Harnessing Helpfulness: Amazon Product Reviews to Elevate Marketplace Experience

DATA 270
Under guidance of Dr. Linsey Pang

Group 5

Presented by:

Eshita Gupta
Monica Lokare
Sneha Karri
Veena Ramesh Beknal

Data Process

Approaches and Steps to Enable Data for Modeling

We used the Amazon dataset, from 1996 onwards. We focused on using data ranging from 2010 to 2018.We restricted to product categories that were objective like appliances, automotive, cell phones and accessories, and tools and home improvement. Our project objective can be achieved based on this subset of the dataset.

Reason for choosing this range:

During this period there was a shift in the consumer behavior where consumers relied on the reviews before buying the products. Also, consumers started giving feedback on the products.

Pre-processing & Transformation methods for modeling

- Data processing and transformation helps us in improvising the quality of the data which then can be used in modeling.
- After converting the raw JSON to csv, we performed data cleaning by removing duplicates, standardizing date formats, imputing missing values, checking non numeric values, plotting graphs to understand the data distribution, and creating helpfulness binary.

Deriving Training, Validation and Test datasets

For transformation we used stratified samples of 50,000 whose distribution is a representative of overall data for each category. We performed transformation on textual contents from reviewTitle (summary) and reviewText.

- The steps involved determining number of words in the review, review length, analysis of sentiment polarity, identifying active reviewers, calculating age of the review.
- Additionally, we also performed tokenization, stop word removal, and lemmatization.
- TF-IDF vectorizer was created and distribution of top n-gram was visualized.
- We have prepared our data for modeling by dividing our cleaned and transformed dataset into 70 % training dataset and 15 % for testing, and 15% for validation.

Data Collection

Data Collection

- For our project, we are using Amazon Review Data (2018). The dataset contains a total of 233.3 million reviews, and the categories we will be using for our project include Appliances, Automotive, Cell Phones and Accessories, and Tools and Home Improvement.
- These categories are selected to provide a broad range of insights from different types of consumer products available on Amazon.
- We can use this data that will help sellers to improve content of the listing and make inform marketing strategies and improve customer service.
- Table includes the data collection plan for our project shown in the next slide.

Data Collection plan for all the four categories

Questions	Dataset	Reason
Why are we gathering this information?	Appliances category	We are gathering information to gain insights into how satisfied our customers are with our products. We want to address any common issues they may be facing and get a sense of how they feel about our appliances overall. This data will be used to make improvements to our products, enhancing customer service, and creating more effective marketing strategies.
	Automotive category	We are collecting information to analyze customer opinions on automotive category available for purchase on Amazon. This involves evaluating customer satisfaction, recognizing typical problems or flaws, and understanding market trends. Our main objective is to improve product quality, customer support, and to inform potential buyers.
	Cell Phones and Accessories category	We are gathering information to understand how customers feel about cell phones and accessories. Our goal is to identify trends, issues, and features that are most important to customers when it comes to these products.
	Tools and Home Improvement category	We are collecting information to better understand customer feedback on tools and home improvement goods available at Amazon. This will help us in identifying trends, features, customer issues, and overall satisfaction with these products.

Data Collection plan (Contd.)

Questions	Dataset	Reason
How the information may be useful?	All the four categories	Analyzing the reviews has many benefits. Sellers can learn about how to improve their products, new products can be designed, and help Amazon to improve categorization and product recommendations. Customers will also be able to make better purchasing decisions by considering the feedback from other users.
What should we do after collecting the data?	All the four categories	Once data is collected, it is important to clean and prepare the data for analysis in order to maintain its quality. This involves eliminating duplicate values, fixing errors, and normalizing text formats. The next step is to analyze the data for patterns, emotions, and helpfulness of reviews. After studying the data, the findings should be communicated to the appropriate parties (such as product teams, customer support, and marketing) for further improvements.
Historical data		Yes, Amazon keeps track of all the historical reviews that customers leave for products. Our goal is to gather feedback from the moment a product is listed all the way, allowing customers to give feedback or opinions of the products.
Operational definition exists?		We will create operational definitions to make sure the data is consistent. For example, we will determine "review helpfulness" by looking at the number of helpful votes compared to total votes.
Duration		8 years of data

Samples of raw data resources

Sample Dataset of Appliances category

	overall	verified	reviewTime	reviewerID	asin	style	reviewerName	reviewText	summary	unixReviewTime	vote	image
0	5	True	08 22, 2013	A34A1UP40713F8	B00009W3I4	{'Style:': ' Dryer Vent'}	James. Backus	I like this as a vent as well as something tha...	Great product	1377129600	NaN	NaN
1	5	True	02 8, 2016	A1AHW6I678O6F2	B00009W3PA	{'Size:': ' 6-Foot'}	kevin.	good item	Five Stars	1454889600	NaN	NaN
2	5	True	08 5, 2015	A8R48NKTGCJDQ	B00009W3PA	{'Size:': ' 6-Foot'}	CDBrannom	Fit my new LG dryer perfectly.	Five Stars	1438732800	NaN	NaN
3	5	True	04 24, 2015	AR3OHHHW01A8E	B00009W3PA	{'Size:': ' 6-Foot'}	Calvin E Reames	Good value for electric dryers	Perfect size	1429833600	NaN	NaN
4	5	True	03 21, 2015	A2CIEGHZ7L1WWR	B00009W3PA	{'Size:': ' 6-Foot'}	albert j. kong	Price and delivery was excellent.	Five Stars	1426896000	NaN	NaN

Sample Dataset of Automotive category

	overall	verified	reviewTime	reviewerID	asin	style	reviewerName	reviewText	summary	unixReviewTime	vote	image
0	4	False	05 1, 2015	A8WEXFRWX1ZHH	0209688726	{'Color:': ' AC'}	Goldengate	After I wrote the below review, the manufactur...	Works well if you place phone in horizontally ...	1430438400	NaN	NaN
1	1	True	04 19, 2018	ABCA1A8E4DGV1	0209688726	{'Color:': ' Blue'}	noe	It sucks barely picks up anything definitely n...	sucks	1524096000	NaN	NaN
2	1	True	04 16, 2018	A1NX8HM89FRQ32	0209688726	{'Color:': ' Black'}	Eduard	Well to write a short one, it blew 2 fuses of ...	Defective	1523836800	NaN	NaN
3	3	True	04 13, 2018	A1X77G023NY0KY	0209688726	{'Color:': ' CA'}	Lauren	I have absolutely no memory of buying this but...	Looks cool! Probably works	1523577600	NaN	NaN
4	5	True	04 8, 2018	A3GK37JO2MGW6Q	0209688726	{'Color:': ' Black'}	danny	it ok it does it job	Five Stars	1523145600	NaN	NaN

Samples of raw data resources (Contd.)

