A

Project Report

On

APPLICATION OF SUPERVISED LEARNING IN FAKE REVIEW DETECTION

By

Name: Abhik De, Roll No: 30018021026, Reg No: 213001818010026

Name: Anubhab Maity, Roll No: 30018021022, Reg No: 213001818010022

Name: Koyel Chakraborty, Roll No: 30018021035, Reg No:213001818010035

Name: Susnato Chakraborty, Roll No: 30018021004, Reg No: 213001818010004

Submitted in Partial Fulfillment of the Requirement for the Degree of

MASTER OF SCIENCE IN APPLIED STATISTICS

MAULANA ABUL KALAM AZAD
UNIVERSITY OF TECHNOLOGY,
WEST BENGAL

Utech

In Pursuit Of Knowledge And Excellence

DEPARTMENT OF APPLIED STATISTICS

MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY (WBUT)

WEST BENGAL, NADIA- 741249

JUNE, 2022

APPLICATION OF SUPERVISED LEARNING IN FAKE REVIEW DETECTION

Project work submitted by

ABHIK DE, ANUBHAB MAITY, KOYEL CHAKRABORTY,

SUSNATO CHAKRABORTY

Under the Supervision of

Ms. ANWESHA SENGUPTA

Department of Applied Statistics

Maulana Abul Kalam Azad University of Technology (WBUT)

West Bengal, Nadia 741249

DEPARTMENT OF APPLIED STATISTICS

MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY (WBUT)

WEST BENGAL, NADIA- 741249

JUNE, 2022

MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY (WBUT)

WEST BENGAL, NADIA- 741249

CERTIFICATE

I hereby forward this project thesis entitled "**Application of Supervised Learning in Fake Review Detection**" by Abhik De(Roll No: 30018021026 ,Reg No: 213001818010026), Anubhab Maity (Roll No: 30018021022 ,Reg No: 213001818010022), Koyel Chakraborty (Roll No: 30018021035 ,Reg No:213001818010035)**,** Susnato Chakraborty (Roll No: 30018021004 ,Reg No: 213001818010004), of 2021-23 in partial fulfillment of the requirement for the degree of MASTER IN APPLIED STATISTICS AND ANALYTICS of the DEPARTMENT OF APPLIED STATISTICS, MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY (WBUT), WEST BENGAL, NADIA- 741249.

This project thesis has been completed under my guidance in the Department of Statistics,

Maulana Abul Kalam Azad University of Technology (WBUT), West Bengal.

Countersigned:

ANWESHA SENGUPTA

Supervisor

Professor & Head of the Department

Department of Applied Statistics

Maulana Abul Kalam Azad University of Technology

West Bengal, Nadia 741249

ACKNOWLEDGEMENT

Acknowledgment is not simply a ritual. It is an expression of heartfelt gratitude and indebtedness to all those who have been associated with the development of the thesis.

I would like to express my profound and deep sense of gratitude to my guide Mr. MAYUKH BHATTACHARJEE for his unending help, guidance, and suggestions without which this thesis would not have been a reality. He has acted as a friend, philosopher, and guide to me. I own great indebtedness for his untiring effort thoughtful the period of my project work.

I express my sincere thanks to Prof. Anwesha Sengupta, Head of the Department of Applied Science, Maulana Abul Kalam Azad University of Technology, Kalyani, West Bengal, India, for his cooperation extended during the period of my project work.

I am great fully indebted to Prof. Prasanta Narayan Dutta, Course Coordinator, and Prof. Sukhendu Samajdar, Director,Department of Applied Science, Maulana Abul Kalam Azad University of Technology, Kalyani, West Bengal, India, for their great inspiration and encouragement for my work and for providing me a favourable environment in the form of infrastructural facilities for the research work.

I am greatly indebted to all the faculty members, who often took pains and stood by me in adverse circumstances. Without their encouragement and inspiration, it was not possible for me to complete this project.

Finally, my earnest thanks go to my friends who were always beside me when I needed them without any excuses and made these three years worthwhile.

(Abhik De)

(Anubhab Maity)

(Koyel Chakraborty)

(Susnato Chakraborty)

Introduction:

Fake reviews are defined as “deceptive reviews provided with an intention to mislead consumers in their purchase decision making, often by reviewers with little or no actual experience with the products or services being reviewed”. In this study, our objective is to detect fake reviews in online e-commerce sites, like Amazon.

Objective:

This study will be helpful to safeguard customers against fake, corrupt and misleading reviews on e-commerce sites and help them to make better purchase decisions.

Data Source:

Secondary data have been collected from Kaggle. The data consists of 2501 reviews on Amazon (UK Region).

A detailed data view is available at https://www.kaggle.com/datasets/akudnaver/amazon-reviews-dataset.

Methodology:

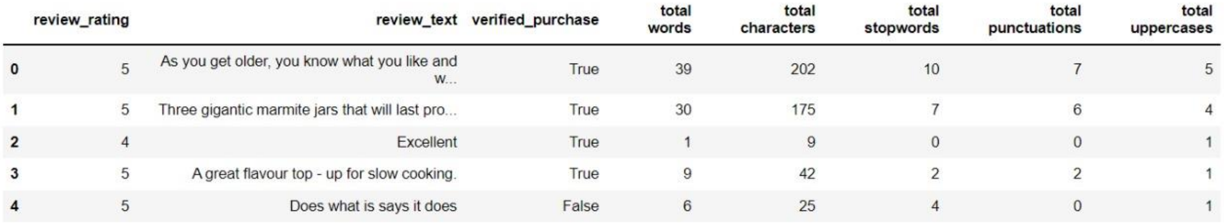
Variables under study:

We have prepared the following columns from the textual review data,

1. Total words, 2. Total characters, 3. Total stopwords, 4. Total punctuations, 5. Total uppercases

These five variables are the independent variables of our study.

We have another column from the data, Verified Purchase (binary) which will be used as dependent variable and our motive is to predict it.



	review_rating	review_text	verified_purchase	total words	total characters	total stopwords	total punctuations	total uppercases
0	5	As you get older, you know what you like and w...	True	39	202	10	7	5
1	5	Three gigantic marmite jars that will last pro...	True	30	175	7	6	4
2	4	Excellent	True	1	9	0	0	1
3	5	A great flavour top - up for slow cooking.	True	9	42	2	2	1
4	5	Does what is says it does	False	6	25	4	0	1

Statistical methods to be used:

We plan to use supervised learning algorithms like Support Vector Machine (SVM) and Logistic Regression to predict whether the reviewer actually purchased the item, i.e., the review is genuine or not. Besides, for a broader understanding, we intend to provide Exploratory Data Analysis (EDA).

- **Logistic Regression:**

Predictive analytics and categorization frequently make use of this kind of statistical model, also referred to as a logit model. Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. Given that the result is a probability, the dependent variable's range is 0 to 1. In logistic regression, the odds—that is, the probability of success divided by the probability of failure—are transformed using the logit formula. The following formulas are used to represent this logistic function, which is sometimes referred to as the log odds or the natural logarithm of odds:

