Project Update: Team Classifier

Complete:	**In Progress:**	**Not Started:**
Updated Project proposal document to implement the review comments.	Place the data into proper data structures	Develop model on train dataset
Identification of the dataset		Anomaly / Outlier Detection
Exploratory Data Analysis		
Clean, Parse and fix class imbalance		

So far no challenges are being faced.

Steps done so far for clean, parse and fix class imbalance for all reviews:

1. Tokenization
2. Remove punctuation
3. Remove stop words
4. Lemmatization
5. Make lower case

Initial feature extraction steps for all reviews:

1. Word count
2. Character count
3. Capital letters count
4. Digit count
5. Punctuation count
6. Sentiment Score

Exploratory Data Analysis:

Sampling data from the input file

Out[4]:

	category	rating	label	text_
40427	Clothing_Shoes_and_Jewelry_5	4.0	OR	I had read some reviews saying that this bra r...
40428	Clothing_Shoes_and_Jewelry_5	5.0	CG	I wasn't sure exactly what it would be. It is ...
40429	Clothing_Shoes_and_Jewelry_5	2.0	OR	You can wear the hood by itself, wear it with ...
40430	Clothing_Shoes_and_Jewelry_5	1.0	CG	I liked nothing about this dress. The only rea...
40431	Clothing_Shoes_and_Jewelry_5	5.0	OR	I work in the wedding industry and have to wor...

```
In [23]: # get the unique classifier
         dataset.label.unique()
```

```
Out[23]: array(['CG', 'OR'], dtype=object)
```

```
In [9]: # get the size of unique classifier
        dataset.label.unique().size
```

```
Out[9]: 2
```

Description of Unique Label Values:

OR - Original Review
CG - Computer Generated

In [11]:
```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40432 entries, 0 to 40431
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   category  40432 non-null  object
 1   rating    40432 non-null  float64
 2   label     40432 non-null  object
 3   text_     40432 non-null  object
dtypes: float64(1), object(3)
memory usage: 1.2+ MB
```

In [12]:
```
dataset.isna().sum()
```

Out[12]:
```
category    0
rating      0
label       0
text_       0
dtype: int64
```

In [13]:
```
dataset = dataset.dropna()
```

In [14]:
```
dataset.isnull().sum()
```

Out[14]:
```
category    0
rating      0
label       0
text_       0
dtype: int64
```

In [16]:
```
#skewness and kurtosis
#The skew result show a positive (right) or negative (left) skew.
#Values closer to zero showless skew.
print("Skewness: ")
dataset.skew()
print("Kurtosis: ")
dataset.kurt()
```

```
Skewness:
Kurtosis:
```

Out[16]:
```
rating    1.476615
dtype: float64
```
