AI CAPSTONE ECOMMERCE

December 15, 2021

[526]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from wordcloud import WordCloud


from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import␣
 ↪confusion_matrix,accuracy_score,classification_report
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.preprocessing import LabelEncoder


from tensorflow.keras.utils import to_categorical

tfidf = TfidfVectorizer(stop_words=set(stopwords.
 ↪words('english')),max_features=100)
```

[490]:
```
train_df = pd.read_csv("Datasets\Ecommerce\\train_data.csv")
test_df = pd.read_csv("Datasets\Ecommerce\\test_data.csv")
test_val_df = pd.read_csv("Datasets\Ecommerce\\test_data_hidden.csv")
train_df.shape,test_df.shape
```

[490]:
```
((4000, 8), (1000, 7))
```

[491]:
```
train_df.duplicated().sum(), test_df.duplicated().sum(), test_val_df.
 ↪duplicated().sum()
```

[491]:
```
(58, 3, 3)
```

[ ]:
```

```

[492]:
```
train_df.describe()
```

[492]:

```
                                                    name   brand  \
count                                               4000    4000   
unique                                                23       1   
top     Amazon Echo Show Alexa-enabled Bluetooth Speak…  Amazon   
freq                                                 676    4000   

                                               categories primaryCategories  \
count                                               4000              4000   
unique                                                23                 4   
top     Electronics,iPad & Tablets,All Tablets,Fire Ta…       Electronics   
freq                                                 628              2600   

                    reviews.date  \
count                       4000   
unique                       638   
top     2017-01-23T00:00:00.000Z   
freq                          99   

                                             reviews.text reviews.title  \
count                                               4000          3990   
unique                                              3598          2606   
top     I bought this kindle for my 11yr old granddaug…  Great tablet   
freq                                                   4           100   

       sentiment  
count       4000  
unique         3  
top     Positive  
freq        3749  
```

[493]:

```
train_df.dtypes
```

[493]:

```
name                 object
brand                object
categories           object
primaryCategories    object
reviews.date         object
reviews.text         object
reviews.title        object
sentiment            object
dtype: object
```

[494]:

```
test_df.head()
```

[494]:

```
                                                name   brand  \
0  Fire Tablet, 7 Display, Wi-Fi, 16 GB - Include…  Amazon   
1  Amazon Echo Show Alexa-enabled Bluetooth Speak…  Amazon   
```

```
2  All-New Fire HD 8 Tablet, 8" HD Display, Wi-Fi…  Amazon
3  Brand New Amazon Kindle Fire 16gb 7" Ips Displ…  Amazon
4  Amazon Echo Show Alexa-enabled Bluetooth Speak…  Amazon

                                          categories     primaryCategories  \
0  Fire Tablets,Computers/Tablets & Networking,Ta…           Electronics
1  Computers,Amazon Echo,Virtual Assistant Speake…  Electronics,Hardware
2  Electronics,iPad & Tablets,All Tablets,Fire Ta…           Electronics
3  Computers/Tablets & Networking,Tablets & eBook…           Electronics
4  Computers,Amazon Echo,Virtual Assistant Speake…  Electronics,Hardware

               reviews.date  \
0  2016-05-23T00:00:00.000Z
1  2018-01-02T00:00:00.000Z
2  2017-01-02T00:00:00.000Z
3  2017-03-25T00:00:00.000Z
4  2017-11-15T00:00:00.000Z

                                        reviews.text  \
0  Amazon kindle fire has a lot of free app and c…
1  The Echo Show is a great addition to the Amazo…
2  Great value from Best Buy. Bought at Christmas…
3  I use mine for email, Facebook ,games and to g…
4  This is a fantastic item & the person I bought…

                       reviews.title
0                  very handy device
1         Another winner from Amazon
2  simple to use and reliable so far
3                         Love it!!!
4                         Fantastic!
```

[495]:
```
test_df.describe()
```

[495]:
```
                                                     name   brand  \
count                                                1000    1000
unique                                                 23       1
top     Amazon Echo Show Alexa-enabled Bluetooth Speak…  Amazon
freq                                                  169    1000

