**SENTIMENT ANALYSIS**

**By Mumtaz Uddin**

**DATA 602 – Prof. Darin Johnson**

**ABSTRACT** –

In this paper, I have worked on the Amazon cellphone and accessories reviews Dataset and have performed Exploratory Data Analysis, Topic Modelling to get a few insights. The main objective of this paper is to predict the ratings of the product from the text of the given review using TF-IF vectorization, Multinomial NB classifier. Further I experimented by remodeling it as a classification problem in which I tried classifying the reviews into sentiments which mocks the overall columns.

**INTRODUCTION** –

Using an existing dataset of cell phones and accessories reviews on amazon here you can see .json zip file which I unzipped which contains 1128437 rows × 12 columns of cell phone reviews and my task was to perform the necessary transformations to train both regression and classification models to predict the 'overall' field in the data set. Which you’ll see I will change it into sentiment.

**EXPLORATORY DATA ANALYSIS**

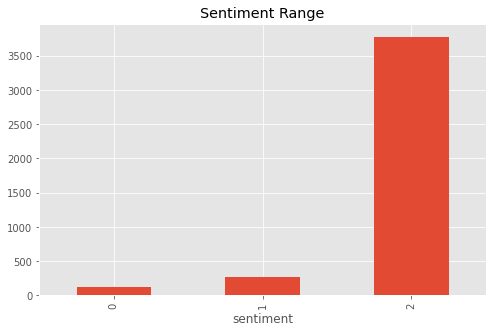
First, I have used sentiment (a new column) for use overall less than 3 as negative, greater than 3 as positive, and 3 as neutral. If you'd prefer a numeric value 0 should be negative, 1 neutral and 2 positives.

# Sentiment

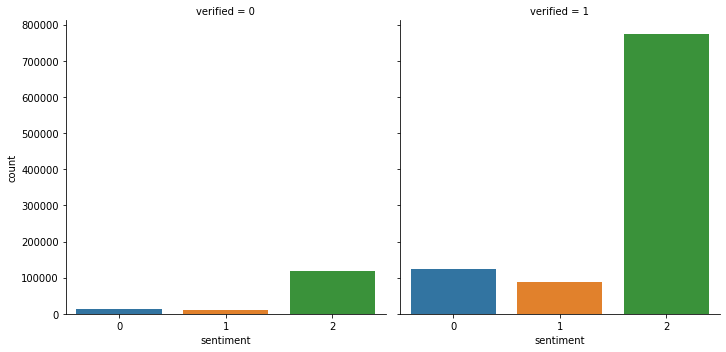
# 2 is Positive, score above 3

# 1 is Neutral, score = 3

# 0 is Negative, score below 3



Then, I plotted the verified/unverified attribute of the reviews against all rating sentiment as overall levels to understand the importance of the verified column in the analysis.

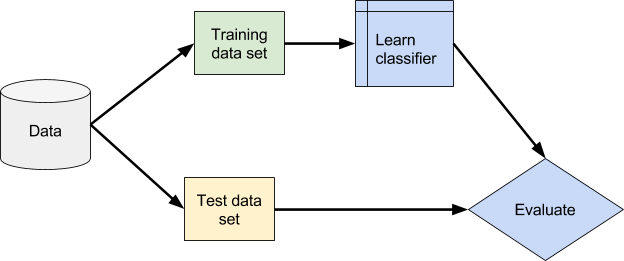


Also, I have tried to predict the sentiments from the text available in the reviews, and I have calculated the accuracy of our model by comparing the predicted ratings with the actual ratings.

**LEARNING FROM REVIEWS TEXT-**

First, I performed some Data cleaning by dropping unnecessary columns and the null values to perform vectorization and further analysis. For this analysis I have only used the reviews column and the correspond rating value, for predicting the rating from text. For these predictions I have used TF-ID vectorizer Logistic regression classifier and MultinomialNB classifier with the dataset split into

X\_train, X\_test, y\_train, y\_test = train\_test\_split(df.drop('sentiment',axis=1), df.sentiment, test\_size=0.3,random\_state=2)



And the accuracy was – 0.7884195645358776

I tried generating the classification reports for these predictions which shows the accuracy of the model

precision recall f1-score support

0 0.00 0.00 0.00 41752

1 0.00 0.00 0.00 29271

2 0.79 1.00 0.88 267099

accuracy 0.79 338122

macro avg 0.26 0.33 0.29 338122

weighted avg 0.62 0.79 0.70 338122

And the confusion matrix for this model is

[[ 712 468 2420]

[ 0 0 0]

[137894 97710 887869]]

Multinomial NB-

This test was to check the predict of x\_test with the actual results after multinomial

Multinomial NB: [2, 2, 2, 2, 2, 2, 2, 2, 2, 2]

Actual: 544549 2

631773 0

294069 2

686835 2

493995 0

358000 1

691075 2

279829 2

753725 2

642989 2

**CONCLUSION** –

I have learned about the feature importance of the features provided on each review and effect of ‘overall’ attribute of the reviews on the cellphone and accessories reviews I also learned to predict the rating(sentiment) as deduced from overall of a product by analyzing the text of the customer reviews.

**REFERENCE-**

* <https://towardsdatascience.com/predicting-sentiment-of-amazon-product-reviews-6370f466fa73>
* <https://stackoverflow.com/questions/23455728/scikit-learn-balanced-subsampling>
* <https://towardsdatascience.com/linear-classifiers-an-overview-e121135bd3bb>
* <http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html>