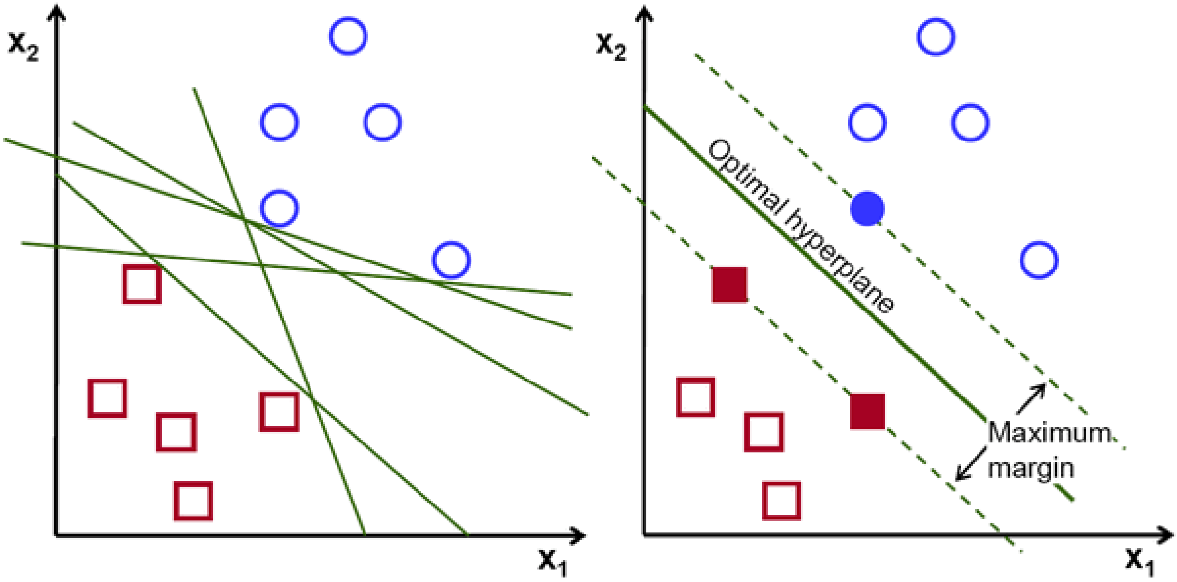
Logit(pi) = 1/(1+ exp(-pi))

ln(pi/(1-pi)) = Beta_0 + Beta_1*X_1 + … + B_k*K_k

Logit(pi) is the dependent or response variable in this logistic regression equation while x is the independent variable. The most frequent method for estimating the beta parameter, or coefficient, in this model is maximum likelihood estimation (MLE). In order to find the best fit for the log odds, this approach iteratively evaluates various beta values. The log likelihood function is created after each of these iterations, and logistic regression aims to maximise this function to get the most accurate parameter estimate. The conditional probabilities for each observation can be calculated, logged, and added together to produce a forecast probability once the best coefficient (or coefficients, if there are multiple independent variables) has been identified. A probability less than.5 will predict 0 in a binary classification, while a probability greater than 0 will forecast 1 before 1. It is recommended to assess the model's goodness of fit, or how well it predicts the dependent variable, once the model has been computed. The Hosmer-Lemeshow test is a well-liked technique for evaluating model fit.

- **SVM:**

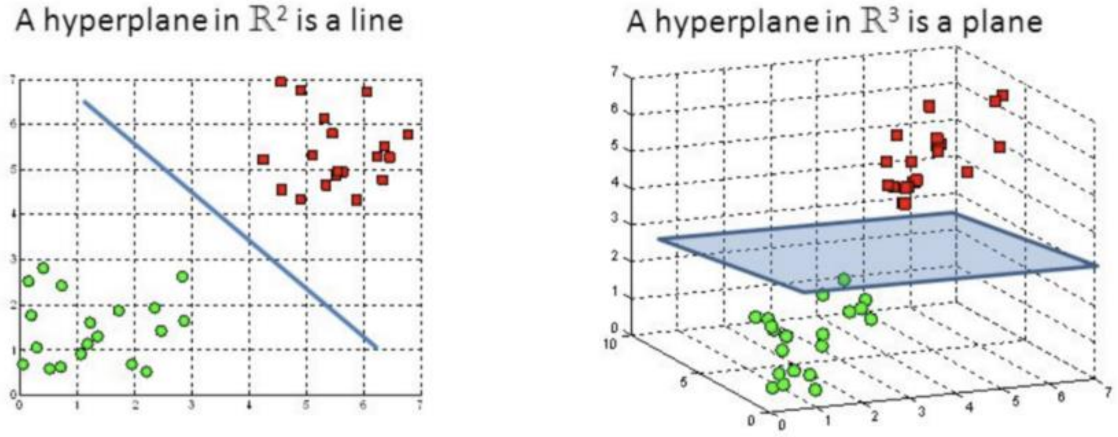
The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.



Possible hyperplanes

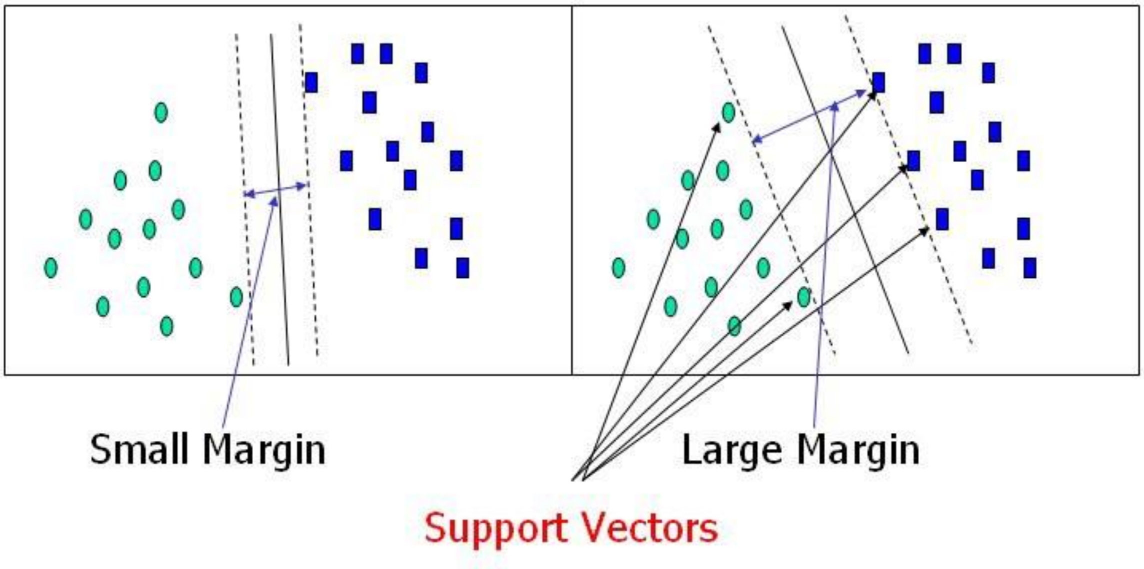
To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Hyperplanes and Support Vectors



Hyperplanes in 2D and 3D feature space

Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3.



Support Vectors

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM.

>Large Margin Intuition

In logistic regression, we take the output of the linear function and squash the value within the range of [0,1] using the sigmoid function. If the squashed value is greater than a threshold value(0.5) we assign it a label 1, else we assign it a label 0. In SVM, we take the output of the linear function and if that output is greater than 1, we identify it with one class and if the output is -1, we identify is with another class. Since the threshold values are changed to 1 and -1 in SVM, we obtain this reinforcement range of values([-1,1]) which acts as margin.

>Cost Function and Gradient Updates

In the SVM algorithm, we are looking to maximize the margin between the data points and the hyperplane. The loss function that helps maximize the margin is hinge loss.

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases} \quad c(x, y, f(x)) = (1 - y * f(x))_+$$

Hinge loss function

The cost is 0 if the predicted value and the actual value are of the same sign. If they are not, we then calculate the loss value. We also add a regularization parameter the cost function. The objective of the regularization parameter is to balance the margin maximization and loss. After adding the regularization parameter, the cost functions looks as below.

$$min_w \lambda \parallel w \parallel^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+$$

Loss function for SVM

Now that we have the loss function, we take partial derivatives with respect to the weights to find the gradients. Using the gradients, we can update our weights.

$$\frac{\delta}{\delta w_k} \lambda \parallel w \parallel^2 \quad = 2\lambda w_k$$

$$\frac{\delta}{\delta w_k} \big(1 - y_i \langle x_i, w \rangle\big)_+ \quad = \begin{cases} 0, & \text{if } y_i \langle x_i, w \rangle \geq 1 \\ -y_i x_{ik}, & \text{else} \end{cases}$$

Gradients

When there is no misclassification, i.e our model correctly predicts the class of our data point, we only have to update the gradient from the regularization parameter.

$$w = w - \alpha \cdot (2\lambda w)$$

Gradient Update — No misclassification

When there is a misclassification, i.e our model make a mistake on the prediction of the class of our data point, we include the loss along with the regularization parameter to perform gradient update.

