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.

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

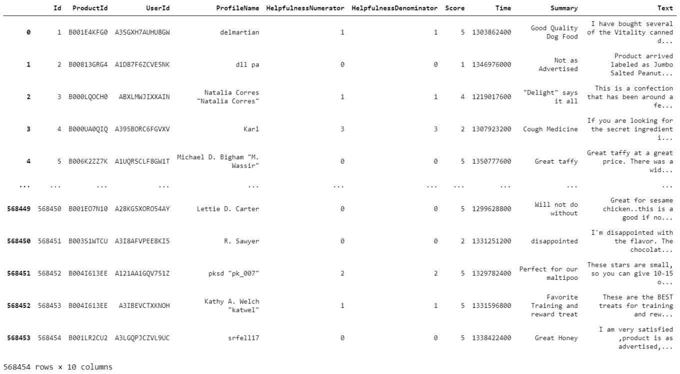
3 Method and Observations

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	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 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

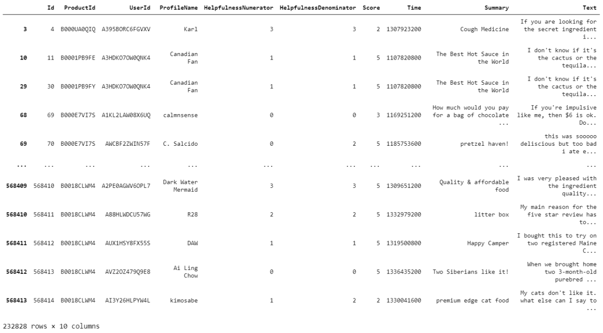
Data preprocessing is an essential part of data analysis. Since most real world problems are not bound in a specific boundary, datasets generally have incomplete data (missing values), noisy data, outliers or discrepancies. All these factors impact the success rate of a model, and it is hence important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

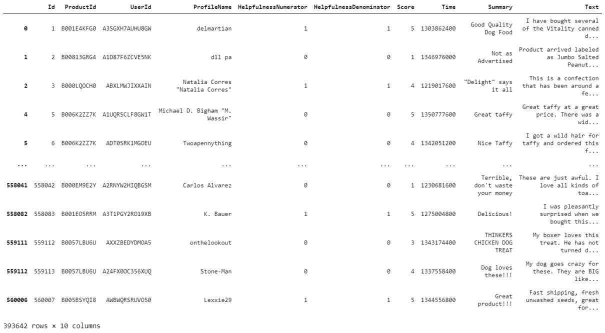


	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
10	11	B0001PB9FE	A3HDKO7OW0QNK4	Canadian Fan	1	1	5	1107820800	The Best Hot Sauce in the World	I don't know if it's the cactus or the tequila...
29	30	B0001PB9FY	A3HDKO7OW0QNK4	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	A1KL2LAW08X6UQ	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	AWCBF2ZWIN57F	C. Salcido	0	2	5	1185753600	pretzel haven!	this was sooooo deliscious but too bad i ate e...
...	...	...	...	...	...	...	...	...	...	...
568409	568410	B0018CLWM4	A2PE0AGWV6OPL7	Dark Water Mermaid	3	3	5	1309651200	Quality & affordable food	I was very pleased with the ingredient quality...
568410	568411	B0018CLWM4	A88HLWDCU57WG	R28	2	2	5	1332979200	litter box	My main reason for the five star review has to...
568411	568412	B0018CLWM4	AUX1HSY8FX55S	DAW	1	1	5	1319500800	Happy Camper	I bought this to try on two registered Maine C...
568412	568413	B0018CLWM4	AVZ2OZ479Q9E8	Ai Ling Chow	0	0	5	1336435200	Two Siberians like it!	When we brought home two 3-month-old purebred ...
568413	568414	B0018CLWM4	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.



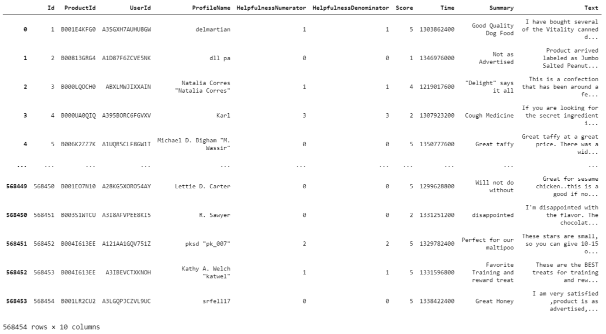
	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	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	ADT0SRK1MGOEU	Twoapennything	0	0	4	1342051200	Nice Taffy	I got a wild hair for taffy and ordered this f...
...	...	...	...	...	...	...	...	...	...	...
558041	558042	B000EM9E2Y	A2RNYW2HIQBGSM	Carlos Alvarez	0	0	1	1230681600	Terrible, don't waste your money	These are just awful. I love all kinds of toa...
558082	558083	B001EO5RRM	A3T1PGY2RO19XB	K. Bauer	1	1	5	1275004800	Delicious!	I was pleasantly surprised when we bought this...
559111	559112	B0057LBU6U	AXXZBEDYDMOA5	onthelookout	0	0	3	1343174400	THINKERS CHICKEN DOG TREAT	My boxer loves this treat. He has not turned d...
559112	559113	B0057LBU6U	A24FX0OC356XUQ	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:

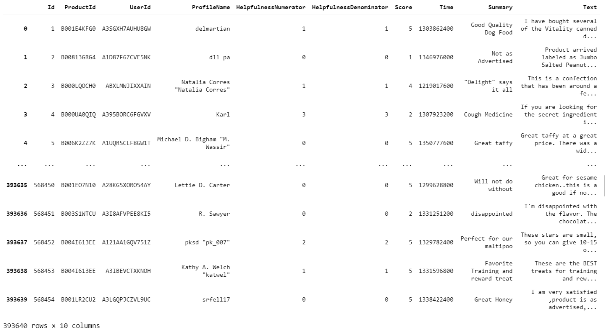


	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	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	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	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	B001EO7N10	A28KG5XORO54AY	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

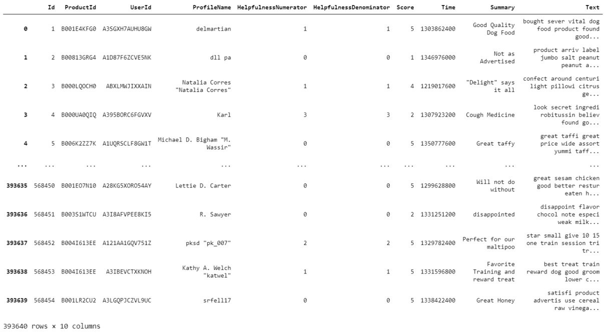
3.2.4 Preprocessing text

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 ge...
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	B001EO7N10	A28KG5XORO54AY	Lettie D. Carter	0	0	5	1299628800	Will not do without	great sesam chicken good better restur eaten h...
393636	568451	B003S1WTCU	A3I8AFVPEE8KI5	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 lower 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 Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

Acknowledgment

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.