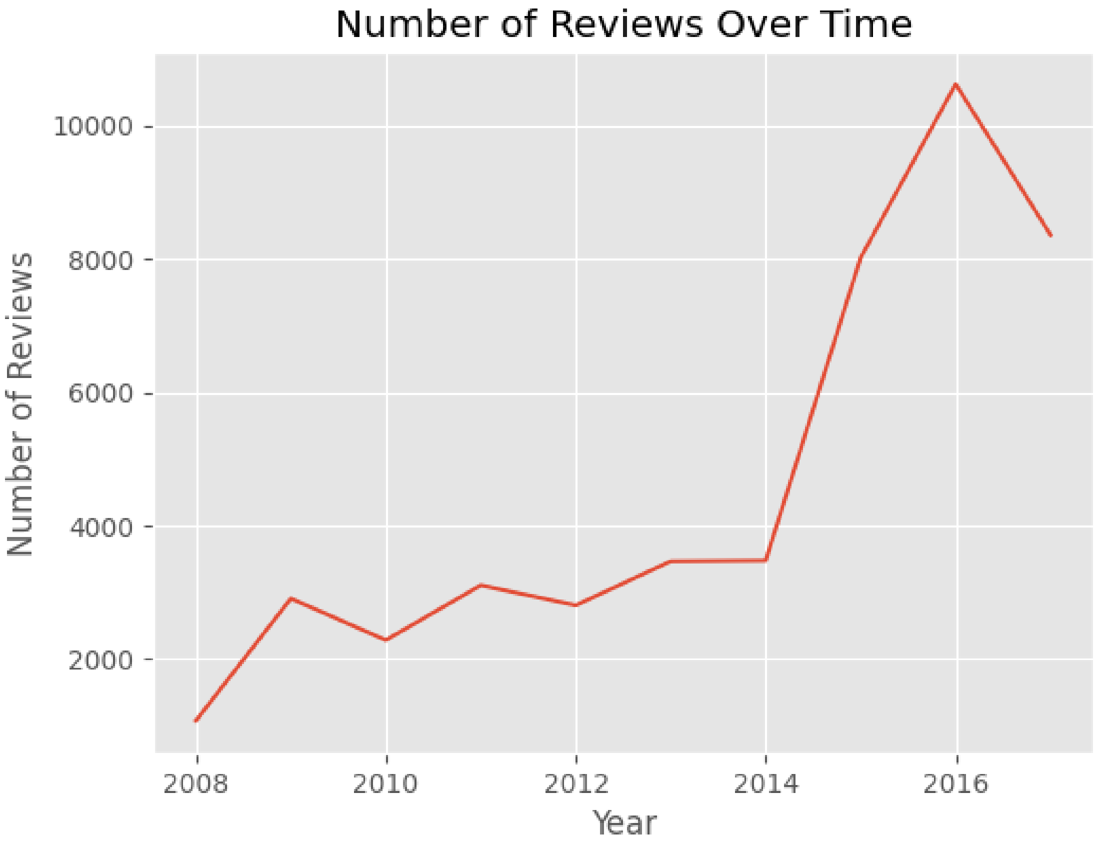
Review trends overtime

[38]:

```
df['date'] = pd.to_datetime(df['date'])
```

[39]:

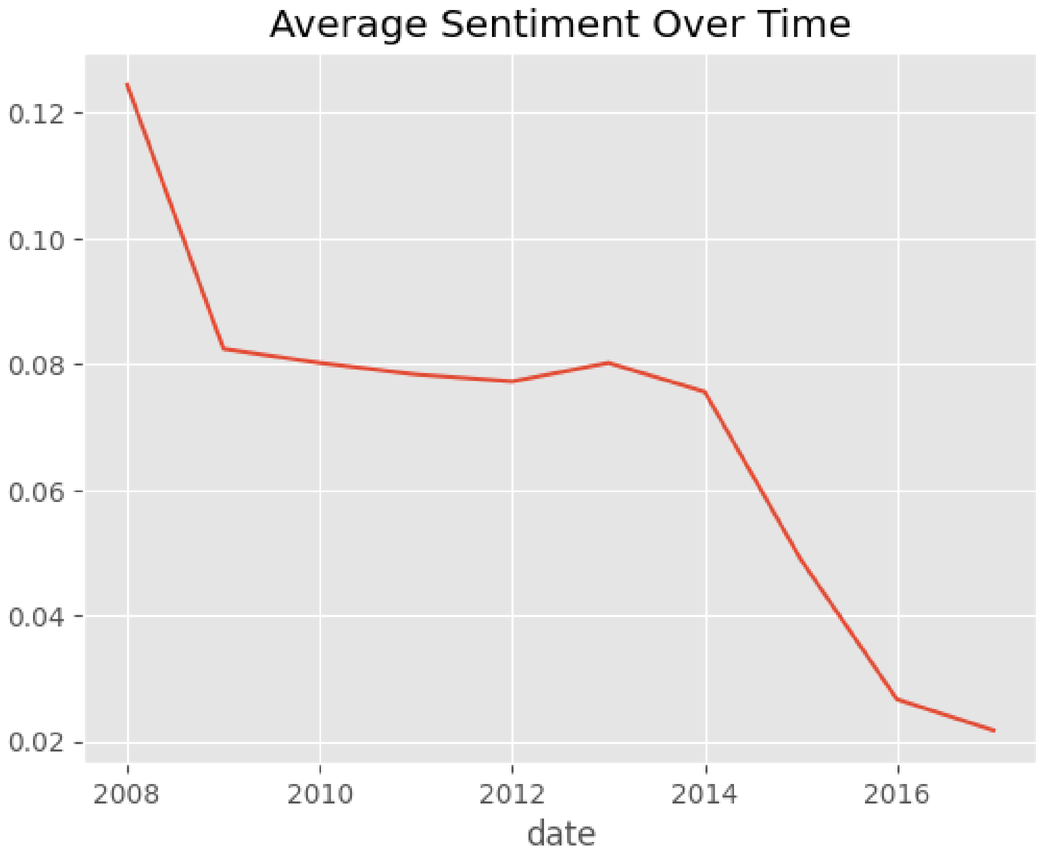
```
reviews_per_year = df.groupby(df['date'].dt.year).size()
reviews_per_year.plot(kind='line', title='Number of Reviews Over Time')
plt.xlabel('Year')
plt.ylabel('Number of Reviews')
plt.show()
```



Observation - This line graph shows a significant increase in the number of reviews over time, peaking sharply around 2016.

Sentiment Overtime

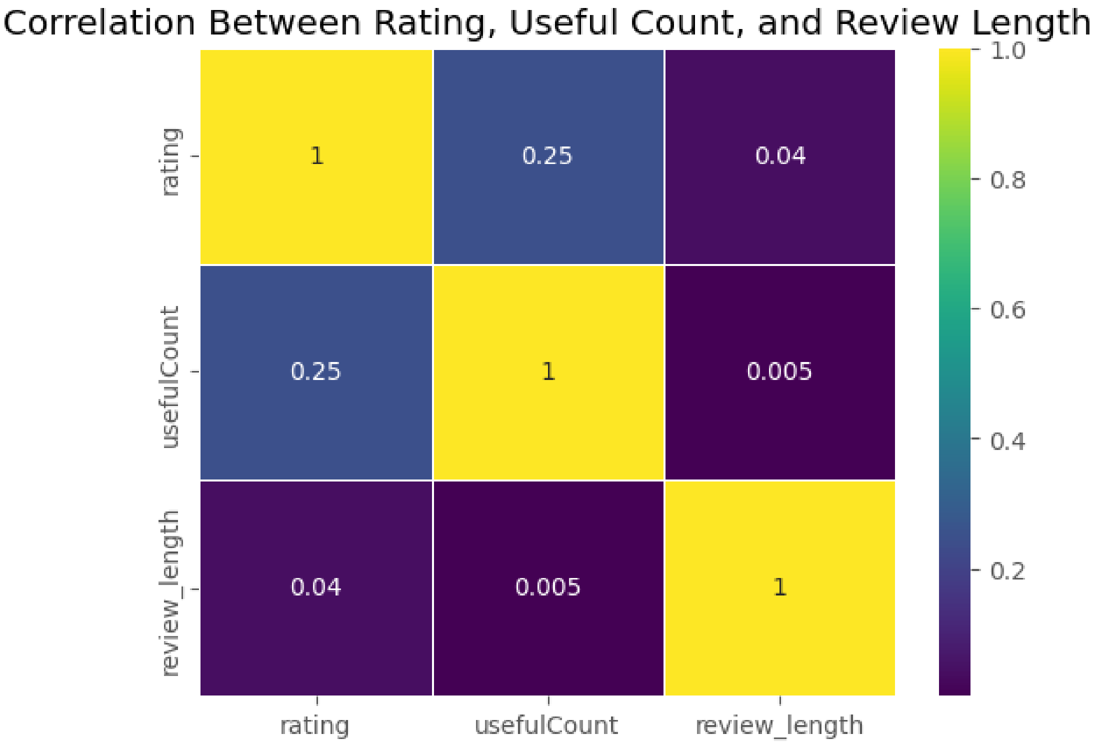
[40]:
```
sentiment_per_year = df.groupby(df['date'].dt.year)['sentiment_score'].mean()
sentiment_per_year.plot(kind='line', title='Average Sentiment Over Time')
plt.show()
```



Observation - The Average Sentiment Over Time illustrates a downward trend in sentiment from 2008 to 2016. Starting just above 0.1 in 2008, the sentiment steadily declines, reaching just above 0.02 by 2016. This suggests a significant negative shift in opinions or attitudes over the eight-year period.