$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w)$$

Gradient Update — Misclassification

- **Analysis:**

There were 32 columns in the initial dataset that was downloaded from Kaggle, which was more than adequate. Further study revealed that the dataset's information was more concerned with the product and the reviews than it was with the reviewers themselves, leading to the conclusion that this project will employ linguistics rather than behavioural analysis to tackle the challenge of identifying phoney reviews.

The 30 additional columns in this dataset, while a bit overkill for the needs of our project, helped us grasp the context of the reviews and the items in-depth, giving us a wealth of information to draw upon when we analyse our findings later in the project. Therefore, we created graphs and analyses before removing the 30 columns.

- **DATA DESCRIPTION**

By examining the columns more closely, we can better understand the dataset. Below is a detailed description of the columns.

1. report_date:when the data was first extracted. At first glance, it appears that the majority were gathered in 2019.

2. online_store:The brand of the shop where these reviews were posted.

3. upc:A barcode symbology called the Universal Product Code (UPC) is frequently used to track trade goods in stores all around the world. A 12-digit UPC that is specific to each trade item is given to it.

4. retailer_product_code:product number from the perspective of the store.

5. brand:the name of the item that is being offered.

6. category:the product's broad category, such as foods

7. sub_category:the category that provides context for the category The item falls under: food category, savoury subcategory

8. product_description:detailed product description to give additional information about the product

9. review_date:time the review was published

10. review_rating:the product reviewer's evaluation of it. determines the review's total rating, which is given to the ostensibly purchased product. From 1 to 5, where 1 is very poor and 5 is great.

11. review_title:the name given to the written review.

12. review_text:the review itself, outlining the product that the buyer allegedly purchased.

13. is_competitor:whether or whether the product is a rival. To fully comprehend this column, more research needs be done.

14. manufacturer:the company that created the goods.

15. market:where these items are located and where the stores are. A closer examination is required to fully comprehend this.

16. matched_keywords:This cannot be inferred from the dataset because it appears that all of the values in this column are NULL.

17. time_of_publication:Since all of the values in this column seem to be NULL, it is impossible to infer this information from the dataset.

18. url:URL of the actual review.

19. review_type:identifies the sort of review. a search for more details

20. parent_review:shows whether the review is coming from a parent or a youngster.

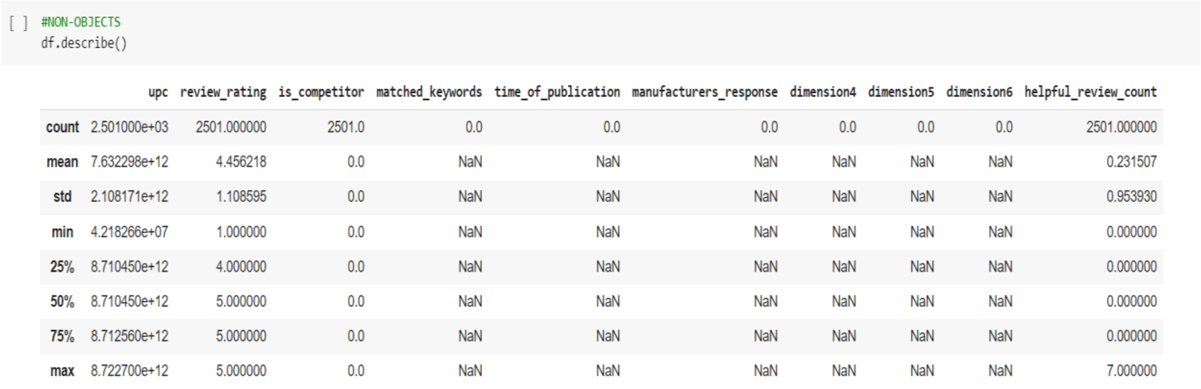
21. manufacturers_response:

22. dimension 1 - 8:many details about the product itself. These columns have a number of missing values, which will be confirmed later.

23. verified_purchase: if Amazon's technology has verified the written reviews or not.

24. helpful_review_count:how many people thought the review was useful

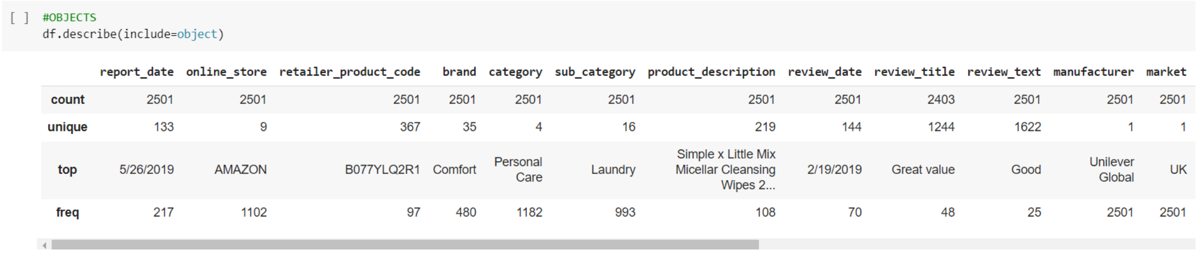
25. review_hash_id:review's special identification number.



```
[ ] #NON-OBJECTS
    df.describe()
```

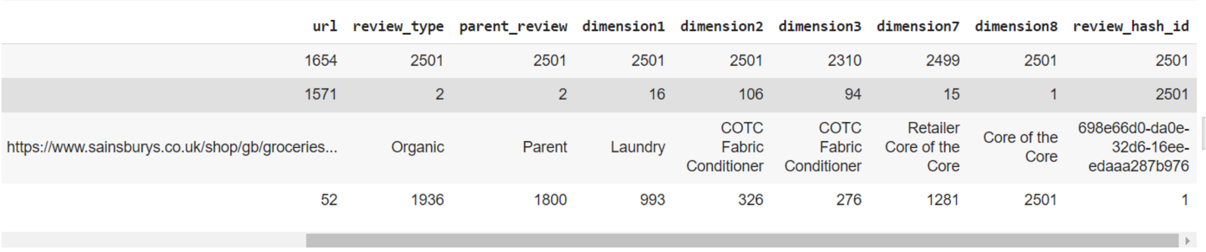
	upc	review_rating	is_competitor	matched_keywords	time_of_publication	manufacturers_response	dimension4	dimension5	dimension6	helpful_review_count
count	2.501000e+03	2501.000000	2501.0	0.0	0.0	0.0	0.0	0.0	0.0	2501.000000
mean	7.632298e+12	4.456218	0.0	NaN	NaN	NaN	NaN	NaN	NaN	0.231507
std	2.108171e+12	1.108595	0.0	NaN	NaN	NaN	NaN	NaN	NaN	0.953930
min	4.218266e+07	1.000000	0.0	NaN	NaN	NaN	NaN	NaN	NaN	0.000000
25%	8.710450e+12	4.000000	0.0	NaN	NaN	NaN	NaN	NaN	NaN	0.000000
50%	8.710450e+12	5.000000	0.0	NaN	NaN	NaN	NaN	NaN	NaN	0.000000
75%	8.712560e+12	5.000000	0.0	NaN	NaN	NaN	NaN	NaN	NaN	0.000000
max	8.722700e+12	5.000000	0.0	NaN	NaN	NaN	NaN	NaN	NaN	7.000000

