Sentiment Analysis of Customer Reviews

Shivani Nandani
CS-306, Data Analysis and Visualization
Dhirubhai Ambani Institute of Information & Communication Technology
201801076@daiict.ac.in

1 Introduction

Sentiment analysis is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques. It helps gauge public opinion, conduct nuanced market research, monitor brand and product reputation, and understand customer experiences. Sentiment analysis, thus, has been an area of interest for the industry that relies heavily on consumer feedback. In this project, we have employed data analysis techniques on the Amazon Fine Food Reviews dataset. All the codes are uploaded on the the GitHub repository - sentiment-analysis.

2 About the dataset

In this project, we have used the Amazon Fine Food Reviews dataset available on Kaggle. The dataset contains 568,454 reviews by 256,059 users for 74,258 products from Oct 1999 - Oct 2012. Around 260 users have more than 50 reviews.

Various parameter in the dataset are:

- Id - review number
- ProductId - unique identifier for the product
- UserId - unique identifier for the user
- ProfileName - profile name of the user
- HelpfulnessNumerator - number of users who found the review helpful
- HelpfulnessDenominator - number of users who indicated whether they found the review helpful or not
- Score - rating between 1 and 5
- Time - timestamp
- Summary - summary of the review
- Text - text of the review

3 Method and Observations

The method used in the project is as given below:

- Step 1 - Read data
- Step 2 - Data Preprocessing
  - check for missing values
  - check for duplicates
  - check for invalid entries
  - preprocessing text
    - remove html tags
    - remove alphanumeric
    - convert to lowercase
    - remove stopwords
    - stemming the words
- Step 3 - Exploratory Data Analysis
  - distribution of score
  - review trend for each year
  - distribution of positive, negative and neutral reviews
  - understanding user data
  - understanding correlation

3.1 Read Data

To read data, we have used `Pandas` library. Output of the read is shown below:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...
...	...	...	...	...	...	...	...	...	...	...
568449	568450	B001EO7N10	A28KG5XORO54AY	Lettie D. Carter	0	0	5	1299628800	Will not do without	Great for sesame chicken..this is a good if no...
568450	568451	B003S1WTCU	A3I8AFVPEE8KI5	R. Sawyer	0	0	2	1331251200	disappointed	I'm disappointed with the flavor. The chocolat...
568451	568452	B004I613EE	A121AA1GQV751Z	pksd "pk_007"	2	2	5	1329782400	Perfect for our maltipoo	These stars are small, so you can give 10-15 o...
568452	568453	B004I613EE	A3I8EVCTXKNOH	Kathy A. Welch "katwel"	1	1	5	1331596800	Favorite Training and reward treat	These are the BEST treats for training and rew...
568453	568454	B001LR2CU2	A3LGQPJCZVL9UC	srfell17	0	0	5	1338422400	Great Honey	I am very satisfied ,product is as advertised,...

568454 rows × 10 columns

Figure 1: `df` without data preprocessing

As is visible in the image, there are 568454 rows and 10 columns in the dataset.

3.2 Data Preprocessing

Data preprocessing is an essential part of data analysis. Since most real world problems are not bound in a specific boundary, datasets generally have incomplete data (missing values), noisy data, outliers or discrepancies. All these factors impact the success rate of a model, and it is hence important to perform data preprocessing to prepare it for further analysis.

3.2.1 Check for missing values

First, we check if any row has a missing value. In case it is present, we will drop that row. On running the test for missing value, we find that the dataset has no missing values.

3.2.2 Check for duplicates

Next, we check for duplicates in the dataset. Two data samples are considered same if they have the same values for UserID, ProfileName and Text.

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
3	4	B000LVA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
10	11	B0001PB9FE	A3HDKO7OWBQNK4	Canadian Fan	1	1	5	1107820800	The Best Hot Sauce in the World	I don't know if it's the cactus or the tequila...
29	30	B0001PB9FY	A3HDKO7OWBQNK4	Canadian Fan	1	1	5	1107820800	The Best Hot Sauce in the World	I don't know if it's the cactus or the tequila...
68	69	B000E7VI7S	A1KL2LAWBKK6UQ	calmnsense	0	0	3	1169251200	How much would you pay for a bag of chocolate ...	If you're impulsive like me, then $6 is ok. Do...
69	70	B000E7VI7S	AWCBF22WIN57F	C. Salcido	0	2	5	1185753600	pretzel haven!	this was sooooo delisicious but too bad I ate e...
568409	568410	B001BCLWM4	A2PE8AGAV6OPL7	Dark Water Mermaid	3	3	5	1309651200	Quality & affordable food	I was very pleased with the ingredient quality...
568410	568411	B001BCLWM4	A88HLWDCU57WG	R2B	2	2	5	1332979200	litter box	My main reason for the five star review has to...
568411	568412	B001BCLWM4	AUXIM5Y8FX55S	DAW	1	1	5	1319500800	Happy Camper	I bought this to try on two registered Maine C...
568412	568413	B001BCLWM4	AVZ2OZ479QYE8	Ai Ling Chow	0	0	5	1336435200	Two Siberians like it!	When we brought home two 3-month-old purrbred ...
568413	568414	B001BCLWM4	AI3Y26HLPYW4L	kimosabe	1	2	2	1330041600	premium edge cat food	My cats don't like it. what else can I say to ...