Sample Dataset of Cell Phones and Accessories category

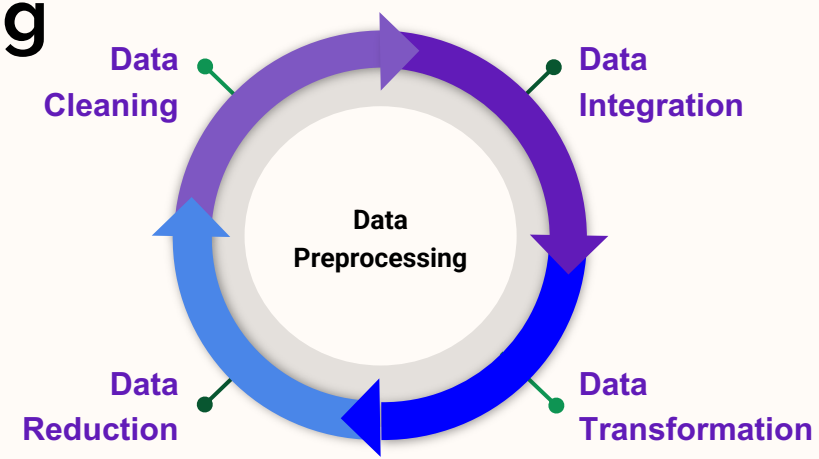
	overall	verified	reviewTime	reviewerID	asin	style	reviewerName	reviewText	summary	unixReviewTime	vote	image
0	5	True	08 4, 2014	A24E3SXTC62LJI	7508492919	{'Color:': ' Bling'}	Claudia Valdivia	Looks even better in person. Be careful to not...	Can't stop won't stop looking at it	1407110400	NaN	NaN
1	5	True	02 12, 2014	A269FLZCB4GIPV	7508492919	NaN	sarah ponce	When you don't want to spend a whole lot of ca...	1	1392163200	NaN	NaN
2	3	True	02 8, 2014	AB6CHQWHZW4TV	7508492919	NaN	Kai	so the case came on time, i love the design. I...	Its okay	1391817600	NaN	NaN
3	2	True	02 4, 2014	A1M117A53LEI8	7508492919	NaN	Sharon Williams	DON'T CARE FOR IT. GAVE IT AS A GIFT AND THEY...	CASE	1391472000	NaN	NaN
4	4	True	02 3, 2014	A272DUT8M88ZS8	7508492919	NaN	Bella Rodriguez	I liked it because it was cute, but the studs ...	Cute!	1391385600	NaN	NaN

Sample Dataset of Tools and Home Improvement category

	overall	verified	reviewTime	reviewerID	asin	style	reviewerName	reviewText	summary	unixReviewTime	vote	image
0	5	True	01 28, 2018	AL19QO4XLBQPU	0982085028	{'Style:': ' 1) IR30 POU (30A/3.4kW/110v)'}	J. Mollenkamp	returned, decided against this product	Five Stars	1517097600	NaN	NaN
1	5	True	11 30, 2017	A1I7CVB7X3T81E	0982085028	{'Style:': ' 3) IR260 POU (30A/6kW/220v)'}	warfam	Awesome heater for the electrical requirements...	Five Stars	1512000000	NaN	NaN
2	5	True	09 12, 2017	A1AQXO4P5U674E	0982085028	{'Style:': ' Style64'}	gbieber2	Keeps the mist of your wood trim and on you. B...	Five Stars	1505174400	NaN	NaN
3	4	True	07 19, 2017	AIRV678P7C4NK	0982085028	NaN	Justin Banner	So far I hooked it up and tested it , filled a...	it is the perfect temp for a shower	1500422400	NaN	NaN
4	1	True	05 25, 2017	A22I5QDNTNECDW	0982085028	{'Style:': ' 3) IR260 POU (30A/6kW/220v)'}	daveparker	i installed this 10 months ago, instructions w...	worked well...for 10 months.	1495670400	16	NaN

Data Pre-processing

Need for Data Pre-Processing



- Data preprocessing is a fundamental step in any data analysis or machine learning project.
- Its primary purpose is to **prepare the data for machines** to learn from and provide meaningful results.
- Without proper pre-processing, **raw data may contain inconsistencies, missing values, or outliers**, which can negatively impact the performance of machine learning models.
- Data pre-processing **lays the foundation for accurate insights and predictions**, making it a crucial stage in the data science pipeline. It involves transforming raw data into a format that is suitable for training and testing models of high quality.
- The **main objectives include removing noise or irrelevant information, removing outliers, cleaning data, handling missing values**, and preparing data well for further analysis and modeling to get higher accuracy and correct predictions. There is a line that is always associated with the field of data analysis: ***"Garbage-In, Garbage-Out".***
- It clearly means that the **quality of the results is linearly related to the quality of the input of data**.

Data Pre-Processing - Techniques

Techniques Used	Usage/Results
Deduplication	Deduplication is needed on large datasets as it's essential to ensure data quality, consistency, and uniformity. Duplicate values can skew the analysis and introduce unnecessary noise.
Handling Missing Values/NA/NaN's/Null	Removing null/missing values from a dataset is very important because many machine learning algorithms and statistical analyses cannot handle missing data efficiently. These values can introduce bias or inaccuracies in the result and can also cause errors or unexpected behavior in computations.
Standardization of Data Types	In order to treat the missing value, understanding the columns and its relevance is important to come up with the appropriate solution.
Imputation	If the feature is of high importance then we may want to impute the values instead of removal. If the missing values are high in number, it is best to impute values using mean, median or mode.
Feature Addition	Based on the use case, we need to add additional columns based on the derivation from existing columns which enhance the solution understanding at hand
Feature Encoding	Converting categorical variables into a format that can be provided to ML algorithms.