1. The average evaluation in this dataset is favourable, as seen by the high rating of 4. 4. A score of 1 is considered to be very poor, and a score of 5 is considered to be exceptional.
2. The low average of 0.2 indicates that the majority of the reviews in this dataset did not receive a helpful review count.
3. Because a UPC is a unique identifier, it will not be included in these statistics.
4. The is competitor column shows that there are no values, which means that nothing has been designated as a competitor. This is supported by the fact that the average value is 0 and both the minimum and maximum values are 0.
5. All of the data for matched keywords, time of publication, manufacturers response, dimension4, dimension5, and dimension6 are NULL.



```
[ ] #OBJECTS
    df.describe(include=object)
```

	report_date	online_store	retailer_product_code	brand	category	sub_category	product_description	review_date	review_title	review_text	manufacturer	market
count	2501	2501	2501	2501	2501	2501	2501	2501	2403	2501	2501	2501
unique	133	9	367	35	4	16	219	144	1244	1622	1	1
top	5/26/2019	AMAZON	B077YLQ2R1	Comfort	Personal Care	Laundry	Simple x Little Mix Micellar Cleansing Wipes 2...	2/19/2019	Great value	Good	Unilever Global	UK
freq	217	1102	97	480	1182	993	108	70	48	25	2501	2501



	url	review_type	parent_review	dimension1	dimension2	dimension3	dimension7	dimension8	review_hash_id
	1654	2501	2501	2501	2501	2310	2499	2501	2501
	1571	2	2	16	106	94	15	1	2501
	https://www.sainsburys.co.uk/shop/gb/groceries...	Organic	Parent	Laundry	COTC Fabric Conditioner	COTC Fabric Conditioner	Retailer Core of the Core	Core of the Core	698e66d0-da0e-32d6-16ee-edaaa287b976
	52	1936	1800	993	326	276	1281	2501	1

1. The dataset contains a total of 9 retailers, the majority of which are from the AMAZON store. Major UK outlet retailers make up the majority of the stores there.

2. The Personal Care area of the Comfort brand's Laundry section is where the majority of the data is found.

3. The reviews with the majority of favourable ratings can be identified by the words "Good" and "great value" that are placed in the review text, suggesting that the terminology used in the positive reviews is consistent. can be confirmed as we learn more about it.

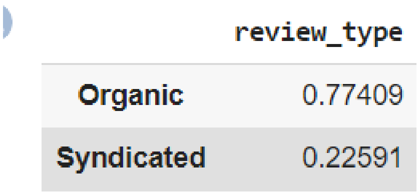
4. Throughout the dataset, one manufacturer caught my attention: Unilever Global

5. The market is limited to the UK.

6. There are two categories of reviews, with parent and organic reviews making up the majority of each.

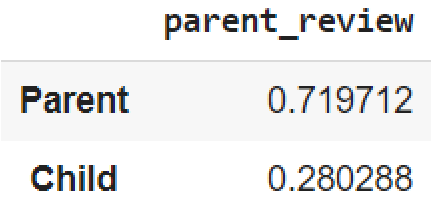
7. As validated, dimensions 1–8 display additional product-specific information. The majority of them are laundry-related items.

8. review hash id can be disregarded because it is a distinctive identifier.



	review_type
Organic	0.77409
Syndicated	0.22591

The majority of reviews (77%) are natural; the lowest percentage of syndicated evaluations, or reviews of the same product disseminated across numerous platforms for greater accessibility, is 22%.



	parent_review
Parent	0.719712
Child	0.280288

The majority of the reviews are allegedly written by parents. Further investigation revealed that this does not clearly state what it stands for, and because of this, it may be dropped.The subcategories are arranged in accordance with the categories.

A deeper look reveals that some of the categories overlap, are written differently even if they refer to the same thing (e.g., Hair and Hair care, Deos and Deodorants & Fragrances), or have the same subcategories but different categories (e.g., Ice Cream under Refreshment and Foods). since handling this will be beneficial for future EDA analysis

As was previously noted, this dataset contains NULL values that need to be handled.

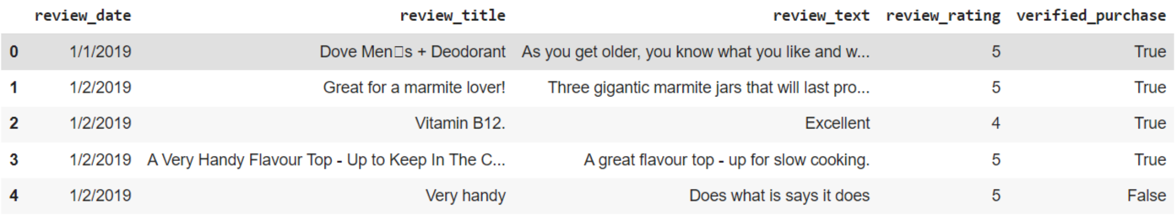
1. Out of almost 2k records, review_title has 98 NULL values, which is a minority.

2. The URL has more than 800 missing values, however this is unimportant given the nature of our project and may be overlooked.

3. Because they don't contain any data, the fields matched_keywords, time_of_publication, manufacturers_response, dimension4, dimension5, and dimension6 will most likely be removed.

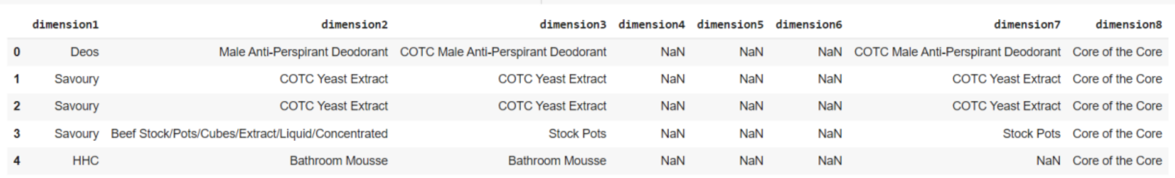
4. Dimension 7 has two missing values, and it is clear from the image above that it only carries further product information.

There aren't any behavioural characteristics in the reviews themselves. Because there is no information about the author, the date the review was published, or the number of reviews the author has written, behavioural context cannot be used to identify phoney reviews of this project.



	review_date	review_title	review_text	review_rating	verified_purchase
0	1/1/2019	Dove Men□s + Deodorant	As you get older, you know what you like and w...	5	True
1	1/2/2019	Great for a marmite lover!	Three gigantic marmite jars that will last pro...	5	True
2	1/2/2019	Vitamin B12.	Excellent	4	True
3	1/2/2019	A Very Handy Flavour Top - Up to Keep In The C...	A great flavour top - up for slow cooking.	5	True
4	1/2/2019	Very handy	Does what is says it does	5	False

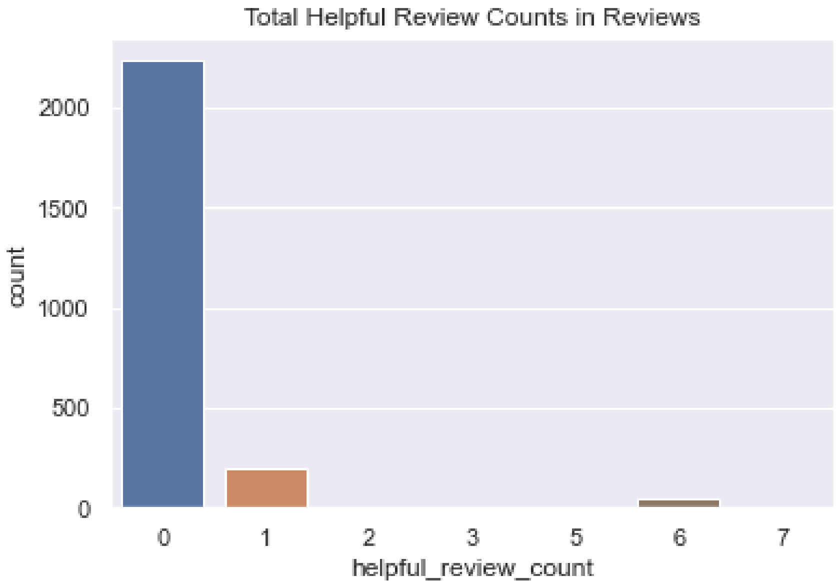
Textual attributes can be used to understand the textual context of these reviews because they are present in this dataset and because, as shown above, they do not contain any NULL values. As a result, by the time we are building the classifier, we will have a better understanding of the review data itself.