232828 rows × 10 columns

Figure 2: all duplicates found in `df`

We see that there are 232828 duplicates values. This includes all the versions of a duplicated data sample. Out of the multiple entries, we will retain the with one with the maximum `HelpfulnessNumerator`. After appending the required samples, we get 393642 samples.

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...
5	6	B006K2ZZ7K	AD7DSRX1MGOEU	Twoapennything	0	0	4	1342051200	Nice Taffy	I got a wild hair for taffy and ordered this f...
...	...	...	...	...	...	...	...	...	...	...
550041	550042	B000EMY0E2Y	A2RNYW2HIQBGSM	Carlos Alvarez	0	0	1	1230681600	Terrible, don't waste your money	These are just awful. I love all kinds of toa...
550082	550083	B001EO5RRM	A3TIPDY2RO19XB	K. Bauer	1	1	5	1275004800	Delicious!	I was pleasantly surprised when we bought this...
559111	559112	B0057LBU6U	AKXZBEDYDMOA5	onthelookout	0	0	3	1343174400	THINKERS CHICKEN DOG TREAT	My boxer loves this treat. He has not turned d...
559112	559113	B0057LBU6U	A24FX0OC3S6XUQ	Stone-Man	0	0	4	1337558400	Dog loves these!!!	My dog goes crazy for these. They are BIG like...
560006	560007	B005BSYQI8	AWBWQRSRUVO50	Lexxie29	1	1	5	1344556800	Great product!!!	Fast shipping, fresh unwashed seeds, great for...

393642 rows × 10 columns

Figure 3: `df` without duplicates

3.2.3 Check for invalid entries

We will now check if the dataset has any invalid entries. An entry will be considered as invalid if the `HelpfulnessNumerator`>`HelpfulnessDenominator`. We find two such entries as shown below:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...
...	...	...	...	...	...	...	...	...	...	...
568449	568450	B001E07N10	A2BGZ52M908X9J	Lettie D. Carter	0	0	5	1299628800	Will not do without	Great for sesame chicken..this is a good if no...
568450	568451	B003S1WTCU	A3I8AFVPEE8KI5	R. Sawyer	0	0	2	1331251200	disappointed	I'm disappointed with the flavor. The chocolat...
568451	568452	B004I613EE	A121AA1GQV751Z	pksd "pk_007"	2	2	5	1329782400	Perfect for our maltipoo	These stars are small, so you can give 10-15 o...
568452	568453	B004I613EE	A3IBEVCTXKNOH	Kathy A. Welch "katwel"	1	1	5	1331596800	Favorite Training and reward treat	These are the BEST treats for training and rew...
568453	568454	B001LR2CU2	A3LGQPJCZVL9UC	srfell17	0	0	5	1338422400	Great Honey	I am very satisfied ,product is as advertised,...

568454 rows × 10 columns

Figure 4: `df` without duplicates

On removing these two entries, we get 393640 entries, which is then re-indexed to match the new state.

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...
...	...	...	...	...	...	...	...	...	...	...
393635	568450	B001E07N10	A2BGZ52M908X9J	Lettie D. Carter	0	0	5	1299628800	Will not do without	Great for sesame chicken..this is a good if no...
393636	568451	B003S1WTCU	A3I8AFVPEE8KI5	R. Sawyer	0	0	2	1331251200	disappointed	I'm disappointed with the flavor. The chocolat...
393637	568452	B004I613EE	A121AA1GQV751Z	pksd "pk_007"	2	2	5	1329782400	Perfect for our maltipoo	These stars are small, so you can give 10-15 o...
393638	568453	B004I613EE	A3IBEVCTXKNOH	Kathy A. Welch "katwel"	1	1	5	1331596800	Favorite Training and reward treat	These are the BEST treats for training and rew...
393639	568454	B001LR2CU2	A3LGQPJCZVL9UC	srfell17	0	0	5	1338422400	Great Honey	I am very satisfied ,product is as advertised,...