Heatmap: Correlation Between Numerical Features

[46]:
```
correlation_matrix = df[['rating', 'usefulCount', 'review_length']].corr()

# Plot a heatmap
sns.heatmap(correlation_matrix, annot=True, cmap='viridis', linewidths=0.5)
plt.title('Correlation Between Rating, Useful Count, and Review Length')
plt.show()
```



[48]:
```
# Check for missing values in each column
df.isna().sum()
```

[48]:
```
Unnamed: 0         0
patient_id         0
drugName           0
condition          0
review             0
rating             0
date               0
usefulCount        0
review_length      0
sentiment_score    0
sentiment          0
dtype: int64
```

Text Preprocessing: Removing Special Characters, Lowercasing, Stopwords, Lemmatization, and Vectorization

[52]:
```
# Initializing the lemmatizer
lemmatizer = WordNetLemmatizer()
stop_words = set(stopwords.words('english'))
```

```
# Function to clean a single review
def clean_text(text):
    # Lowercase the text
    text = text.lower()

    # Remove punctuation and non-alphabetical characters
    text = re.sub(r'[^a-z\s]', '', text)

    # Tokenize and remove stopwords, then lemmatize
    tokens = text.split()
    tokens = [lemmatizer.lemmatize(word) for word in tokens if word not in␣
 ↪stop_words]

    # Join tokens back into a single string
    return ' '.join(tokens)

#cleaning the column containing the reviews
df['cleaned_reviews'] = df['review'].apply(clean_text)

# Now, 'cleaned_reviews' contains the preprocessed reviews
cleaned_reviews = df['cleaned_reviews'].tolist()

# Apply CountVectorizer on the cleaned reviews
vectorizer = CountVectorizer(max_df=0.95, min_df=2, stop_words='english')
review_matrix = vectorizer.fit_transform(cleaned_reviews)

print(review_matrix.shape)
```

```
(46108, 18205)
```

Modeling with Latent Dirichlet Allocation (LDA)

```
[54]: from sklearn.decomposition import LatentDirichletAllocation

      lda = LatentDirichletAllocation(n_components=5, random_state=42)   #␣
       ↪n_components is the number of topics
      lda.fit(review_matrix)
```

```
[54]: LatentDirichletAllocation(n_components=5, random_state=42)
```

```
[56]: #Displaying LDA Topics with Top Words
      def display_topics(model, feature_names, no_top_words):
          for topic_idx, topic in enumerate(model.components_):
              print(f"Topic {topic_idx}:")
              print([feature_names[i] for i in topic.argsort()[:-no_top_words - 1:
       ↪-1]])

      display_topics(lda, vectorizer.get_feature_names_out(), 10)
```

```
Topic 0:
['mg', 'day', 'anxiety', 'effect', 'feel', 'taking', 'year', 'week', 'im',
'started']
Topic 1:
['day', 'effect', 'time', 'year', 'taking', 'hour', 'blood', 'took', 'doctor',
'like']
Topic 2:
['day', 'week', 'skin', 'month', 'acne', 'got', 'period', 'im', 'year', 'like']
Topic 3:
['pain', 'day', 'year', 'taking', 'work', 'time', 'mg', 'doctor', 'effect',
'medication']
Topic 4:
['month', 'period', 'pill', 'ive', 'im', 'control', 'weight', 'birth', 'day',
'week']
```

Observation - The LDA modeling reveals key themes in drug reviews, including anxiety treatment, medication effects, skin conditions, pain management, and birth control, offering insights into common user concerns and experiences with their treatments.

[58]:

```python
## Mapping sentiment categories to numerical values: neutral = 1, negative = 0, positive = 2
df['target'] = df['sentiment'].map({'neutral': 1, 'negative': 0, 'positive': 2})
```