Data Pre-Processing - Original Features

Column	Summary	Original Datatype	Modification
Overall	Rating for the Product	Int64	As is
verified	If the purchase is verified or not. Also, useful for getting helpfulness dependent variable	Bool	Imputed Missing values with False
reviewTime	Time at which the review was posted in “mm dd, yyyy” format	Object	Dtype changed to standardized datetime format, and created additional columns for day, year, month
reviewerID	ID of the reviewer who posted the review	Object	As is
asin	Amazon Standard Identification Number	Object	As is
style	Describes the SKU or variant	Object	Empty values replaced by an empty string
reviewName	Name of the reviewer	Object	Missing values imputed with “unknown”

Data Pre-Processing - Original Features

Column	Summary	Original Datatype	Modification
reviewText	Actual review text	Object	Dropped the rows with missing values as instance were low
summary	Review title	Object	Dropped the missing values
vote	Number of votes for helpfulness	Object	Imputed these values with “0”
image	URL of review image if available	Object	Convert to Boolean values T/F
unixReviewTime	Time of the review (Unix format)	Int64	As is

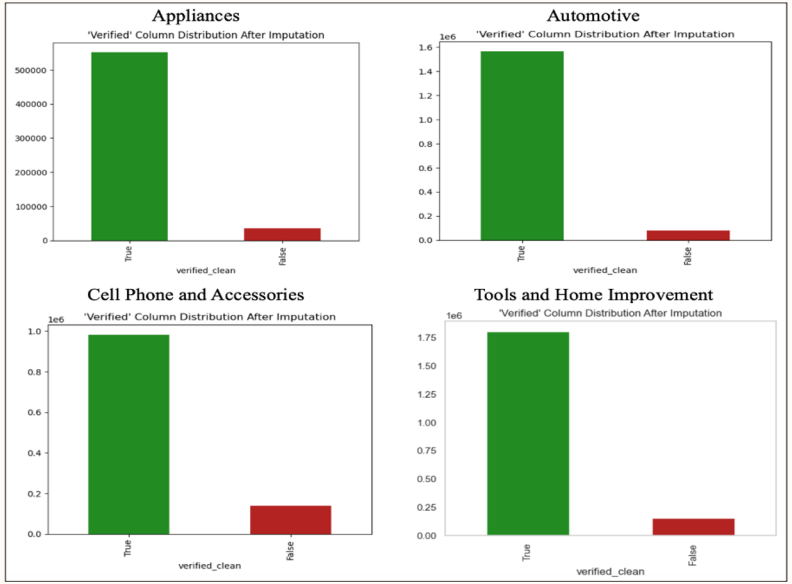
Data Pre-Processing - New Features

Column	Summary	Original Datatype	Usage
reviewTime_cleaned	Cleaned and Formatted review time	DateTime	Used for fetching review age and year
review_year	Review Year extracted from cleaned review time	Object	Used for filtering the data as we are only considering reviews posted after year 2010
review_month	Review Month extracted from cleaned review time	Object	Used for review age determination
review_day	Review Day extracted from cleaned review time	Object	Used for review age determination
verified_clean	Verified Review column after imputation	Bool	Used to generate target feature “helpfulness_binary”
vote_clean	Votes column after imputation	Int64	Used to generate target feature “helpfulness_binary”
image_available	Image available column converted to bool	Bool	Used for stratified Sampling

Data Pre-Processing - Target Variable

Out target feature called "**helpfulness _binary**" was created using the "vote_clean" and "verified_clean" features with the certain conditions. Helpfulness was tagged with a value of 1 only if the number of votes was greater than 0 and if the review was verified, for all other combinations it was tagged with a value of 0. ***Our target variable can be interpreted as "1" being helpful, and "0" being non-helpful. Chart below shows that majority of the reviews are verified.*** However, the table on left shows that percentage of helpful reviews is only 8-10%. This is because the number of votes on the reviews plays a major role in our target feature. Using a combination of the two, we are able to setup our target variable.

Category	Helpfulness	Percentage
Appliances	0: 540331 **1: 47005**	0: 92% **1: 8%**
Automotive	0: 1474453 **1: 166434**	0: 90% **1: 10%**
Cell Phones and Accessories	0: 1050626 **1: 68505**	0: 94% **1: 6%**
Tools and Home Improvement	0: 1726619 **1: 231963**	0: 88% **1: 12%**



Data Pre-Processing - Stratified Sampling

Stratified sampling is a technique employed in data statistics and research that ensures the sampled data is an accurate representation of the original population. Based on certain characteristics we create subgroups called strata and then take random samples from each stratum. We used the features “helpfulness_binary”, “review_year”, “review_month” and “image_available” for defining strata. This ensured that we had accurate representation in the sample. Following is distribution plot for target variable on overall and sample data.

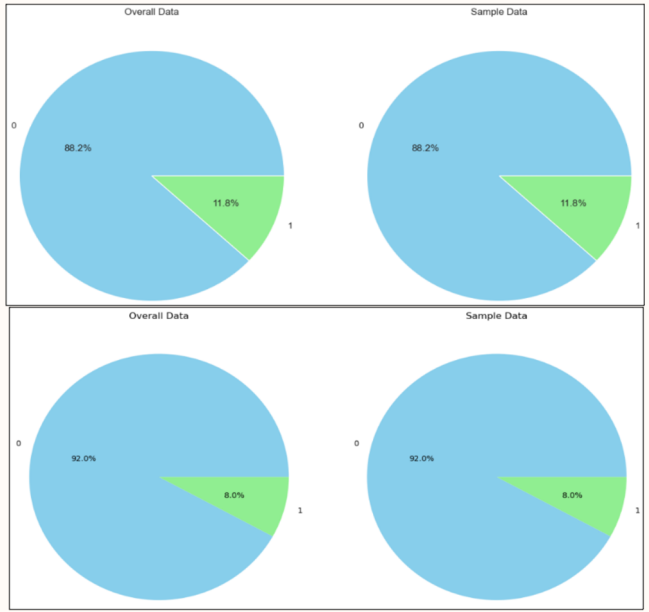


Image Source: https://www.scribbr.com/methodology/stratified-sampling/

Data Pre-Processing - Data Shape

The original size of our datasets was huge with millions of rows. There were also a quite a few duplicate rows along with redundant data. We had to perform data reduction to have a more manageable size and avoid processing constraints. Following is a comparison of the data shape across categories for our dataset. We did stratified sampling on the cleaned data to consider a sample size of 50000 across each category.

Category	Original Size	After Deduplication	After Sampling
Appliances	602777	591371	**49995**
Automotive	1711519	1647280	**50002**
Cell Phones & Accessories	1128437	1124986	**50006**
Tools and Home Improvement	2070831	1982666	**49996**

Data Transformation

Data Transformation: review title

- Data transformation involves **cleaning, integration, organizing, aggregation, and enriching the data for modeling**
- This step is **crucial in the Machine Learning lifecycle**, as building the models will be easy with transformed data
- We extracted **latent features of the text data which we refer to as metadata features** and there are **two primary columns with the text data – review title and review text**
- The **title is** intended to be a **simple summary** while the **review text is a more elaborate description**
- We extracted the **number of words in the title**, the **length of the original title**, the **sentiment associated with the title**, **unique word count**, **emoji and non-ASCII character** count from review text feature
- The **sentiment of the review title is essential** to understand the sentiments and **they can take three values such as positive, negative, and neutral**
- The length of the title, the number of words in the title and unique word count can help in understanding the relationship between long and verbose titles on helpfulness
- Insights from analyzing the metadata features of titles can be crucial in predicting helpfulness and can also be used by the sellers to improve customer experience

