Report

About Dataset

The dataset used for this project is the Amazon Review/Product Dataset, provided by Julian McAuley and Jianmo Ni, University of California, San Diego (UCSD). It includes data reviews for the range May 1996 - October 2018. It has a total number of 233.1 million reviews of several different categories. In this project, we have worked on two such categories, namely "Cell Phones and Accessories" and "Video Games". Both of these had two datasets under them, the "review data" and the "metadata". The former comprised of information such as the identity of the review and the product, text summary and helpful votes of the review, etc. The metadata included descriptions, price, sales-rank, brand info, and co-purchasing links. The Cell Phones and Accessories dataset had 1,128,437 rows and 12 columns, and its metadata had 590,071 rows and 19 columns. The Video Games dataset had 231,780 rows and 9 columns, and its metadata had 84819 rows and 19 columns. Note that the above figures represent the original raw data, without any data cleaning.

Data Pre-processing and Cleaning

Both the rating data and the metadata were initially JSON files, and hence were present in python dictionary form. To do data manipulation and further analysis, these were converted to python dataframes. Our data had an abundant amount of duplicate values, they are redundant and may contaminate the training data with the test data or vice versa. Hence, these were dropped.

Now, to integrate the product details along with the reviews, the review data and metadata, for each of the two categories, were merged based on the same Amazon Standard Identification Number (ASIN) number of the products.

There were a few columns in this combined dataframe which were not of much importance and did not contribute significantly as a factor while deciding the value of the target variable ("votes" or "helpfulness count"), hence these were removed as well. Several rows in the column which contained the detailed text of the reviews, i.e. the "reviewText" column, had NaN (Not a Number) values in them. Since no analysis can be done without the reviews of the products, such rows were removed immediately. Some of them also had numbers as the input, which does not make any sense and no evaluation can possibly be done for such rows, hence these were removed too. Various rows in the "votes" column were also filled with NaN values. Since this is the target variable, it cannot be dropped, however it can be easily inferred that a row having NaN value as its vote could simply mean that the review got no votes at all, and hence could be replaced by zero.

After the above data cleaning, the merged Cell Phones and Accessories dataset had 1,041,169 rows and 18 columns, while the merged Video Games dataset had 28133 rows and 28 columns.

Exploratory Data Analysis (EDA)

With the clean data in hand, some processing can be done on the “reviewText” column to analyze the reviews and extract features from them, which can further be analysed to predict the corresponding number of votes.

One of the trivial features that was extracted was the number of sentences of all the reviews in the dataset. This was followed by the removal of “stopwords” (a set of commonly used words in any language). This is a very critical step as it will eliminate unimportant words, and will allow the model to focus on the important words instead. This also significantly reduces the size of the overall corpus without sacrificing its quality and semantics. In our case, the presence of numbers would not hold any vital information in the reviews, and hence they were removed as well. Punctuation marks were also removed and all the text was converted to lowercase so that the texts/words get treated equally by the model. We also applied Lemmatization to our review texts, as it maps multiple words to a common root word. That way, these words are treated similarly and the model learns that they are being used in similar contexts. After this, two others features that were calculated were the number of words and the number of unique words in a review.

