Naan Mudhalva Phase3 Development: Part1

Data Pre Processing

```
from PIL import Image
from sklearn import svm
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics import accuracy_score
from sklearn.metrics import roc_curve
from sklearn.naive_bayes import MultinomialNB
from sklearn.neighbors import KNeighborsClassifier
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator

import collections
import matplotlib as mpl
import matplotlib.pyplot as plt
import numpy as np
import operator
import pandas as pd


tweets = pd.read_csv('Tweets.csv')


tweets.head(5)
```

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereas
0	570306133677760513	neutral	1.0000	N
1	570301130888122368	positive	0.3486	N
2	570301083672813571	neutral	0.6837	N
3	570301031407624196	negative	1.0000	Bad Fli
4	570300817074462722	negative	1.0000	Can't T

```
tweets['negativereason_gold'].value_counts()
```

```
Customer Service Issue                      12
Late Flight                                  4
Can't Tell                                   3
Cancelled Flight                             3
Cancelled Flight\nCustomer Service Issue     2
Late Flight\nFlight Attendant Complaints     1
Late Flight\nLost Luggage                    1
Bad Flight                                   1
Lost Luggage\nDamaged Luggage                1
Late Flight\nCancelled Flight                1
Flight Attendant Complaints                  1
Customer Service Issue\nLost Luggage         1
Customer Service Issue\nCan't Tell           1
Name: negativereason_gold, dtype: int64
```

```
tweets['airline_sentiment_gold'].value_counts()
```

```
negative    32
positive     5
neutral      3
Name: airline_sentiment_gold, dtype: int64
```

```
tweets['retweet_count'].value_counts()
```

```
0     12780
1       609
```

```
2        60
3        21
4        17
5         5
7         3
6         3
22        2
8         1
32        1
28        1
9         1
18        1
11        1
31        1
15        1
44        1
Name: retweet_count, dtype: int64
```

```
tweets.drop('negativereason_gold', axis=1, inplace=True)
tweets.drop('airline_sentiment_gold', axis=1, inplace=True)
tweets.drop('retweet_count', axis=1, inplace=True)
tweets.drop('tweet_coord', axis=1, inplace=True)

tweets.drop('tweet_location', axis=1, inplace=True)
tweets.drop('tweet_created', axis=1, inplace=True)
tweets.drop('user_timezone', axis=1, inplace=True)
tweets.drop('name', axis=1, inplace=True)

list(tweets.columns)
```

```
['tweet_id',
 'airline_sentiment',
 'airline_sentiment_confidence',
 'negativereason',
 'negativereason_confidence',
 'airline',
 'text']
```

```
unmeaningful = ['i', 'you', 'me', 'to', 'the', 'a', 'my', 'is', 'in', 'and', 'for', 'on', 'of',
                'your', 'so', 'was', 'have', 'it', 'at', 'with', 'that', 'from', 'do', 'get',
                'but', 'this', 'can', 'just', 'they', 'we', 'are', 'an', 'be', "i'm", 'will',
                'if', 'had', 'our', 'about', 'there', 'has', 'been', '-', 'by', 'like', 'or',
                'as', 'he', 'she', 'it', 'us', 'has' ,"i've", "it's", "don't", 'would', 'am',
               'flight', 'customer', 'any', 'very', "didn't", "you've", 'thing', 'take',
               'other', 'u', '', ' ']


def clean_text(str_in):

    res = ""
    str_in = str_in.lower()
    str_arr = str_in.split(' ')
    for word in str_arr:

        word = word.lower()

        if '@' in word or word == '' or word[:1] == '&':
            continue
        if word.lower() in unmeaningful:
            continue
        if word.isnumeric():
            continue
        res = res + " " + word
    return res


tweets["text"] = tweets["text"].apply(clean_text)


tweets.head(5)
```

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	p
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	real
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	