                                             categories primaryCategories  \
count                                                1000              1000
unique                                                 23                 4
top     Electronics,iPad & Tablets,All Tablets,Fire Ta…       Electronics
freq                                                  169               676

                    reviews.date  \
count                       1000
unique                       366
top     2017-01-23T00:00:00.000Z
freq                          26

                                             reviews.text reviews.title
count                                                1000           997
unique                                                979           796
top     This device meets the needs of my grandson. He…  Great tablet
freq                                                    2            22
```

[496]:
```
test_val_df.head()
```

[496]:
```
                                                name   brand  \
0  Fire Tablet, 7 Display, Wi-Fi, 16 GB - Include…  Amazon
1  Amazon Echo Show Alexa-enabled Bluetooth Speak…  Amazon
2  All-New Fire HD 8 Tablet, 8" HD Display, Wi-Fi…  Amazon
3  Brand New Amazon Kindle Fire 16gb 7" Ips Displ…  Amazon
4  Amazon Echo Show Alexa-enabled Bluetooth Speak…  Amazon

                                          categories     primaryCategories  \
0  Fire Tablets,Computers/Tablets & Networking,Ta…           Electronics
1  Computers,Amazon Echo,Virtual Assistant Speake…  Electronics,Hardware
2  Electronics,iPad & Tablets,All Tablets,Fire Ta…           Electronics
3  Computers/Tablets & Networking,Tablets & eBook…           Electronics
4  Computers,Amazon Echo,Virtual Assistant Speake…  Electronics,Hardware

               reviews.date  \
0  2016-05-23T00:00:00.000Z
1  2018-01-02T00:00:00.000Z
2  2017-01-02T00:00:00.000Z
3  2017-03-25T00:00:00.000Z
4  2017-11-15T00:00:00.000Z

                                        reviews.text  \
0  Amazon kindle fire has a lot of free app and c…
1  The Echo Show is a great addition to the Amazo…
2  Great value from Best Buy. Bought at Christmas…
3  I use mine for email, Facebook ,games and to g…
4  This is a fantastic item & the person I bought…

                       reviews.title sentiment
0                  very handy device  Positive
1         Another winner from Amazon  Positive
2  simple to use and reliable so far  Positive
3                         Love it!!!  Positive
4                         Fantastic!  Positive
```

[497]:

```
test_val_df.describe()
```

[497]:

```
                                                     name   brand  \
count                                                1000    1000
unique                                                 23       1
top     Amazon Echo Show Alexa-enabled Bluetooth Speak…  Amazon
freq                                                  169    1000

                                               categories primaryCategories  \
count                                                1000              1000
unique                                                 23                 4
top     Electronics,iPad & Tablets,All Tablets,Fire Ta…       Electronics
freq                                                  169               676

                    reviews.date  \
count                       1000
unique                       366
top     2017-01-23T00:00:00.000Z
freq                          26

                                             reviews.text reviews.title  \
count                                                1000           997
unique                                                979           796
top     This device meets the needs of my grandson. He…  Great tablet
freq                                                    2            22

       sentiment
count       1000
unique         3
top     Positive
freq         937
```

[498]:

```
train_df.isnull().sum()
```

[498]:

```
name                  0
brand                 0
categories            0
primaryCategories     0
reviews.date          0
reviews.text          0
reviews.title        10
sentiment             0
dtype: int64
```

[499]:

```
test_df.isnull().sum()
```

```
[499]: name                0
       brand               0
       categories          0
       primaryCategories   0
       reviews.date        0
       reviews.text        0
       reviews.title       3
       dtype: int64
```

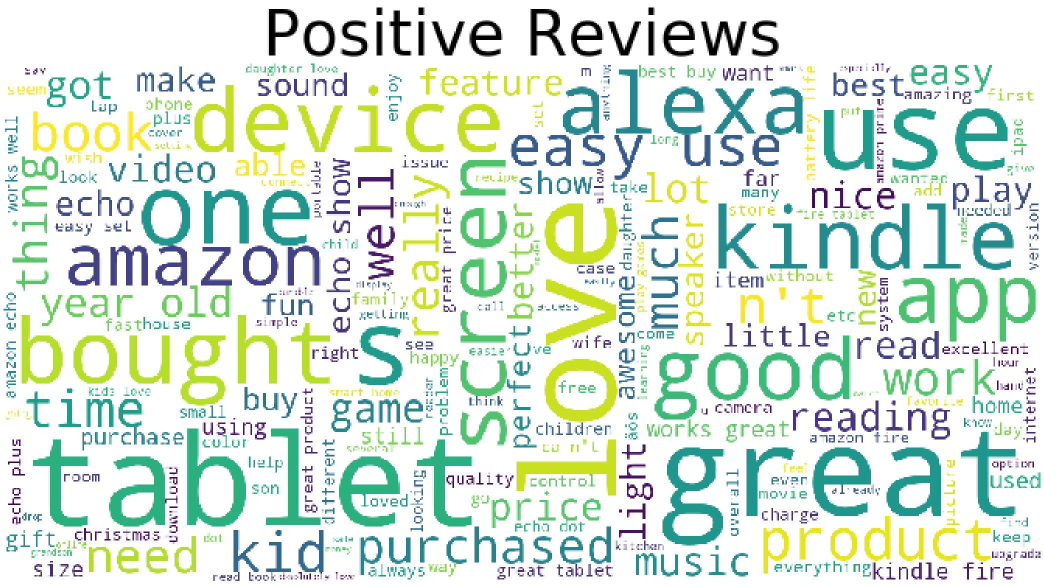
```
[500]: train_df["sentiment"].value_counts()
```

```
[500]: Positive    3749
       Neutral      158
       Negative      93
       Name: sentiment, dtype: int64
```

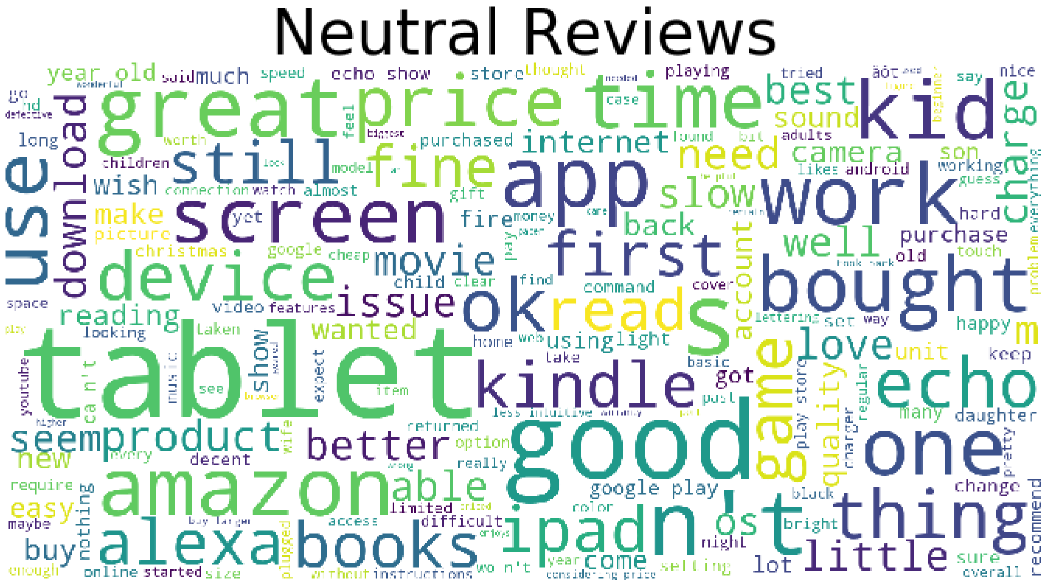
```
[501]: Positive_Review_Text = ""
       for review in  train_df[train_df["sentiment"]=="Positive"]["reviews.text"]:
         Positive_Review_Text += " " +review.lower()