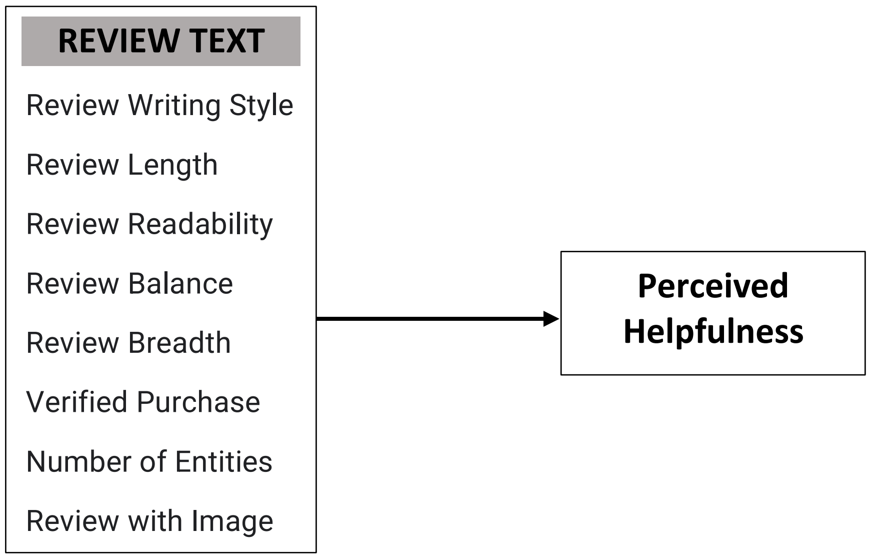
The “rank” column in the dataset gave the rank of any particular product, along with the the category and sub-categories name. Some rows were not in the proper format, so after some more text mining, this rank and and the names of main category and the three sub-categories under it were also extracted.

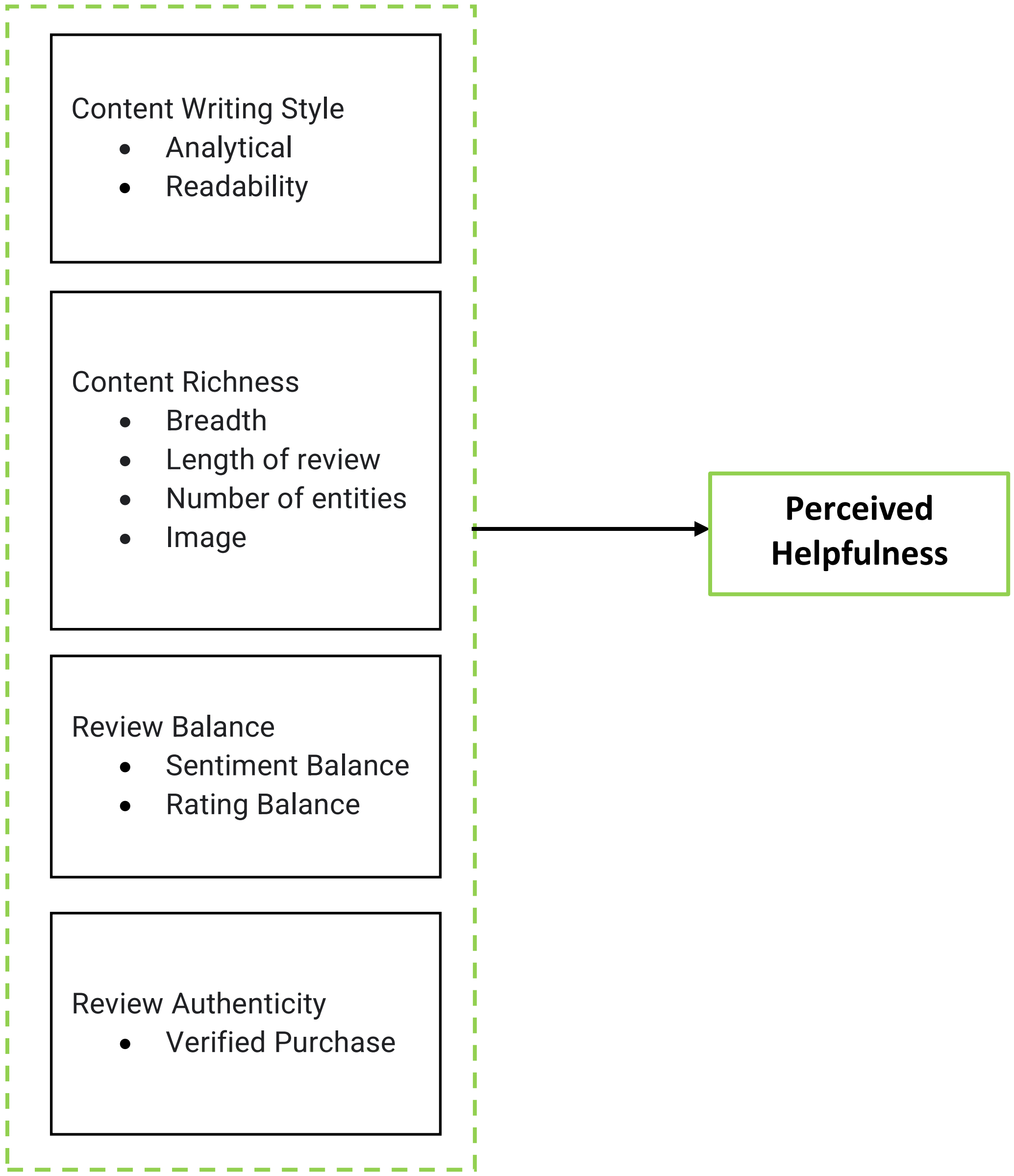
Linguistic Inquiry and Word Count (LIWC)