393640 rows × 10 columns

Figure 5: `df` without duplicates and invalid entries

3.2.4 Preprocessing text

For preprocessing the text (i.e., the `Review`), we do the following:

- remove html tags
- remove alphanumeric
- convert to lowercase
- remove stopwords
- stemming the words

HTML tags are introduced when the user adds links other such components to the review. Alphanumeric characters (such as numbers, special characters etc.) are also removed as they do not add any *sentimental value* to the review. To correctly judge the words we change all the letters to lowercase so that they can all be treated equally, without creating unnecessary duplicates. Finally, we deal with stopwords and stemming of words. Stopwords are the words that do not add any meaning to the sentence and thus can be ignored without sacrificing the essence of the sentence. Examples of stopwords are words like *the*, *are*, *have* etc. Stemming of words involves reducing the words to their base form. For example, on stemming,

eating and *eats* are both changed to *eat*. Stemming is an important part of text preprocessing as it gives us the root for each value, which allows the model to consider all valid versions of the word.

The final dataset after complete data preprocessing has 393640 values as shown below:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	bought sever vital dog food product found good...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	product arriv label jumbo salt peanut peanut a...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	confect around centuri light pillowi citrus go...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	look secret ingredi robitussin believ found go...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	great taffi great price wide assort yummi taff...
...	...	...	...	...	...	...	...	...	...	...
393635	568450	B001ED7N10	A28KG5XORO54AY	Lettie D. Carter	0	0	5	1299628800	Will not do without	great sesam chicken good better restur eaten h...
393636	568451	B003S1WTCU	A3IBAFVPEBKI5	R. Sawyer	0	0	2	1331251200	disappointed	disappoint flavor chocol note especi weak milk...
393637	568452	B004I613EE	A121AA1GQV751Z	pksd "pk_007"	2	2	5	1329782400	Perfect for our maltipoo	star small give 10 15 one train session tri tr...
393638	568453	B004I613EE	A3IBEVCTXKNOH	Kathy A. Welch "katwel"	1	1	5	1331596800	Favorite Training and reward treat	best treat train reward dog good groom lover c...
393639	568454	B001LR2CU2	A3LGQPJCZVL9UC	srfell17	0	0	5	1338422400	Great Honey	satisfi product advertis use cereal raw vinega...

393640 rows × 10 columns

Figure 6: `df` after complete data preprocessing

3.3 Exploratory Data Analysis

3.3.1 Distribution of scores

From the below-given histogram, we can see that most reviews are positive.

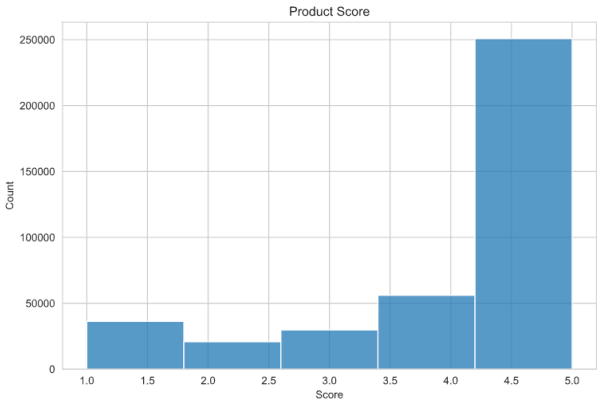

Figure 7: score distribution among various reviews

3.3.2 Review trend for each year

Here we can see that in the initial years, i.e., from 2001 to 2006, the number of reviews remain almost constant. However, the rate increases after 2006 and most reviews after that period are positive.

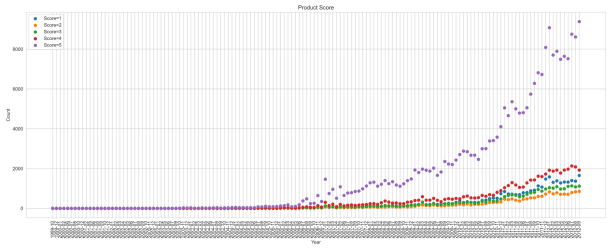

Figure 8: review trend for each year (month-wise manner)

3.3.3 Distribution of positive, negative and neutral reviews

We divided the reviews in three categories based on their score – any review with score greater than 3 has be considered positive; any review with score equal to 3 is consider neutral; any review with score less than 3 has be considered negative.

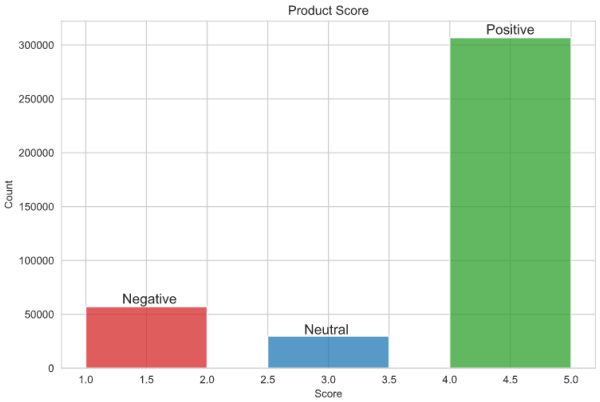

Figure 9: histogram showing number of positive, negative and neutral reviews

3.3.4 Understanding user data

We can see that the there are 256055 unique users. Since the mean is 1.537326, we can say that most users bought the product only once.

```
count    256055.000000
mean          1.537326
std           2.732791
min           1.000000
25%           1.000000
50%           1.000000
75%           1.000000
max         329.000000
Name: UserId, dtype: float64
```

Figure 10: user data

3.3.5 Wordclouds

Wordclouds are display the most frequently occurring words in the given sample. Here, the size of a word is proportional to its frequency.