       Negative_Review_Text = ""
       for review in  train_df[train_df["sentiment"]=="Negative"]["reviews.text"]:
         Negative_Review_Text += " " +review.lower()

       Neutral_Review_Text = ""
       for review in  train_df[train_df["sentiment"]=="Neutral"]["reviews.text"]:
         Neutral_Review_Text += " " +review.lower()
```

```
[502]: class WordCloudGeneration:
           def preprocessing(self, data):
               data = data.split(".")
               # convert all words to lowercase
               data = [item.lower() for item in data]
               # load the stop_words of english
               stop_words = set(stopwords.words('english'))
               # concatenate all the data with spaces.
               paragraph = ' '.join(data)
               # tokenize the paragraph using the inbuilt tokenizer
               word_tokens = word_tokenize(paragraph)
               # filter words present in stopwords list
               preprocessed_data = ' '.join([word for word in word_tokens if not word␣
        ↪in stop_words])
               return preprocessed_data

           def create_word_cloud(self, final_data,title=""):
               final_data=self.preprocessing(final_data)
               wordcloud = WordCloud(width=1600, height=800, max_font_size=200,␣
       ↪background_color="white").generate(final_data)
               plt.figure(figsize=(12,10))
               plt.imshow(wordcloud)
               plt.axis("off")
               plt.title(title,fontsize=40)
               plt.show()


       wordcloud_generator = WordCloudGeneration()
```

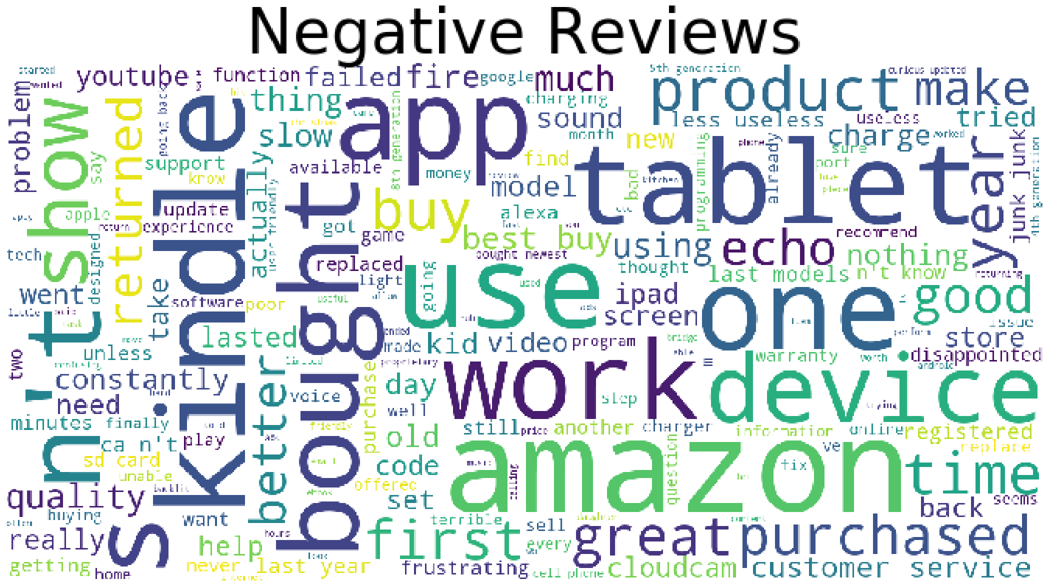
```
[503]: wordcloud_generator.create_word_cloud(Positive_Review_Text,"Positive Reviews")
```



```
[504]: wordcloud_generator.create_word_cloud(Neutral_Review_Text,"Neutral Reviews")
```



```
[505]: wordcloud_generator.create_word_cloud(Negative_Review_Text,"Negative Reviews")
```



1 Observations

Duplicates found in all datasets: Only one brand. - brand column can be dropped:

name, categories, primaryCategories, and sentiment are categorical: LabelEncoder

reviews.date to be converted to DateTime (Drop or not?):

reviews.text and reviews.title are text: TFIDF

null values in reviews.title:

class imbalance issue: undersampling oversampling

2 Tasks based on observations

Remove Duplicates

[506]:

```python
train_df=train_df[train_df.duplicated()==False]
test_df=test_df[test_df.duplicated()==False]
test_val_df=test_val_df[test_val_df.duplicated()==False]

train_df.reset_index(inplace=True)
test_val_df.reset_index(inplace=True)
test_df.reset_index(inplace=True)
```

fill null values

[507]:

```python
train_df['reviews.title'].fillna(value='',inplace=True)
test_val_df['reviews.title'].fillna(value=' ',inplace=True)
test_df['reviews.title'].fillna(value=' ',inplace=True)
```

Drop Brand Category

[508]:

```python
train_df.drop("brand",inplace=True,axis=1)
test_df.drop("brand",inplace=True,axis=1)
test_val_df.drop("brand",inplace=True,axis=1)
```

Encode categories

[509]:

```python
def to_labels(series):
  le=LabelEncoder()
  return le.fit_transform(series)

categories = ["name","categories","primaryCategories","sentiment"]

for cat in categories:
  train_df[cat]=to_labels(train_df[cat])
  test_val_df[cat]=to_labels(test_val_df[cat])
  if not cat=="sentiment":
    test_df[cat]=to_labels(test_df[cat])


train_df.shape,test_df.shape,test_val_df.shape
```

[509]:
```
((3942, 8), (997, 7), (997, 8))
```

Vectorize text with Tfidf

[537]:
```
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
def get_tfidf(series):
  new_series=[]
  for review in series:
    toks = word_tokenize(review)
    toks_sans_stopwords = [word for word in toks if not word in set(stopwords.
 ↪words('english'))]
    review_lemma = lemmatizer.lemmatize(" ".join(toks_sans_stopwords))
    new_series.append(review_lemma)
  result=pd.DataFrame(tfidf.fit_transform(new_series).toarray())
  return result
```

[538]:
```
train_df=pd.concat((train_df,get_tfidf(train_df["reviews.text"])),axis=1).
 ↪drop("reviews.text",axis=1)
test_df=pd.concat((test_df,get_tfidf(test_df["reviews.text"])),axis=1).
 ↪drop("reviews.text",axis=1)
test_val_df=pd.concat((test_val_df,get_tfidf(test_val_df["reviews.
 ↪text"])),axis=1).drop("reviews.text",axis=1)

train_df=pd.concat((train_df,get_tfidf(train_df["reviews.title"])),axis=1).
 ↪drop("reviews.title",axis=1)
test_df=pd.concat((test_df,get_tfidf(test_df["reviews.title"])),axis=1).
 ↪drop("reviews.title",axis=1)
test_val_df=pd.concat((test_val_df,get_tfidf(test_val_df["reviews.
 ↪title"])),axis=1).drop("reviews.title",axis=1)
```

[539]:
```
X_train = np.array(train_df.drop(["sentiment","reviews.date"],axis=1))
y_train = np.array(train_df["sentiment"])
y_test = np.array(test_val_df["sentiment"])

X_test = np.array(test_df.drop(["reviews.date"],axis=1))
```

3 Multinomial Naive Bayes Classification

[540]:
```
model = MultinomialNB()
```

[541]:
```
model.fit(X_train,y_train)
```

[541]:
```
MultinomialNB()
```

[542]:
```
preds=model.predict(X_test)
```

[543]:
```
pd.
 ↪DataFrame(confusion_matrix(y_test,preds),columns=["negative","neutral","positive"],index=["
```

[543]:
```
          negative  neutral  positive
negative         2        7        15
neutral          0       12        27
positive        51      248       635
```

[544]:
```
print(classification_report(y_test,preds))
```

```
              precision    recall  f1-score   support

           0       0.04      0.08      0.05        24
           1       0.04      0.31      0.08        39
           2       0.94      0.68      0.79       934

    accuracy                           0.65       997
   macro avg       0.34      0.36      0.31       997
weighted avg       0.88      0.65      0.74       997
```

[ ]:
```

```