Data Transformation: review text

- **The metadata features applied on the title are also applied on review text**
  - Additionally, **stop word count** and **URL presence features** are extracted
- Stop word count may also be an interesting feature since we will be removing stop words from the review text
  - **Stop words are words that are often used for grammatical purposes** like “a”, “the”, “is”, “an”, etc. but do not add value to the underlying meaning of the text and can be excluded from text processing
  - Also, identifying the **count of stop words helps understand the ratio of signal versus noise in a given text**
- **URL presence may be an interesting feature since we can understand if reviews, wherein customers share links to other products or websites, are helpful or not**
- We also calculated the **review age** by considering the time difference in days between a review being posted and the end of the data, that is 31st Dec, 2018
  - **Review age feature brings temporal data directly into the dataset** so that more recent reviews are represented by smaller numbers, which can be more **informative to assessments of recent product purchases**
- We perform **text cleaning on the review text feature to create review_text_clean feature** and this is done in three stages
  - First stage is **converting all the review text to lower case to offset the effect of different textual cases** on the raw data, allows algorithms to identify repetitions of the same words
  - In the second stage, we **expand contractions to make the text clearer for machines** to read by retaining the actual semantics and structure
  - In the third stage, we used some functions to clean up the review text: for instance, **removing non-ASCII characters, and special characters such as emojis from the review text feature**

Data Transformation: TF-IDF

- Finally, we **convert the cleaned review text data into a TF-IDF matrix**
  - **TF-IDF stands for Term Frequency-Inverse Document Frequency** and is a method of measuring the relevance of a word in a document and a collection of documents
  - The TF part stands for Term Frequency which is the count of occurrences of the word
  - The IDF part stands for Inverse Document Frequency and is the log of the ratio of the total number of documents to the number of documents in which a term appears; **the final metric is a product of TF and IDF**
  - This method ensures that words that occur too frequently get a low score and terms that are rare or relevant get a higher score and also ensures that we're able to score terms based on their relative importance and not just the number of times they are used
- As mentioned earlier, we remove special characters, stop words and digits from the review text since these are unimportant
- Later, we expand contractions and colloquial abbreviations, turning shorthand notations into complete descriptions like "tbh" to "to be honest" so that the algorithms understand the contextual usage
- We performed **Lemmatization, which is the process of reducing different inflected forms or variants of the same word in text data to a common root word** - this is a kind of text normalization technique
- We have visualized the review text clean feature by creating n-gram distributions across all the categories as shown in the following slide

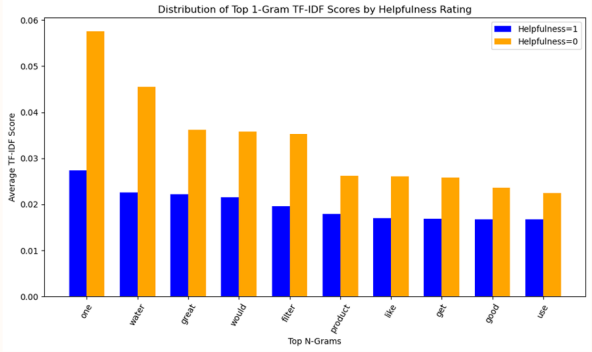
Data Transformation: ngrams visualization

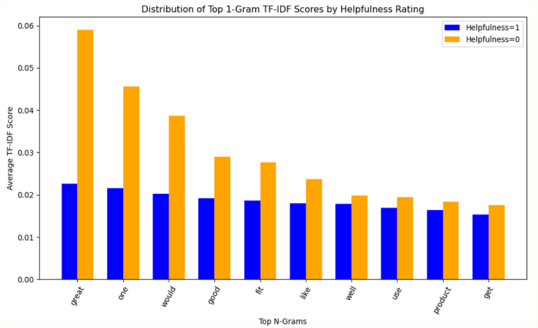
Data snapshot

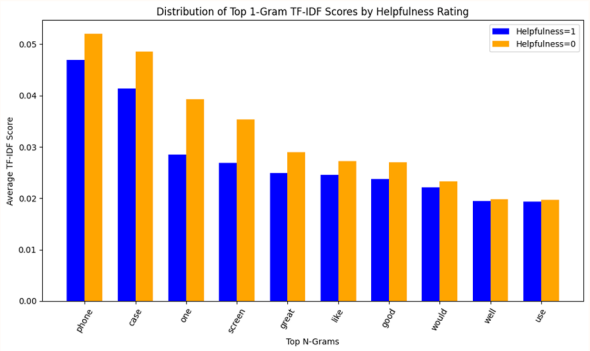
```
Index(['overall', 'vote', 'verified', 'reviewTime', 'reviewerID', 'asin',
       'style', 'reviewerName', 'reviewText', 'summary', 'unixReviewTime',
       'image', 'reviewTime_cleaned', 'review_year', 'review_month',
       'review_day', 'verified_clean', 'vote_clean', 'image_available',
       'helpfulness_binary', 'num_words_review_title', 'title_length',
       'title_sentiment', 'unique_word_count_title',
       'emoji_non_ascii_count_title', 'num_words_review_text', 'review_length',
       'review_sentiment', 'unique_word_count_review', 'url_count_review',
       'stop_word_count_review', 'review_age_days', 'review_text_clean',
       'review_text_clean_lemmatized'],
      dtype='object')
```

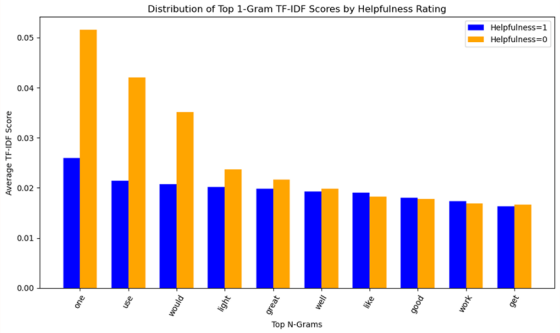
num_words_review_text	review_length	review_sentiment	unique_word_count_review	url_count_review	stop_word_count_review	review_age_days	review_te
172	880	0.186111	95	0	50	3280	cartridges work c work
60	300	-0.011111	48	0	29	3280	repl produc glove how
789	3841	0.121429	339	0	296	3261	heirarchy induc althou
310	1612	-0.057292	190	0	103	3282	bosch sh reputation en
23	95	0.377778	20	0	11	3279	power length qua ea