Figure 11: wordcloud for all reviews

Figure 12: wordcloud for positive reviews

Figure 13: wordcloud for neutral reviews

Figure 14: wordcloud for negative reviews

3.3.6 Understanding correlation

We checked the correlation between various fields:

Parameter 1	Parameter 2	Correlation
Score	ProductID	-0.03879
Score	UserId	-0.0139827
Score	HelpfulnessNumerator	-0.03620
Score	Length of the review	-0.05968

4 Conclusions

- In FIG[7] and FIG[9], we see that most reviews are positive. This can be due to the reasons discussed below. However, a consequence of this is that the dataset is not balanced. Thus, in this case, if a model to predict if a review is positive or negative is trained using this dataset, is will be inherently biased towards positive reviews. In such cases, accuracy will give a false sense of correctness of the model.
- In FIG[8], we see that most reviews are positive after 2006. This can be interpreted in two ways:
  1. The number of reviews grew with the increases number of users of Amazon, the graph of which is give below[1]. The nature of this graph is similar to that of FIG[8].

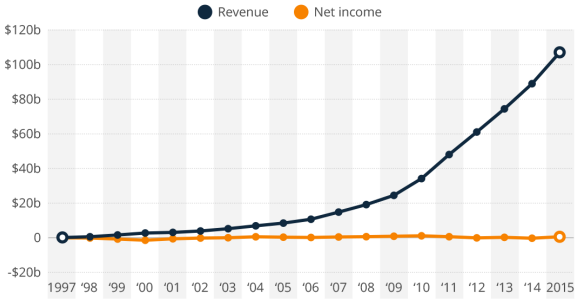

Figure 15: growth of Amazon

  2. This trend seen is FiG[8] can also be explained by the marketing tactics used by the companies in which the users are either paid to write fake positive reviews, or company employees write positive reviews using fake accounts.

  However, the actual reason can be a combination of above-mentioned reasons.
- In FIG[10], we see that the maximum number of reviews posted by a single user is 329. This *UserId* acts as an outlier in this case since it is extremely far from the mean and variance of the data. In real life context, this can be related to the one of the reasons for the sudden growth mentioned above. We checked the distribution of scores for this users, we find that most of the reviews are rated 4. This, thus, can be a proof of the fraudulent ways in which companies try to increases their customer ratings.

	Id	ProductId	HelpfulnessNumerator	HelpfulnessDenominator	Score
count	329.000000	329.000000	329.000000	329.000000	329.000000
mean	270291.358663	38124.367781	1.124620	1.234043	3.659574
std	170106.232415	19451.560950	2.728514	2.968831	0.657676
min	110.000000	17.000000	0.000000	0.000000	1.000000
25%	99740.000000	22090.000000	0.000000	0.000000	3.000000
50%	271766.000000	40085.000000	0.000000	0.000000	4.000000
75%	409657.000000	54438.000000	1.000000	1.000000	4.000000
max	566680.000000	67248.000000	20.000000	22.000000	4.000000

Figure 16: details for UserId with 329 reviews

- From FIG[11], FIG[12], FIG[??] and FIG[14], we see that most words associated with the complete set are similar to the ones associated with positive and neutral sets. This can be attributed

[1]Source - One simple chart that shows Amazon's relentless focus on long-term growth

to the imbalanced dataset. In the wordcloud for negative set, we see words such as good and great. This is because they were used with the word 'not' to imply dissatisfaction. However, these were removed as a part of text preprocessing during the removal of stopwords. Finally, we see that the negative set contains a lot of conditional words (such as 'however', 'though' etc.). This can show that most users were not completely dissatisfied with the product but the same did not match their expectations.

- From the correlation data we can see that the parameters are not dependent on each other. Thus, an individual doesn't always rate a product good or bad, but do so based on their experiences with the products. Similarly, most products have their rating fluctuating. Even with *Sentiment* as a parameter instead of *Score*[2], we see that correlation does not improve.

References

[1] Sentiment Analysis: Concept, Analysis and Applications

[2] Sentiment Analysis Explained

[3] Techniques and applications for sentiment analysis

[4] A Gentle Introduction to Imbalanced Classification

[5] Learning from imbalanced data.

[2] given in the GitHub repository