[59]:

```python
df.head()
```

[59]:

```
   Unnamed: 0  patient_id         drugName                     condition  \
0           0      163740       Mirtazapine                    depression
1           1      206473        Mesalamine  crohn's disease, maintenance
2           2       39293          Contrave                   weight loss
3           3       97768  Cyclafem 1 / 35                 birth control
4           4      208087           Zyclara                     keratosis

                                              review  rating       date  \
0  "i've tried a few antidepressants over the yea…    10.0 2012-02-28
1  "my son has crohn's disease and has done very …     8.0 2009-05-17
2  "contrave combines drugs that were used for al…     9.0 2017-03-05
3  "i have been on this birth control for one cyc…     9.0 2015-10-22
4  "4 days in on first 2 weeks.  using on arms an…     4.0 2014-07-03

   usefulCount  review_length  sentiment_score sentiment  \
0           22             68         0.000000   neutral
1           17             48         0.566667  positive
2           35            143         0.139063  positive
3            4            149         0.260926  positive
4           13             60         0.341667  positive

                                    cleaned_reviews  target
0  ive tried antidepressant year citalopram fluox…       1
```

```
1  son crohn disease done well asacol complaint s…       2
2  contrave combine drug used alcohol smoking opi…       2
3  birth control one cycle reading review type si…       2
4  day first week using arm face put vaseline lip…       2
```

Filtering Positive Reviews

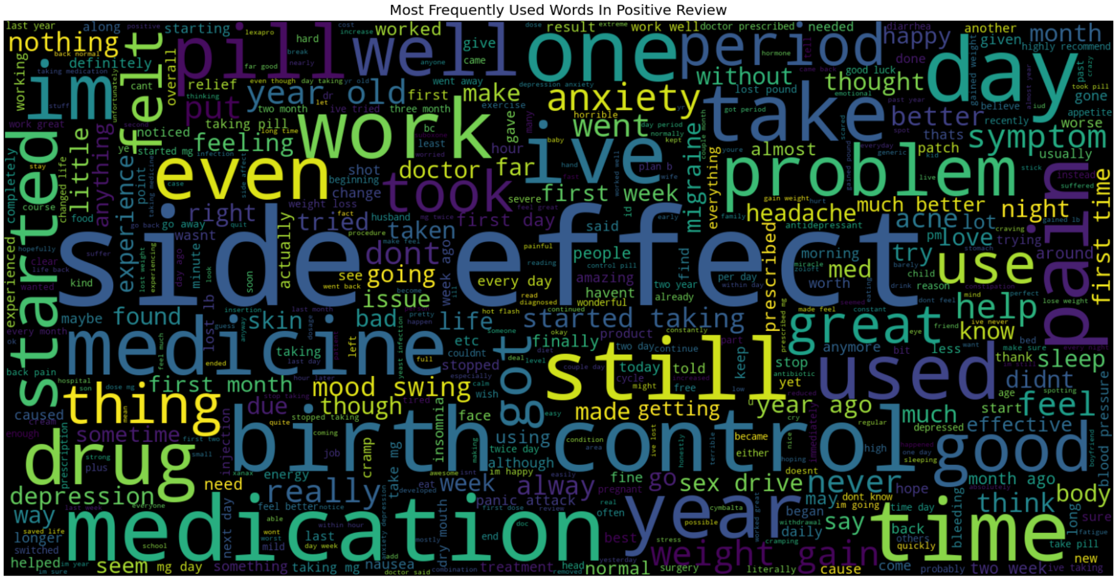
```
[60]: pos_reviews = df[df.target==2]
      pos_reviews.head()
```

```
[60]:    Unnamed: 0  patient_id         drugName                     condition  \
      1           1      206473       Mesalamine  crohn's disease, maintenance
      2           2       39293         Contrave                   weight loss
      3           3       97768  Cyclafem 1 / 35                 birth control
      4           4      208087          Zyclara                     keratosis
      6           6      169852    Amitriptyline          migraine prevention

                                                    review  rating       date  \
      1  "my son has crohn's disease and has done very …     8.0 2009-05-17
      2  "contrave combines drugs that were used for al…     9.0 2017-03-05
      3  "i have been on this birth control for one cyc…     9.0 2015-10-22
      4  "4 days in on first 2 weeks.  using on arms an…     4.0 2014-07-03
      6  "this has been great for me. i've been on it f…     9.0 2009-04-21

         usefulCount  review_length  sentiment_score sentiment  \
      1           17             48         0.566667  positive
      2           35            143         0.139063  positive
      3            4            149         0.260926  positive
      4           13             60         0.341667  positive
      6           32             64         0.185417  positive

                                           cleaned_reviews  target
      1  son crohn disease done well asacol complaint s…       2
      2  contrave combine drug used alcohol smoking opi…       2
      3  birth control one cycle reading review type si…       2
      4  day first week using arm face put vaseline lip…       2
      6  great ive week last week headache went away ty…       2
```

Positive Wordcloud Vizualization

```
[69]: text = ' '.join([word for word in pos_reviews['cleaned_reviews']])
      plt.figure(figsize=(20,60), facecolor=None)
      wordcloud = WordCloud(max_words=500, width=1600, height=800).generate(text)
      plt.imshow(wordcloud, interpolation='bilinear')
      plt.axis('off')
      plt.title('Most Frequently Used Words In Positive Review')
```

```
[69]: Text(0.5, 1.0, 'Most Frequently Used Words In Positive Review')
```