Unigrams for all four categories









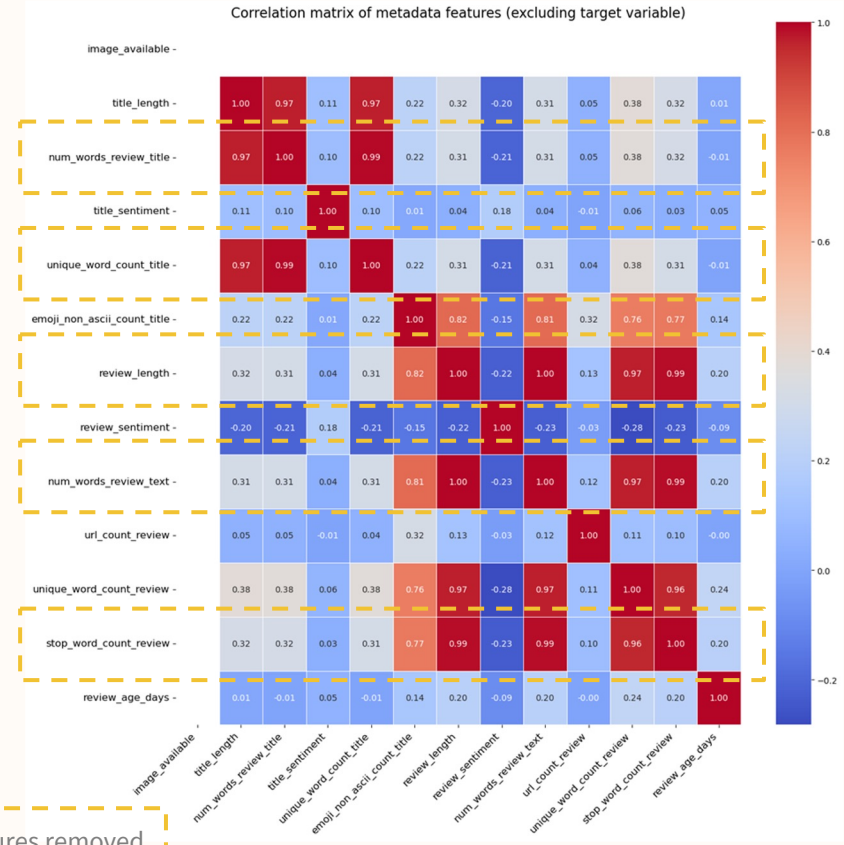
Data Preparation

Data Preparation: metadata feature correlation

- TF IDF features are too large - we cannot manually look for and address correlation between them and hence, we look at the correlation between the metadata features only
- After analyzing the correlation matrix, there are a lot of correlated metadata features; **we decided the absolute correlation cut-off as 0.85**
- This means that **any feature that is more than +0.85 is highly positively correlated, any feature that is less than -0.85 is highly negatively correlated**
- Finally, we removed the correlated features based on relative importance
- After removing highly correlated features, the final correlation matrix shows the metadata variables to be considered for modeling

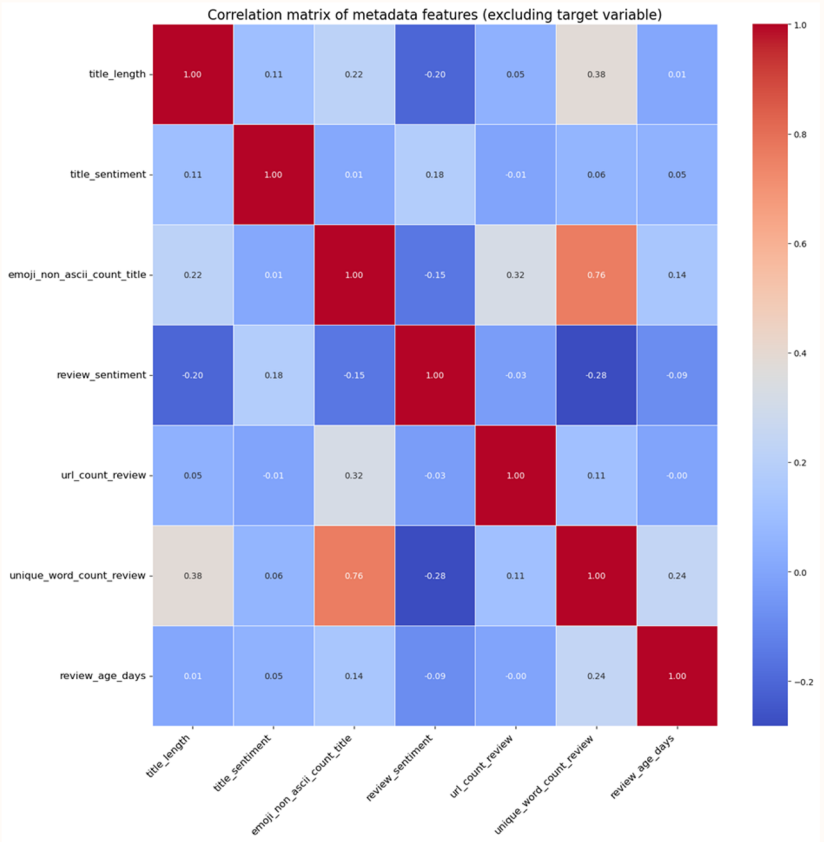
Data Preparation: correlation heatmap

Correlation HeatMap before



Features removed

Correlation HeatMap after



Data Preparation: modeling dataset creation

- After the data is pre-processed and transformed, it is ready for modelling
- We have a huge volume of data, and have reduced the data to get a stratified sample of 50K reviews
- Sampled data has same distribution as that of the overall data to avoid bias in modeling
- We are considering **modeling data as a combination of metadata features and top 1000 words from the TF IDF matrix**
- We have split the transformed data into 70:15:15 split i.e, we have considered **70% as training set, 15% as testing test and 15% as validation set**
- We will build the baseline model on the training dataset, tune model on the validation dataset and test these models on the testing dataset
- We will calculate the metrics like accuracy, precision, recall, ROC, AUC and F1-score on the baseline and tuned models and compare them to pick the best performing model

Rows and columns of the data

```
X_train shape: (39996, 1008)
y_train shape: (39996,)
X_test shape: (7499, 1008)
y_test shape: (7499,)
X_validation shape: (7500, 1008)
y_validation shape: (7500,)
```

`X_train.head(2)`

	able	absolutely	access	accurate	across	actual	actually	adapter	add	added	...	yet	youtube	title_length	title_sentiment	emoji_non_ascii_co
31984	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	10	1.0	
47286	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	83	0.7	

2 rows × 1008 columns

`X_test.head(2)`

	able	absolutely	access	accurate	across	actual	actually	adapter	add	added	...	yet	youtube	title_length	title_sentiment	emoji_non_
1659	0.116317	0.0	0.0	0.0	0.163125	0.0	0.0	0.0	0.0	0.162023	...	0.0	0.0	26	0.166667	
23287	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	56	0.000000	