Linguistic Inquiry and Word Count (LIWC) is a text analysis software that gives more profound insights into the words in the corpus. When run through the dataset, it gives certain features for any review, such as Authentic, Emotional Tone, Past Focus, etc. In order to calculate these statistics, each dictionary word is measured as a percentage of total words per text (Cronbach’s α) or, alternatively, in a binary “present versus absent” manner (Kuder–Richardson Formula 20; Kuder & Richardson, 1937). These LIWC features, along with the attributes discussed above, could be used to run through a Linear Regression or a Neural Network model to potentially find a pattern/correlation between these review features and the associated votes.

Model

The various features of a review that affect the perceived helpfulness are as follows:


REVIEW TEXT
Review Writing Style
Review Length
Review Readability
Review Balance
Review Breadth
Verified Purchase
Number of Entities
Review with Image
Perceived Helpfulness


Content Writing Style
Analytical
Readability
Content Richness
Breadth
Length of review
Number of entities
Image
Review Balance
Sentiment Balance
Rating Balance
Review Authenticity
Verified Purchase
Perceived Helpfulness

1. Review Writing Style (Analytic)

Customers can form impressions based on the structure and style of online content. They avoid noisy and chaotic information and seek clarity and certainty in a review. Hence, a formal, logically ordered, specific, and consistent writing gives a them convincing impression. This measure in a review was assessed by the LIWC-22 summary variable “Analytical thinking” (abbrev. Analytic), which defines it as a “Metric of logical, formal thinking”.

The LIWC Analytic score ranges from 0 to 100. A higher Analytic score refers to a higher degree of formal, logical, and analytical thinking in the text, and is correlated with things like grades and reasoning skills, whereas a lower score means a more narrative, intuitive, and personal writing style.

Conclusion – Analytical writing showed a **positive correlation** with helpful votes count in Amazon reviews, indicating that structured, logical, and formal writing enhances perceived helpfulness. This suggests that readers value clarity and reasoning in reviews, as it reduces cognitive effort and increases trust in the information presented.

2. Review Length (WPS)

The review length was calculated by dividing the total number of words by the total number of sentences in a review. The two mentioned features were calculated by LIWC.

Conclusion – Review length (WPS) exhibited a **curvilinear relationship** with helpful votes count, meaning extremely short or excessively long reviews were perceived as less helpful, while moderate-length reviews received the most helpful votes. This suggests that readers prefer reviews that are detailed enough to provide useful insights but not overwhelmingly lengthy, balancing informativeness and readability.

3. Review Readability

Readability refers to the ease with which a reader can understand a written text. Readability of the reviews was assessed by using the Flesch Reading Ease Index (FRE). In this test, a higher score indicates ease of readability (i.e., the higher the score, the easier it is to read and comprehend the text).