	dimension1	dimension2	dimension3	dimension4	dimension5	dimension6	dimension7	dimension8
0	Deos	Male Anti-Perspirant Deodorant	COTC Male Anti-Perspirant Deodorant	NaN	NaN	NaN	COTC Male Anti-Perspirant Deodorant	Core of the Core
1	Savoury	COTC Yeast Extract	COTC Yeast Extract	NaN	NaN	NaN	COTC Yeast Extract	Core of the Core
2	Savoury	COTC Yeast Extract	COTC Yeast Extract	NaN	NaN	NaN	COTC Yeast Extract	Core of the Core
3	Savoury	Beef Stock/Pots/Cubes/Extract/Liquid/Concentrated	Stock Pots	NaN	NaN	NaN	Stock Pots	Core of the Core
4	HHC	Bathroom Mousse	Bathroom Mousse	NaN	NaN	NaN	NaN	Core of the Core

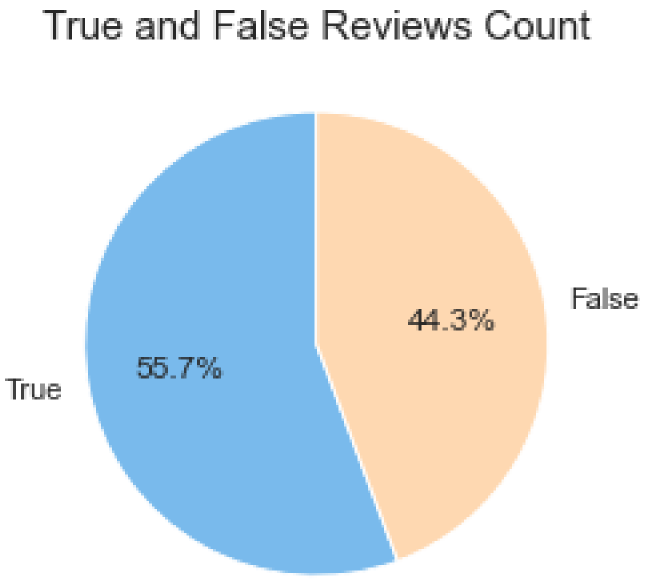
The remaining characteristics are regarded as supplemental data that have helped us comprehend the context of both the evaluations and the actual items. For instance, dimensions 1 through 8 provided additional product-specific information that, while it helped us comprehend the context of the product, was not useful in separating out fraudulent evaluations from genuine ones.

EXPLORATORY DATA ANALYSIS:

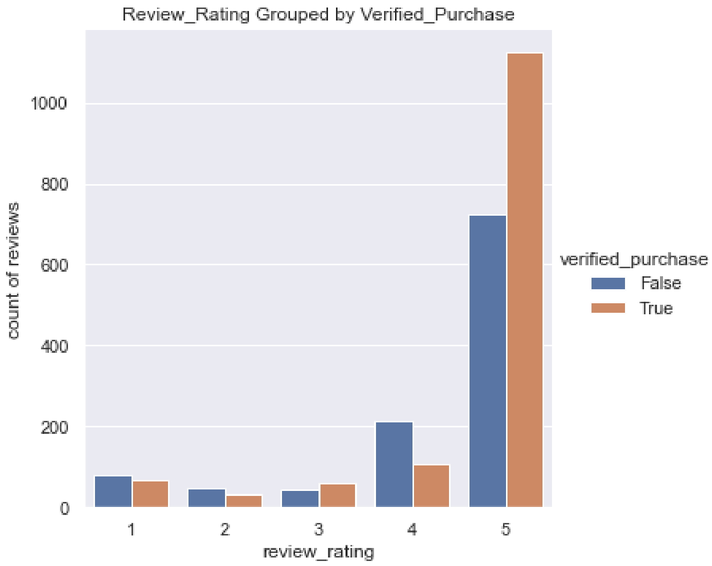


helpful_review_count can assist us determine which reviews have influenced consumers' purchasing of a product on a regular basis. The majority of the evaluations in this dataset, however, have a count of 0 helpful reviews, while some have a meagre number of helpful reviews. In this instance, this will actually distort our understanding of how to distinguish between false and legitimate reviews, thus in order to remove bias, this column will not be taken into account while creating our model.

The goal variable for this project is the Verified_purchases column. The countplot beside shows that the proportions of true and false VP are about similar (56% and 44%, respectively).

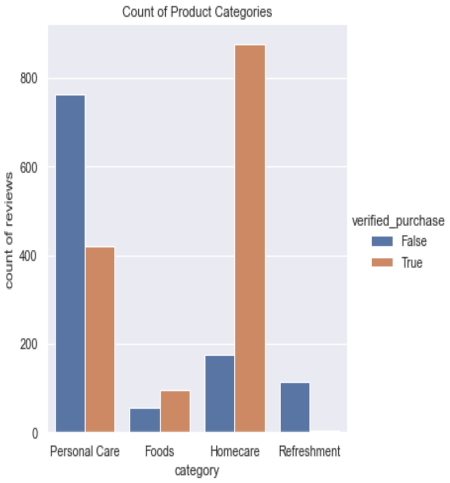


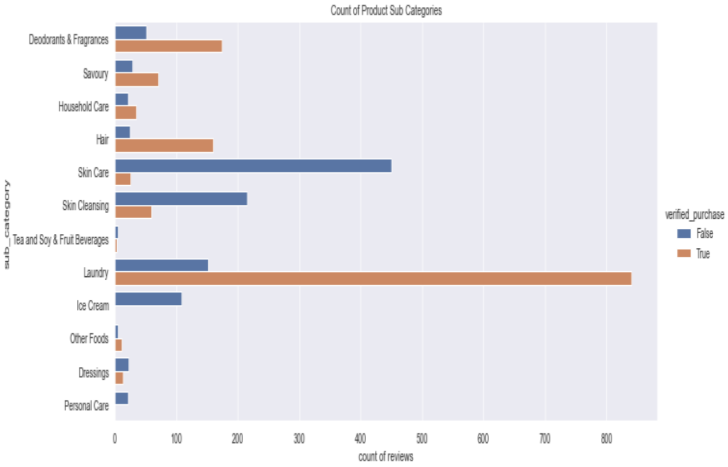
By introducing this column, Verified Purchases, where reviewers must go through a number of verification processes to confirm that the review they are posting has actually been purchased from the site, Amazon has presented its response to counteract phoney reviews. As a result, because the review was written after the product was purchased, there is security about the veracity of the reviews thanks to Amazon's response to false reviews.



Fake reviews, according to earlier literature, are those that disparage or encourage a product without having actually used the service or product in question.

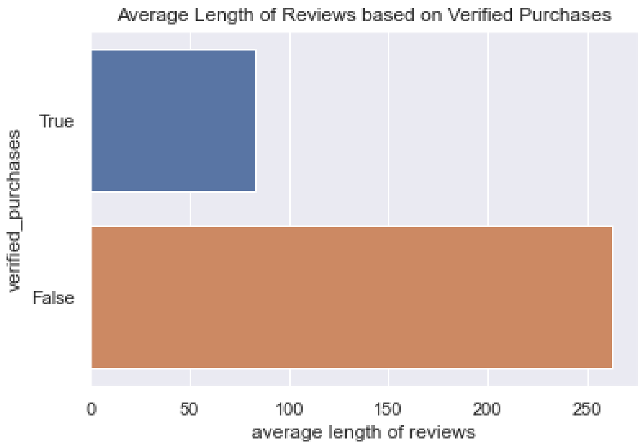
If you look at the graph, you can see that False VP has the highest percentage of 5 ratings compared to its 1 rating, which shows that phoney reviews have been utilised to promote the product without the customer's actual purchase. This is typically done to raise the product's ratings. Additionally, the True Vp has a higher rating of 5 than the False, as can be seen. When we look at the 1 star rating, the False VP is more than the True VP, which can suggest that the reviews were attempting to downgrade the said product—again, without a prior purchase—is clear from the False VP.The majority of the ratings in this dataset are positive overall, which should be taken into consideration as we move forward with the model building.