2 rows × 1008 columns

`X_validation.head(2)`

	able	absolutely	access	accurate	across	actual	actually	adapter	add	added	...	yet	youtube	title_length	title_sentiment	emoji_non_ascii_co
37352	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	4	0.7	
9052	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	10	0.0	

2 rows × 1008 columns

Data Statistics

Data Statistics

- The raw data set that was gathered included 13,677 records from the Appliances category, 1,711,519 records from the Automotive category, 1,128,437 records from the Cell Phones and Accessories category, and 2,070,831 records from the Tools and Home Improvement category.
- First, unprocessed data is gathered, where the number of rows and columns indicates the amount of the dataset for each category. Data preparation modifies the dataset dimensions by removing duplicates, lemmatizing terms to their basic forms, etc.
- After that, TF-IDF feature extraction is used to convert the data, creating a homogeneous feature space with 1000 columns for every category. Ultimately, the data is separated into training with 70%, testing and validation each with 15% sets, each of which has a distinct function in the creation and evaluation of machine learning models.
- Notably, the Tools and Home Improvement category has a testing set that is significantly larger than the others. This methodology guarantees a thorough examination of the evaluations with the goal of using machine learning techniques to extract significant insights.
- Table includes summary of data sizes after each step of data processing shown in the next slide.

Summary of datasets after data processing

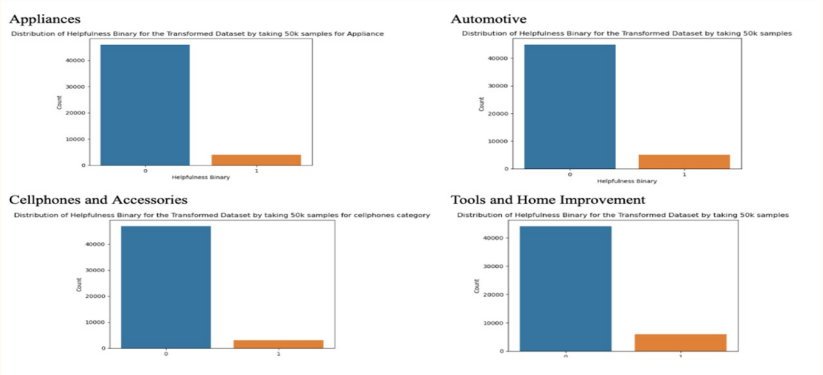
Stage	Process	Categories	Rows x Columns
Raw data collection	Appliances dataset		602777 x 12
	Automotive dataset		1711519 x 12
	Cell Phones and Accessories dataset		1128437 x 12
	Tools and Home Improvement dataset		2070831 x 12
Data Preprocess	After Dedupliaction	Appliances	591371 x 12
		Automotive	1647280 x 12
		Cell Phones and Accessories	1124986 x 12
		Tools and Home Improvement	1982666 x 12
	Lemmatization	Appliances	49995 x 34
		Automotive	50002 x 34
		Cell Phones and Accessories	50006 x 34
		Tools and Home Improvement	49996 x 34

Summary (Contd.)

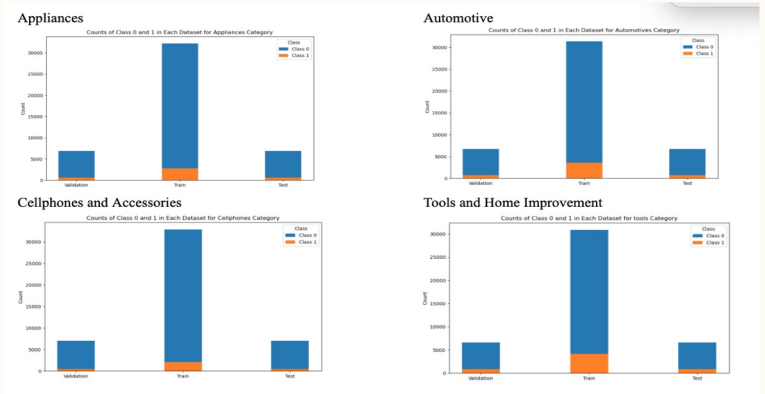
Stage	Process	Categories	Rows x Columns
Data Transformation	Feature extraction using TF-IDF	Appliances	49995 x 1000
		Automotive	49872 x 1000
		Cell Phones and Accessories	50006 x 1000
		Tools and Home Improvement	49996 x 1000
Data Preparation	Training set	Appliances	39996 x 1008
		Automotive	34910 x 1008
		Cell Phones and Accessories	35004 x 1008
		Tools and Home Improvement	39996 x 1008
	Validation set	Appliances	7499 x 1008
		Automotive	7481 x 1008
		Cell Phones and Accessories	7501 x 1008
		Tools and Home Improvement	7499 x 1008
	Testing set	Appliances	7500 x 1008
		Automotive	7481 x 1008
		Cell Phones and Accessories	7501 x 1008
		Tools and Home Improvement	7500 x 1008

Data Statistics (Contd.)

Bar Charts showing the distribution of helpfulness binary for the transformed dataset by taking 50k samples for various categories



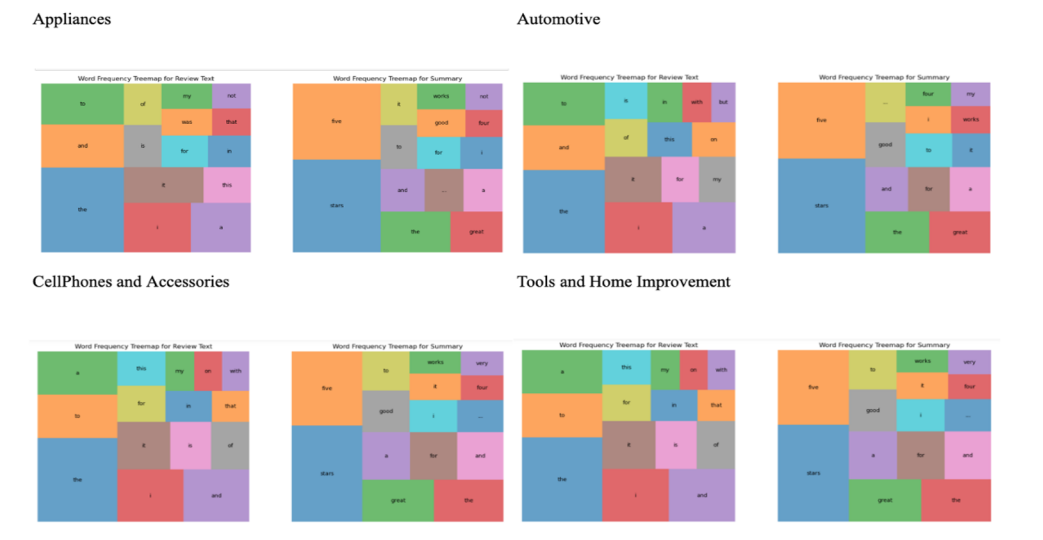
Bar charts showing count of class 0 and 1 in train, valid and test dataset for various categories



Data Statistics (Contd.)