Most Frequently Used Words In Positive Review

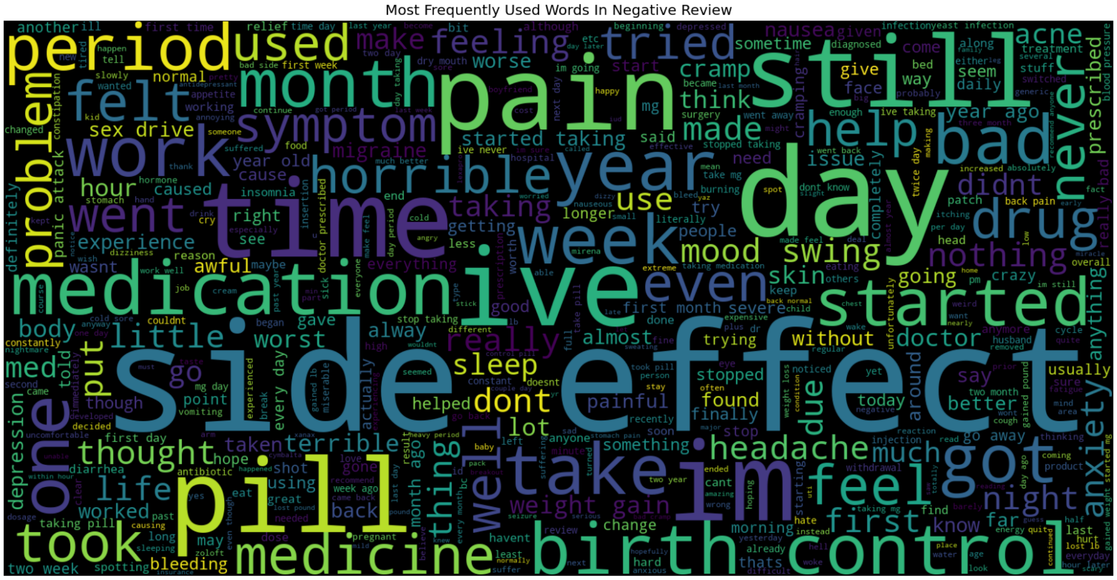
Observation - The word cloud highlights key terms in positive reviews, with “side effect,” “birth control,” “still,” “medication,” and “drug” standing out prominently. This suggests that while users are discussing medication positively, side effects and birth control are recurring themes in their experiences. Other notable terms like “take,” “pain,” “work,” and “time” reflect common concerns or observations about the efficacy and usage of these medications.

Negative Wordcloud Vizualization

[63]:
```
neg_reviews = df[df.target==0]
```

[68]:
```
text = ' '.join([word for word in neg_reviews['cleaned_reviews']])
plt.figure(figsize=(20,60), facecolor=None)
wordcloud = WordCloud(max_words=500, width=1600, height=800).generate(text)
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('Most Frequently Used Words In Negative Review')
```

[68]:
```
Text(0.5, 1.0, 'Most Frequently Used Words In Negative Review')
```



Observation - The word cloud for negative reviews highlights "side effect," "pain," "medication," "still," and "pill" as the most frequent terms, indicating that side effects and pain are major concerns in users' negative experiences with medications. Words like "anxiety," "horrible," "bad," "issue," and "problem" suggest dissatisfaction with both the medication's effectiveness and the adverse effects experienced. The prominence of terms such as "year," "month," and "day" may indicate a focus on the duration of use and the long-term impact of these side effects.