Looking further into the categories, we can observe that the dressings, skin care, skin washing, and personal care subcategories have the highest percentage of unverified reviews. With the exception of dressings, which are classified under the Foods division, the most of them belong within the Personal care category.

According to the graph above, laundry is the subcategory that belongs under the Home care category because it has the most purchases there as well as the most confirmed transactions.



One of the most important conclusions to be drawn from this graph is that the average length of the incorrect values actually exceeded the ones that are marked as confirmed. This is supported by a number of postings, particularly one on a website that specialises in spotting fake reviews, where it was noted that the length of these evaluations tends to be longer on average than that of real reviews.

Here, we can observe that the average length of a bogus VP is more than 250 characters, whereas the average length of a legitimate VP is between 50 and 100 characters.

We have finished using the other features to deepen our understanding of the dataset, therefore it is time to think about what will be required for model creation. As previously said, we are using linguistics to distinguish between genuine and fraudulent reviews, thus we will only preserve the data for the review text, review date, review rating, review title, and verified purchase that are directly related to the reviews' nature.

Verified purchases will help us determine which value is the ground false value that can be used to train the classifier. This is done because we are attempting to extract the features from the review text. In the following stage of data pre-processing, where we will concentrate more on the review text itself and less on the other attributes, the additional review-related columns will help us.

1. The following fields should be dropped: matched keywords, time of publication, manufacturers response, dimension4, dimension5, and dimension6 all contain NULL values.

2. According to the summary statistics, the variables is competitor and helpful review count have very low or zero values. As a result, they do not assist us understand the dataset or create models, thus they should be removed.

3. The columns report date, online store, brand, category, sub category, and market have helped us grasp the context of the data and its source, but they will not be of any further use to us in developing models.

4. Because they are unique identifiers, the upc, retailer product code, review hash id, and url will not help us develop models.

5. Additional information on the product itself is provided in the form of the product description, parent review, review type, manufacturer, dimension1, dimension2, dimension3, dimension4, and dimension5 columns. However, because we are not conducting our research from the perspective of the product, these details are useless and will not help us.

The fact that the duplicate reviews were not found in our first EDA using the complete dataset may have been caused by the other columns' slightly differing values. We need to confirm whether there are duplicate reviews within the dataset so that we can remove them appropriately to remove potential bias since we have removed the other columns and are left with only the review centric values.

Removing Duplicates

We initially attempted to identify duplicates within this dataset during the initial data exploration in the old csv file. However, it did not provide any results at first. The chosen columns were added to the dataset to help determine whether there are genuinely any duplicates present in order to double-check that there aren't any.

Since the backdrop of the Amazon dataset has already been thoroughly examined, this time the review text will also be examined in more detail. A few columns will be added to help us comprehend some of the instances the sentences have throughout our pre-processing. There are also counts of:

Word

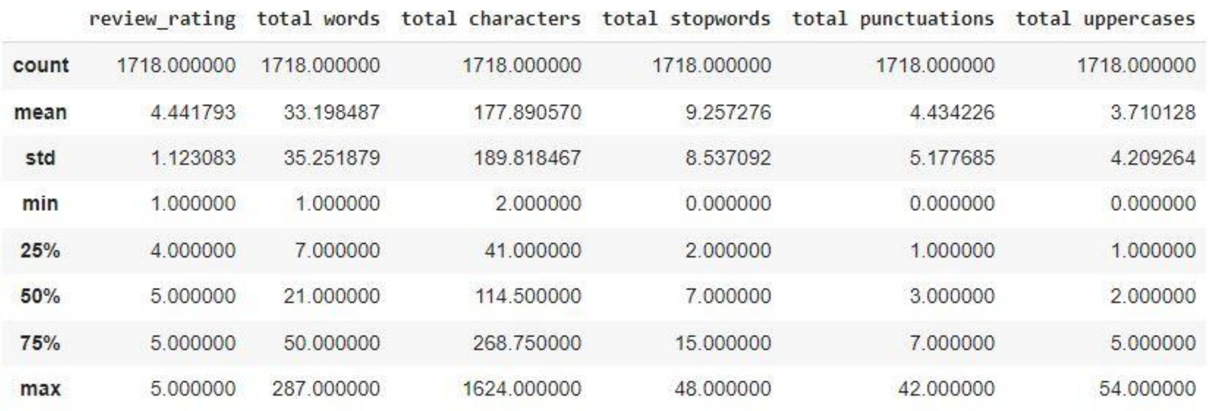
Characters (with spaces) (with spaces)

Stopwords

Punctuations

capitalised characters

Following the addition of the columns, the relevant summary statistics will be run to determine how the pre-processing will proceed.



	review_rating	total words	total characters	total stopwords	total punctuations	total uppercases
count	1718.000000	1718.000000	1718.000000	1718.000000	1718.000000	1718.000000
mean	4.441793	33.198487	177.890570	9.257276	4.434226	3.710128
std	1.123083	35.251879	189.818467	8.537092	5.177685	4.209264
min	1.000000	1.000000	2.000000	0.000000	0.000000	0.000000
25%	4.000000	7.000000	41.000000	2.000000	1.000000	1.000000
50%	5.000000	21.000000	114.500000	7.000000	3.000000	2.000000
75%	5.000000	50.000000	268.750000	15.000000	7.000000	5.000000
max	5.000000	287.000000	1624.000000	48.000000	42.000000	54.000000

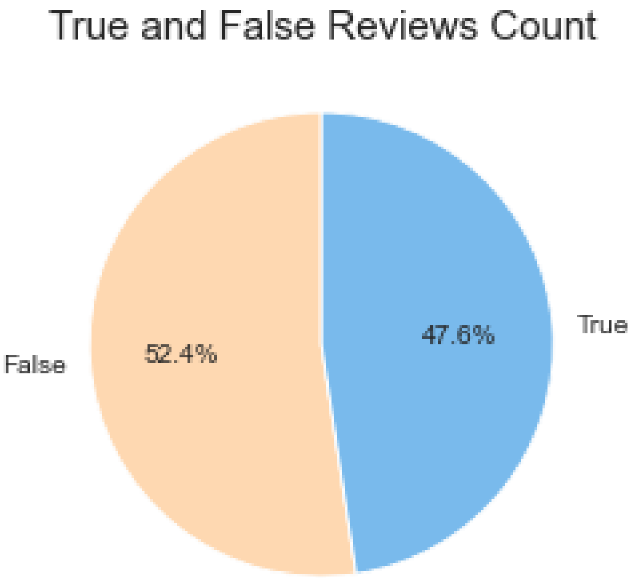
1)The average number of characters across the entire dataset is 177, or roughly 33 words per review.

2)There are typically nine stop words and four punctuations within each phrase.

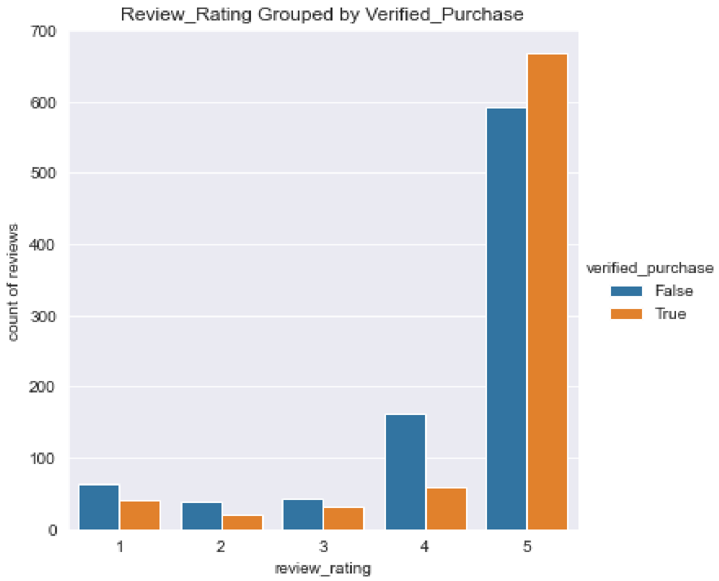
3)From the mean number, it is safe to deduce that the majority of reviews used capital letters as sentence case.