Word Frequency Treemap for review text and summary

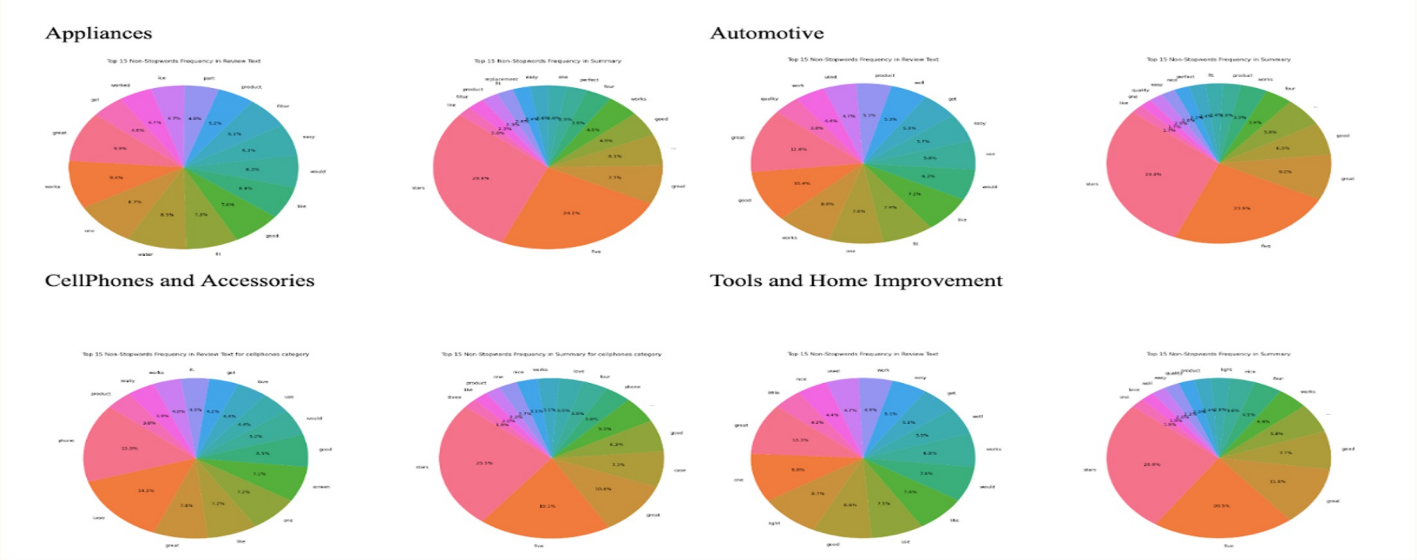
The treemaps show that stop words that are commonly used in English, such as ‘and’, ‘the’, ‘for’ dominate in these charts. The rectangles in each treemap show the frequency of each word, with bigger sizes indicating higher frequency.



Data Statistics (Contd.)

Pie charts showing top 15 non-stopwords frequency in and review text and summary

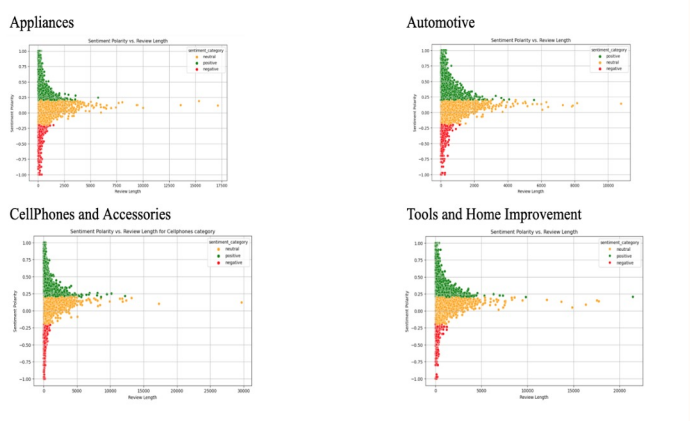
We can see these charts provide insight into the most common terms used by customers in their reviews. For instance, words like ‘great’, ‘easy’, ‘perfect’ are frequently mentioned suggesting the areas of customer focus or satisfaction. These terms can highlight what customers value most in their products and can be used to understand areas for product improvement.



Data Statistics (Contd.)

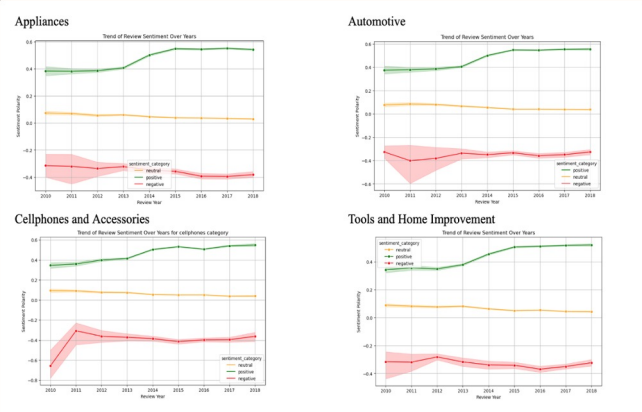
Scatterplot showing sentiment polarity vs. review length for various categories

Positive sentiments are more frequent in shorter reviews, while negative sentiments appear across all lengths, and neutral sentiments are less common overall. The trend suggests that people often write positively in brief reviews, while longer reviews may contain more detail, often associated with negative feedback.



Line graphs showing the trend of review sentiment over years

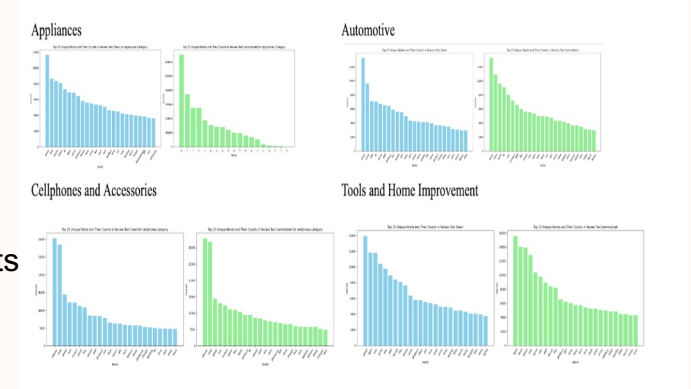
The high percentage of positive ratings indicates that customers are generally satisfied. While the categories for Automotive, Tools and Home Improvement indicate progress, there is a small increase in negative sentiment in the Appliances, Cellphones and Accessories categories, suggesting dissatisfaction in recent years. Neutral sentiment remains relatively unchanged.