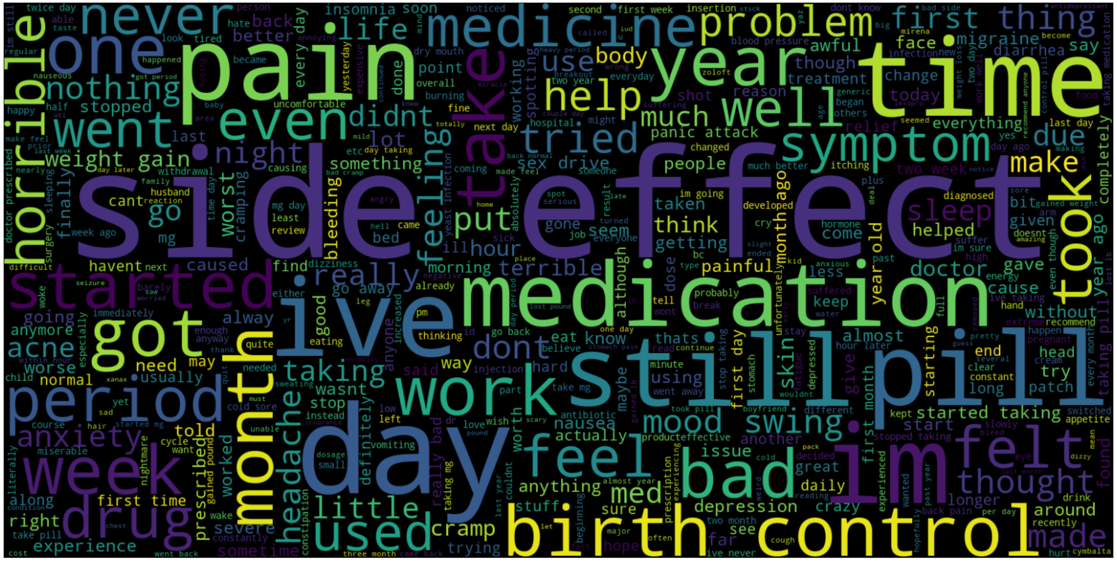
Neutral Wordcloud Vizualization

```
[65]: neu_reviews = df[df.target==0]
```

```
[67]: text = ' '.join([word for word in neu_reviews['cleaned_reviews']])
      plt.figure(figsize=(20,60), facecolor=None)
      wordcloud = WordCloud(max_words=500, width=1600, height=800).generate(text)
      plt.imshow(wordcloud, interpolation='bilinear')
      plt.axis('off')
      plt.title('Most Frequently Used Words In Neutral Review')
```

```
[67]: Text(0.5, 1.0, 'Most Frequently Used Words In Neutral Review')
```

Most Frequently Used Words In Neutral Review



Observation - The word cloud for neutral reviews highlights “side effect,” “birth,” “control,” “still,” and “pill” as the most frequent terms, indicating a focus on side effects and birth control. Words like “medication,” “time,” “pain,” “week,” “and”month” suggest ongoing experiences with the medication. Terms such as “feel,” “work,” and “drug” may indicate mixed feelings about the medication’s effectiveness. Overall, the neutral reviews provide a balanced view of ongoing medication experiences.

[71]:

```
#TF-IDF vectorization converts text into numerical data by highlighting␣
 ↪important words while downplaying common ones.
#This helps the model focus on relevant features from the reviews for better␣
 ↪prediction

# Vectorizing training data.
tfidf = TfidfVectorizer()
x = tfidf.fit_transform(df['cleaned_reviews'])
y = df['target']
## Applying Tf-Idf vectorizer on the Text column.
```

Splitting the Data for Training and Testing

[73]:

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.30,␣
 ↪random_state=42)
```

[74]:

```
print("Size of x_train:", (x_train.shape))
print("Size of y_train:", (y_train.shape))
print("Size of x_test:", (x_test.shape))
print("Size of y_test:", (y_test.shape))
```

```
Size of x_train: (32275, 36636)
Size of y_train: (32275,)
Size of x_test: (13833, 36636)
Size of y_test: (13833,)
```

Logistic Regression

[76]:

```python
# Fit Logistic Regression model
log_reg = LogisticRegression().fit(x_train, y_train)

# Predict on train
train_preds = log_reg.predict(x_train)

# Accuracy on train
print("Model accuracy on train is: ", accuracy_score(y_train, train_preds))  # Use y_train instead of x_train

# Predict on test
test_preds = log_reg.predict(x_test)

# Accuracy on test
print("Model accuracy on test is: ", accuracy_score(y_test, test_preds))

# Precision, Recall, F1 score (use 'macro' for multiclass)
precision = precision_score(y_test, test_preds, average='macro')
recall = recall_score(y_test, test_preds, average='macro')
f1 = f1_score(y_test, test_preds, average='macro')

print(f"Precision: {precision:.4f}")
print(f"Recall: {recall:.4f}")
print(f"F1 score: {f1:.4f}")
```

```
Model accuracy on train is:  0.9132455460883037
Model accuracy on test is:  0.8612737656329068
Precision: 0.5730
Recall: 0.5669
F1 score: 0.5686
```

K-Fold Cross-Validation for Logistic Regression Model

[77]:

```python
k = 15

# Initialize K-Fold with the specified number of splits
kfold = KFold(n_splits=k, shuffle=True, random_state=20)

# Perform cross-validation
K_results = cross_val_score(log_reg, x, y, cv=kfold)

# Calculate the mean accuracy
accuracy = np.mean(K_results)  # Use K_results directly, not absolute values
```

```
print("Mean Cross-Validation Accuracy:", accuracy)
```

```
Mean Cross-Validation Accuracy: 0.8678317804669614
```

Result Explanation: - The model performs well overall, with a 91% accuracy on the training data and 86% on new, unseen data. - It also does well when tested on different parts of the data, with an average accuracy of 87%. However, it's only about 57% accurate in correctly identifying positive cases, and it sometimes misses or misclassifies them.

Decision Tree Classifier

[85]:

```
# Fit the model on the training data only
DTree = DecisionTreeClassifier().fit(x_train, y_train)

# Predict on the test set (we avoid predicting on the training set to prevent␣
  ↪data leakage)
test_preds2 = DTree.predict(x_test)

# Accuracy on test
print("Model accuracy on test is: ", accuracy_score(y_test, test_preds2))

# Precision, Recall, F1 score (use 'macro' for multiclass)
precision = precision_score(y_test, test_preds2, average='macro')
recall = recall_score(y_test, test_preds2, average='macro')
f1 = f1_score(y_test, test_preds2, average='macro')

# Print the evaluation metrics
print(f"Precision: {precision:.4f}")
print(f"Recall: {recall:.4f}")
print(f"F1 score: {f1:.4f}")
```

```
Model accuracy on test is:  0.7646931251355454
Precision: 0.6099
Recall: 0.5876
F1 score: 0.5974
```

[86]:

```
k = 15

# Initialize K-Fold with the specified number of splits
kfold = KFold(n_splits=k, shuffle=True, random_state=20)

# Perform cross-validation
K_results = cross_val_score(DTree, x, y, cv=kfold)

# Calculate the mean accuracy
accuracy = np.mean(K_results)  # Use K_results directly, not absolute values
print("Mean Cross-Validation Accuracy:", accuracy)
```

```
Mean Cross-Validation Accuracy: 0.7719264964586515
```

Result Explanation

- The model exhibits a reasonable performance, achieving an accuracy of 76.47% on the test data, indicating it correctly predicts about three-quarters of the instances. The mean cross-validation accuracy of 77.19% suggests that the model maintains consistent performance across different subsets of the training data. However, the precision of 60.99% implies that while the model identifies a decent portion of positive predictions correctly, it also misclassifies some negative cases as positive. The recall of 58.76% indicates that the model is missing a significant number of actual positive cases. The F1 score of 59.74% provides a balanced measure of the model's performance, highlighting the trade-off between precision and recall.

HyperParameter Tuning Decision tree

[88]:

```
from sklearn.model_selection import GridSearchCV


# Define the parameter grid for hyperparameter tuning
param_grid = {
    'criterion': ['gini', 'entropy'],  # Criteria for splitting
    'max_depth': [None, 5, 10, 15, 20],  # Maximum depth of the tree
    'min_samples_split': [2, 5, 10],  # Minimum samples required to split an␣
 ↪internal node
    'min_samples_leaf': [1, 2, 5],  # Minimum samples required to be at a leaf␣
 ↪node
}

# Initialize the Decision Tree Classifier
DTree = DecisionTreeClassifier(random_state=42)

# Initialize GridSearchCV
grid_search = GridSearchCV(estimator=DTree, param_grid=param_grid,
                           scoring='f1_macro', cv=5, verbose=1, n_jobs=-1)

# Fit the Grid Search to the training data
grid_search.fit(x_train, y_train)

# Best parameters found
print("Best parameters found: ", grid_search.best_params_)

# Predict on the test set using the best estimator
best_model = grid_search.best_estimator_
test_preds2 = best_model.predict(x_test)

# Accuracy on test
print("Model accuracy on test is: ", accuracy_score(y_test, test_preds2))

# Precision, Recall, F1 score
precision = precision_score(y_test, test_preds2, average='macro')
```

```
recall = recall_score(y_test, test_preds2, average='macro')
f1 = f1_score(y_test, test_preds2, average='macro')

# Print the evaluation metrics
print(f"Precision: {precision:.4f}")
print(f"Recall: {recall:.4f}")
print(f"F1 score: {f1:.4f}")
```

```
Fitting 5 folds for each of 90 candidates, totalling 450 fits
Best parameters found:  {'criterion': 'gini', 'max_depth': None,
'min_samples_leaf': 1, 'min_samples_split': 2}
Model accuracy on test is:  0.7628858526711487
Precision: 0.6103
Recall: 0.5892
F1 score: 0.5985
```

Hyperparameter Tuning Result Explanation

The hyperparameter tuning process involved fitting the model using 5-fold cross-validation across 90 different combinations of parameters, resulting in a total of 450 model fits. The best parameters identified for the Decision Tree model were:

```
Criterion: 'gini' - This metric is used to measure the quality of a split, with the Gini impuri
Max Depth: None - Allowing the tree to expand fully, which can lead to overfitting if not contr
Min Samples Leaf: 1 - This indicates that a leaf can have as few as one sample, potentially all
Min Samples Split: 2 - This means a node must have at least two samples to be considered for sp
```

Despite finding optimal parameters, the model's performance on the test set shows an accuracy of 76.29%, with precision at 61.03%, recall at 58.92%, and an F1 score of 59.85%. These results indicate's the model retains similar performance to its previous iteration.

Random Forest Classifier

[90]:

```
# Fit the model on the training data only
RF = RandomForestClassifier().fit(x_train, y_train)

# Predict on train (this should use x_train, not x_test)
train_preds3 = RF.predict(x_train)

# Accuracy on train
print("Model accuracy on train is: ", accuracy_score(y_train, train_preds3))

# Predict on test
test_preds3 = RF.predict(x_test)

# Accuracy on test
print("Model accuracy on test is: ", accuracy_score(y_test, test_preds3))

# Precision, Recall, F1 score (use 'macro' for multiclass)
precision = precision_score(y_test, test_preds3, average='macro')
```

```python
recall = recall_score(y_test, test_preds3, average='macro')
f1 = f1_score(y_test, test_preds3, average='macro')

# Print the evaluation metrics
print(f"Precision: {precision:.4f}")
print(f"Recall: {recall:.4f}")
print(f"F1 score: {f1:.4f}")
```

```
Model accuracy on train is:  1.0
Model accuracy on test is:  0.7996096291476903
Precision: 0.8975
Recall: 0.5430
F1 score: 0.5898
```

[92]:

```python
k = 15

# Initialize K-Fold with the specified number of splits
kfold = KFold(n_splits=k, shuffle=True, random_state=20)

# Perform cross-validation
K_results = cross_val_score(RF, x, y, cv=kfold)

# Calculate the mean accuracy
accuracy = np.mean(K_results)  # Use K_results directly, not absolute values
print("Mean Cross-Validation Accuracy:", accuracy)
```

```
Mean Cross-Validation Accuracy: 0.8146525982414611
```

Result Explanation - The Random Forest model achieved a perfect accuracy of 100% on the training data, indicating potential overfitting. In contrast, its accuracy on the test data is 79.96%, which, while solid, shows a significant drop. The precision of 89.75% indicates that positive predictions are mostly correct, but the recall of 54.30% reveals that many actual positive cases are missed. The F1 score of 58.98% suggests a need for improvement in capturing positive instances. The mean cross-validation accuracy of 81.47% indicates the model maintains consistent performance across different training subsets.

Conclusion

Through a detailed analysis of drug reviews, this project uncovered several key insights and results:

Sentiment Distribution: The sentiment analysis revealed that the majority of reviews were positive (around 63.7%), followed by negative reviews (34.9%) and a small percentage of neutral reviews (1.4%). This indicates a generally favorable sentiment towards the drugs being reviewed.

Rating Distribution: The distribution of ratings showed a strong preference for higher ratings, with the most common rating being 10, followed by 9. This suggests that most users had positive experiences with the drugs they reviewed.

Top Reviewed Drugs: Drugs like Escitalopram and Mirena received the highest number of reviews, indicating their popularity and widespread use. Other highly reviewed drugs were those used for birth control, mental health treatments, and weight loss.

Useful Count vs. Rating: The scatter plot of useful counts against ratings showed no clear correlation, suggesting that users found reviews helpful regardless of whether they rated the drug highly or poorly.

Review Length: The histogram of review lengths showed that most reviews were relatively short, with only a few lengthy reviews. This suggests that users tend to leave brief feedback, but longer reviews may contain more detailed experiences.

Sentiment Trends Over Time: An analysis of sentiments over time showed a downward trend, indicating that user opinions may have become slightly more negative or less enthusiastic in recent years.

Machine Learning Results:

- The Logistic Regression model achieved an accuracy of 86.1% on the test set, showing strong performance in predicting user sentiment. However, precision and recall were moderate, indicating challenges in correctly identifying some positive and negative cases.
- The Decision Tree model achieved 76.4% accuracy, with a lower F1 score compared to Logistic Regression, suggesting room for improvement.
- The Random Forest model showed signs of overfitting, achieving 100% accuracy on the training data but dropping to 79.9% on the test data. While - precision was high (89.75%), recall remained relatively low, indicating that many actual positive cases were missed.
- Topic Modeling: Using Latent Dirichlet Allocation (LDA), key themes were identified in the reviews, such as anxiety treatment, medication side effects, skin conditions, pain management, and birth control. These topics provide valuable insights into the primary concerns and experiences of drug users.

```
[ ]:
```