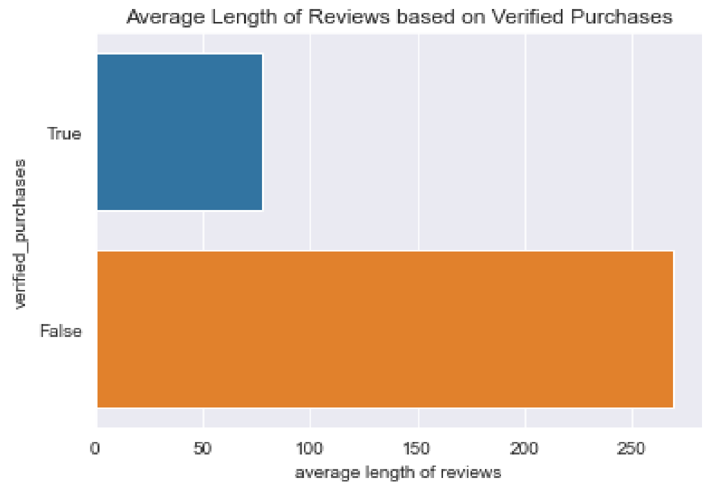
After Group of Verified Purchase Column, The findings are

1) Overall, we can observe that False values have an average of 50 words and 268 characters, while True values have an average of 14 words and 77 characters each review. This means that False values have more words per character than True Values.
2) It is noticeable that there are more stopwords in the fake reviews as well as the genuine evaluations.
3) It is clear that there are more punctuation marks and sentence cases in False values than in True values since the sentences are lengthier.



We may argue that the dataset is balanced since, even after eliminating the duplicates, the percentages of True and False values are still very close to being equal. A closer inspection of the graph reveals that the dataset now contains more False values and True values.





As we can see from the image above, the highest rating is still a 5, and honest reviews continue to outweigh fake ones. Within this dataset, sentiment is still overwhelmingly positive.

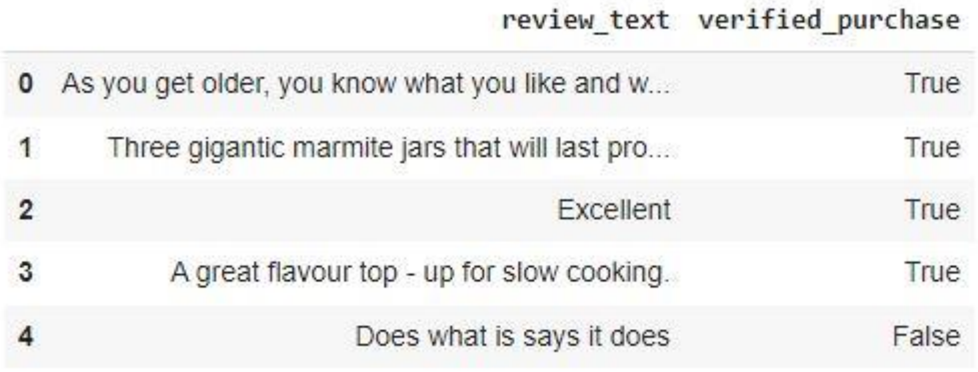
PRE-PROCESSING

Text preprocessing is a method for decluttering and getting ready text data for a model. Text data includes, among other things, noise in the form of punctuation, emotions, and text written in a different case. There are numerous methods to communicate the same idea in human language, but this is simply the beginning of the challenge. Machines can't understand words; they only understand numbers, thus we need to effectively transform text to numbers.

We can see from the summary statistics that the noise stated occurs in the review text, thus pre-processing will be done in accordance with that.

To Do

1) Remove unnecessary columns
2) Lowercasing
3) Eliminate Stopwords
4) Eliminate punctuation and any special characters.
5) Stemming



	review_text	verified_purchase
0	As you get older, you know what you like and w...	True
1	Three gigantic marmite jars that will last pro...	True
2	Excellent	True
3	A great flavour top - up for slow cooking.	True
4	Does what is says it does	False

We will just use review text and verified purchase as our classifiers for the time being.

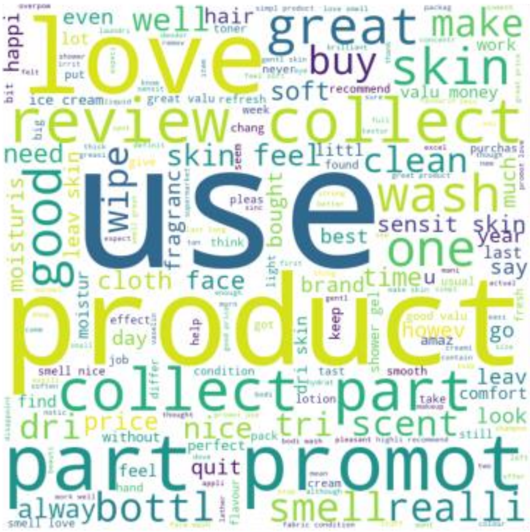
Text Pre-Processing

So that the model may be applied and optimized to its full potential, the review text will be cleaned and standardized. As it is based on trial and error, this process takes the longest.

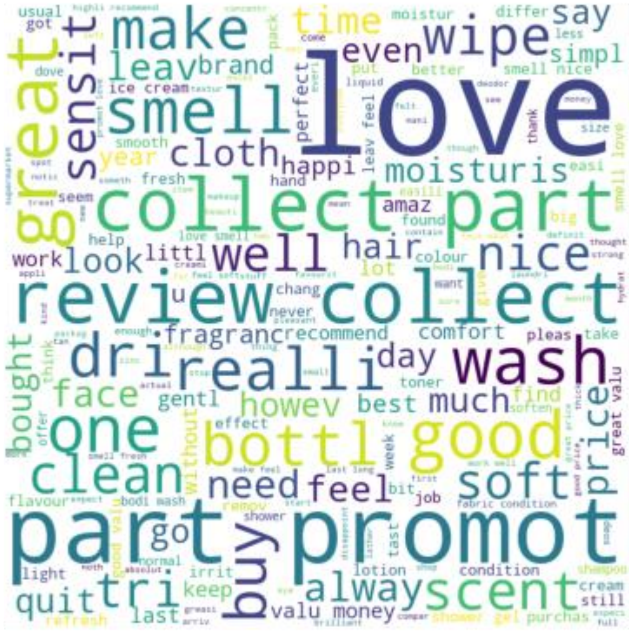
PERFORMED AT THIS STAGE:

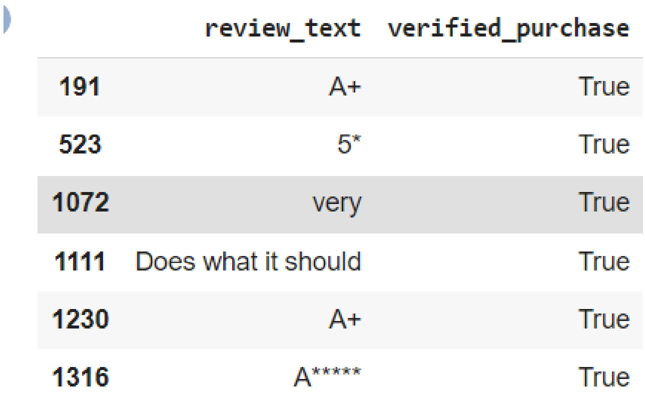
Tokenization is used to improve the spelling and remove stop words, punctuation, and special characters. Eliminating the top three common and uncommon terms and lowercasing them.