Data Statistics (Contd.)

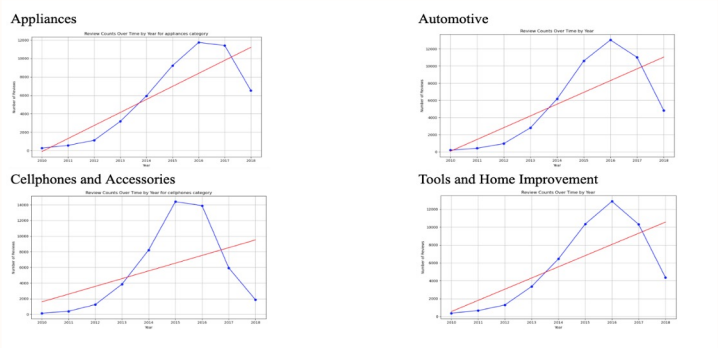
Bar charts showing the top 25 unique words and their counts in review text clean and review text Lemmatized

The charts indicate that post-lemmatization, the word distribution does not significantly differ from the cleaned text data. This suggests that lemmatization has not changed the frequency of the top words.



Line graphs showing review counts over time by year for various categories

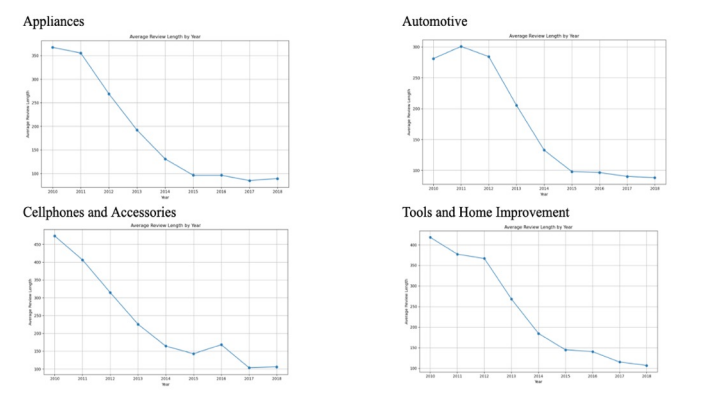
Every graph shows the count of reviews between 2010 and 2018, which reflects patterns in consumer buying and reviewing habits. There are notable peaks in a few categories, signifying years with a large spike or drop in the number of reviews.



Data Statistics (Contd.)

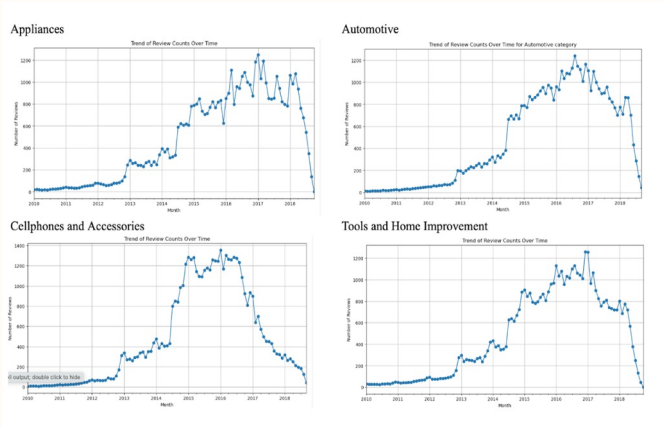
Line graphs showing average review length by year

Every category shows a distinct downward trend, suggesting that customer ratings have been shorter over time.



Line graphs showing trend of review counts over time

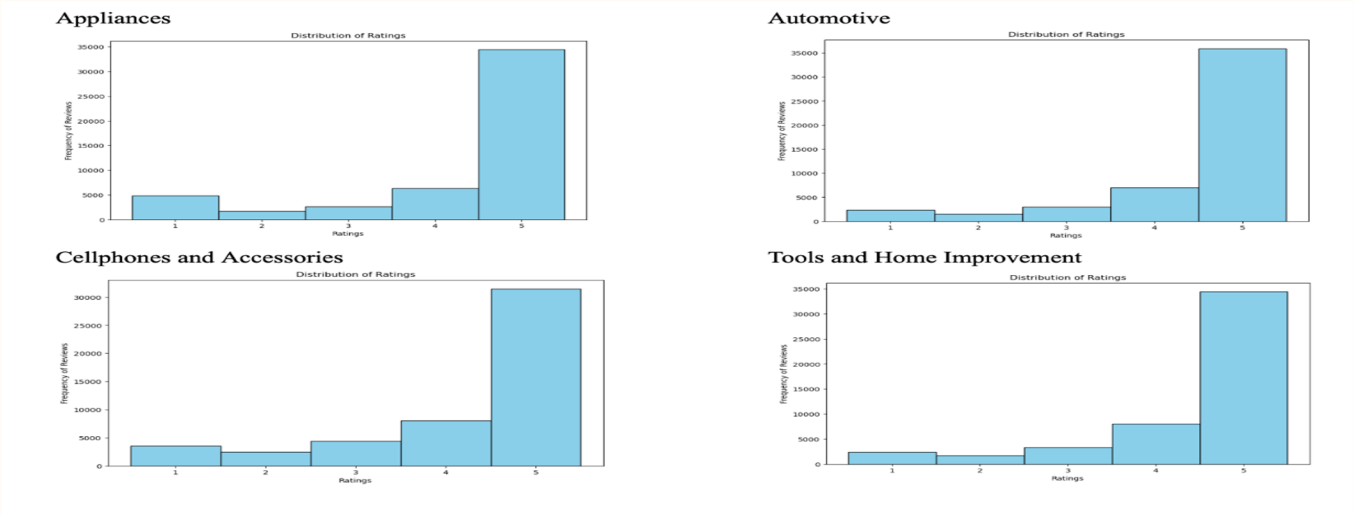
In the middle of the timeframe, there is a sharp peak for cellphones and accessories, which is followed by a decrease. There is a general growth in Tools and Home Improvement, showing an overall increase with some fluctuations and a peak before a decline.



Data Statistics (Contd.)

Histogram showing distribution of ratings for various categories

Every histogram plots a frequency count against the associated rating number to display the frequency of ratings on a range of 1 to 5. The distribution is heavily skewed in towards the higher rating of 5, which is the highest possible rating provided by the majority of consumers in all categories. Across all categories, ratings of 1 are the least common, and ratings of 5 are the most common.



Thank You