This dataset has 3150 Amazon Alexa customer comments (text) on amazon Alexa products in about 3 months.

AIM:

We perform sentiment analysis on the above stated dataset. We detect positive and negative feedback and predict the feedback from the reviews using different machine learning models and then compare the accuracy of each to get the best fit model for our data analysis.

STEPS FOLLOWED FOR DATA ANALYSIS:

1)READING TSV FILE IN PYTHON USING PANDAS

	rating	date	variation	verified_reviews	feedback
0	5	31-Jul-18	Charcoal Fabric	Love my Echo!	1
1	5	31-Jul-18	Charcoal Fabric	Loved it!	1
2	4	31-Jul-18	Walnut Finish	Sometimes while playing a game, you can answer...	1
3	5	31-Jul-18	Charcoal Fabric	I have had a lot of fun with this thing. My 4 ...	1
4	5	31-Jul-18	Charcoal Fabric	Music	1

2)UNDERSTANDING THE DATA

- There are 5 columns. The feedback column is the target variable and rest 4 are features.
- There are no NULL values or missing values in our dataset
- The mean rating is 4.46 on a scale of 5 which denotes that most of the reviews are positive. This can be said as there is high correlation between ratings and reviews.
- The ratings are not affected by the date on which the comment was made. Therefore, the date column does not add any new insights to the data.

```
corr_matrix = data.corr()
corr_matrix['feedback'].sort_values(ascending=False)
```

```
feedback    1.000000
rating      0.861968
Name: feedback, dtype: float64
```

- The dataset is imbalanced. The number of 0 is less as compared to number of 1's, which can be seen from the following observation:

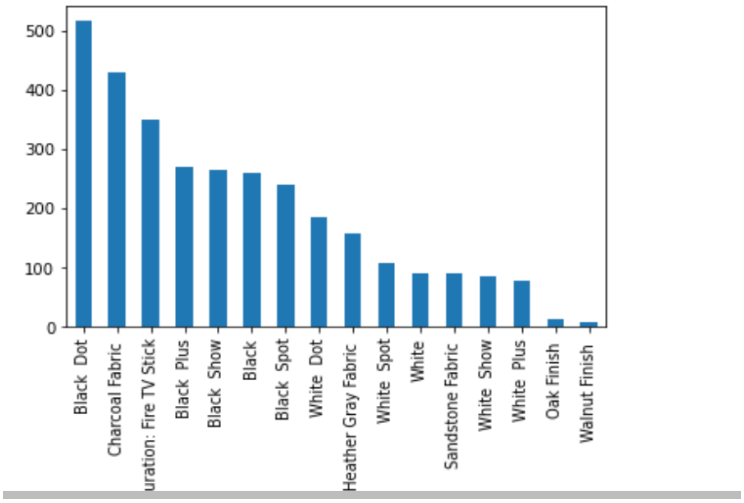
```
data['feedback'].value_counts()
#imbalanced dataset
```

```
1    2893
0     257
```

- The product has very good ratings. Majority of people are very happy with the product since they have given 5 stars. Reviews for ratings less than 3 are classified as negative.

	feedback		
	count	5%	max
rating			
1	161.0	0.0	0.0
2	96.0	0.0	0.0
3	152.0	1.0	1.0
4	455.0	1.0	1.0
5	2286.0	1.0	1.0

- Black dot is the most popular and walnut finish least popular from the graph on basis of the number of ratings received from the customers.



3)CLEANING OF DATA

The following steps are followed:

- Remove missing values and replace null values by median.
- Converting text to lowercase
- Removing punctuation
- Removing emoticons
- Tokenisation
- Removing stop words
- Lemmatization

4)SPLITTING DATA INTO TRAINING AND TEST DATA

5)BUILDING DIFFERENT MODELS AND CHECKING FOR ACCURACY AND F1 SCORE

Here we built two datasets: One using COUNTVECTORISER and second using TFIDFVECTORIZER. These 2 are different ways in which we can convert our text to numbers which can be understood by the computer and then it can be fed to different machine learning models.

6)BUILD DIFFERENT MODELS

OUR FINDINGS:

The table shows the different f1 scores for different models we used for data analysis.

	MULTINOMIAL NAIVE BAYES	RANDOM FOREST	SVM	LOGISTIC REGRESSION
BOW MODEL	0.96	0.96	0.96	0.95
TFIDF MODEL	0.95	0.96	0.96	0.96

The table shows the different accuracy scores for different models we used for data analysis.

	MULTINOMIAL NAIVE BAYES	RANDOM FOREST	SVM	LOGISTIC REGRESSION
BOW MODEL	0.93	0.94	0.93	0.91
TFIDF MODEL	0.91	0.94	0.93	0.93

CONCLUSIONS:

The evaluation metric used is F1 SCORE and confusion matrix. The reason being, our dataset is imbalanced as it contains more no of 1's than zeroes. So, it becomes necessary to evaluate the false positive and false negative rate accurately rather than true positive and true negative because the distribution is imbalanced.

From the above tables, we can see that both BOW MODEL and TFIDF MODEL perform nearly same for different machine learning models with their F1 scores very near to each other. But does that mean we can take any model as best fit for this dataset?

The answer is No.

We now analyse the confusion matrix of all the different models to see which one suit the best:

Out of TFIDF and BOW model, TFIDF model is preferred usually. The reason being,

For BOW MODEL: FP is higher than FN and rate of FN is very less but it exists.

For TFIDF MODEL: FP is higher than FN and rate of FN is 0% which means there is no value which is predicted falsely. Therefore, the f1 score goes high. This applies to all the machine learning models we used.

Now we have to select the machine learning model which minimises our FP and FN because this percentage is high, which means the model is predicting a negative review as a positive which is not good for a business as it leads to misleading numbers and product will not be improved.

RESULT:

The random forest classifier for TFIDF model is the best fit for the data as it has minimum FP, maximum TN with TP percentages remaining more or less constant among all TFIDF models.