Generate Word cloud on clean review data



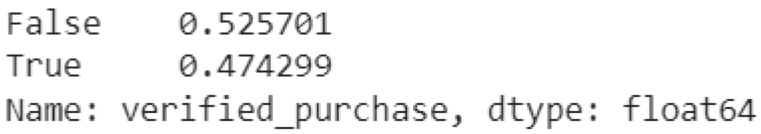
We are left with the current top 10 words after deleting the top 3 common words (which were deleted because doing so would render the entire list meaningless). As can be observed from the image above, the sentiment is largely positive, indicating that this dataset has a lot of reviews that are favourable in nature. Thus, the overall polarity is positive and should be kept in mind for further research. It should be noted that the absence of negative reviews in this instance can lead to inconsistencies when, for example, a negative value is set to be identified as "false" or "genuine," and can therefore be listed as a limitation to this study.





	review_text	verified_purchase
191	A+	True
523	5*	True
1072	very	True
1111	Does what it should	True
1230	A+	True
1316	A*****	True

When looking at the original csv file, it is clear why the review text for the five rows was completely removed. That's because only words with meaning were saved during the previous step of text processing, and the second table shows that the majority of those words were either stopwords or contained symbols and numbers. These will be removed later because they are meaningless either way.



```
False     0.525701
True      0.474299
Name: verified_purchase, dtype: float64
```

The T/F numbers were little affected by the alteration, thus we may move further.

There are presently just two columns in the dataset. Review text will be used as the input variable and verified purchases as the target variable out of the two. The data will subsequently be divided appropriately.

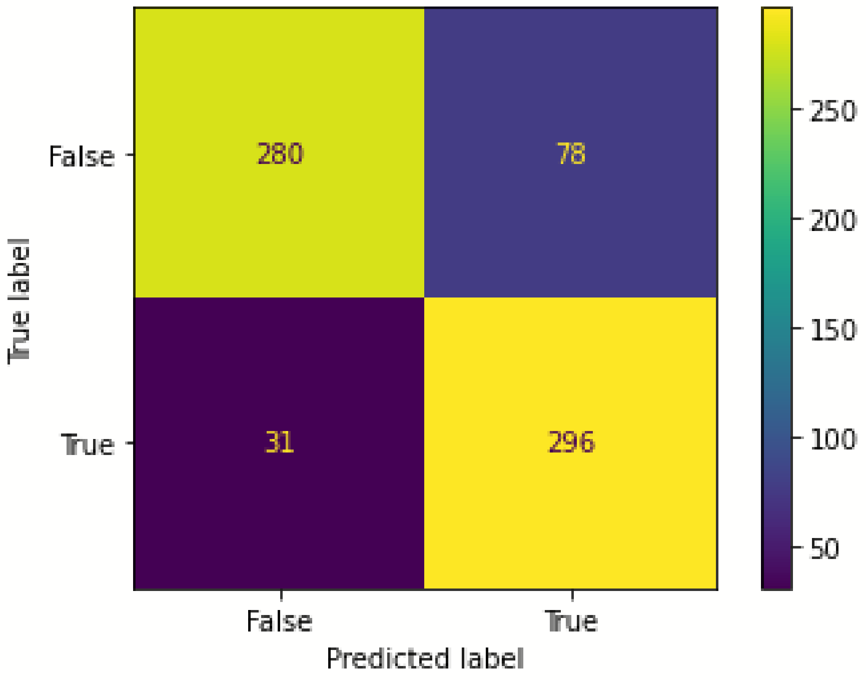
Trial and error led to the decision to divide the data into 60 and 40 groups. We are going to use this splitting because it gives the models the best accuracy.

COUNT VECTORIZER AND MODELING:

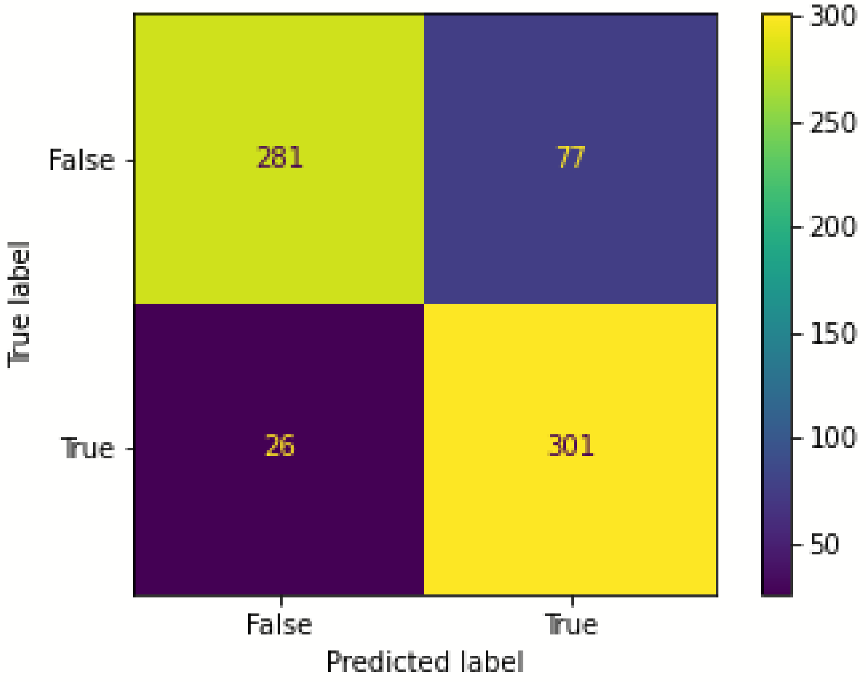
Due to the fact that models can only comprehend numerical input, word vectorization converts words or phrases from a lexicon into a matched vector of real numbers that may be used to derive word predictions and semantics.

Two vectorization techniques will be used, the first of which is the count vectorizer. In CountVectorizer, we merely count the number of times a word appears in the document, which favours the most prevalent terms.

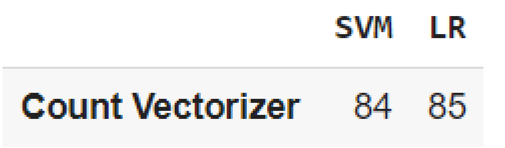
Confusion matrix for Support Vector Machine model:



Confusion matrix for Logistic Regression model:

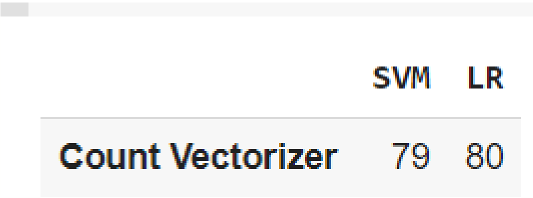


COMPARING ACCURACY



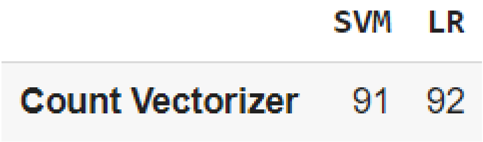
	SVM	LR
Count Vectorizer	84	85

COMPARING PRECISION



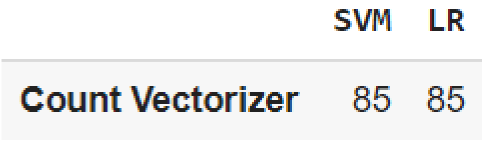
	SVM	LR
Count Vectorizer	79	80

COMPARING RECALL



	SVM	LR
Count Vectorizer	91	92

COMPARING F1 SCORE



	SVM	LR
Count Vectorizer	85	85

Conclusion:
The best has been determined to be LR with count vectorizer after careful consideration. The accuracy of the remaining models has all been above 80%, and the other measures have all been above 79%.