The formula for the Flesch Reading Ease index is:

Flesch Reading Ease = 206.835 - (1.015 * Average words per sentence) - (84.6 * Average syllables per word)

Conclusion – Readability, measured by the Flesch Reading Ease Index (FRE), exhibited a **curvilinear relationship** with helpful votes count, where reviews with moderate readability were rated as most helpful. Extremely easy-to-read reviews may lack depth and detail, while overly complex reviews may be difficult to comprehend, suggesting that readers prefer a balance between simplicity and informativeness.

4. Review Balance

Review balance refers to the degree to which the review's tone was positive or negative. A positive sentimental tone might convey pleasant information to the consumer, whereas a negative sentimental tone sends a disappointing or unpleasant message to the consumer. A positive review can provide a consumer with the reassurance and confidence they need to make a purchase, while a negative review can serve as a warning to steer clear of the product and consider alternative options, ultimately aiding in their decision-making process. This was calculated by taking the ratio of the two LIWC variables, positive tone (tone_pos) and negative tone (tone_neg) of a review.

Conclusion – Review balance has a **curvilinear relationship** with helpful votes count. Extremely positive or extremely negative reviews may be perceived as biased or lacking critical insight, while more balanced reviews—those that acknowledge both pros and cons—are often seen as more credible and useful. Readers may trust reviews that present a nuanced perspective, helping them make informed decisions.

5. Review Breadth

Review breadth refers to the various number of topics a review discusses. This was achieved by implementing Topic Modeling using Latent Dirichlet Allocation (LDA), discussed below.

Topic Modeling using LDA

Introduction

Topic Modeling is an unsupervised approach of recognizing or extracting different topics in various documents by extracting the patterns of word clusters and frequencies of words in the document. As this doesn't have any outputs through which it can do this task hence it is an unsupervised learning method. This type of modeling is very much useful when there are many documents present and when we want to get to know what type of information is present in it. This takes a lot of time when done manually and this can be done easily in very little time using Topic Modeling.

There are various techniques for implementing it; a few popular ones include Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (pLSA), and Latent Dirichlet Allocation (LDA). Due to its ability to build valid dictionaries and use previous learnings to predict topics in new sets of documents, LDA is the recommended model for advanced topic modeling, and hence will be used for the all the datasets in this paper.

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a Bayesian version of pLSA. The core concept is replaced by Dirichlet allocations where the distribution is sampled over a probability simplex. A probability simplex represents a set of numbers that add up to 1. When the set comprises three numbers, it is called a three-dimensional Dirichlet distribution.

The total desired number of topics is set as ‘k’ in the dimensional Dirichlet distribution. The LDA model reads every document, assigns each word to one of the ‘k’ topics, and provides a representation of the words and documents for a given topic. As the assignment of topics is random, the representation is not optimal. Through different equations, the Latent Dirichlet Allocation model can provide the following results:

- Percentage of words within a given document that are assigned to a particular topic.
- The number of times a particular word in all the documents has been assigned to a particular topic.
- Movement of a word from topic A to topic B, i.e. (topic A | d) * P (w | topic A) < P (topic B | d) * P (w | topic B), where ‘w’ means word and ‘d’ means document.

These results denote the optimal number of topics and the assignment of words to each topic. The model can learn from a given set of documents and its Dirichlet distribution and, later, predict topics for a new set of documents.

Choosing the Optimal Number of Topics

Although there is no “hard science” or a single best way or any standard practice to select the optimal number of topics, a reliable way is to compute the topic coherence for different number of topics and choose the model that gives the highest topic coherence. However, this dataset is giving the same coherence value for all the topics from ranging from 6 to 20. So, another method, as used here too, is manually trying out different values of k and select the one that has the largest likelihood. This can be done by using pyLDAvis library in Python to visualize our LDA model and hence the different topics.

A good topic model visualization has the following features:

- The larger the bubble, the higher percentage of the number of words in the corpus is about that topic.
- The further the bubbles are away from each other, the more different they are.

Hence, we would want a model which gives big and non-overlapping bubbles scattered throughout the chart.

Trying all the values from 4 to 10 with our model, k=7 as the number of topics gives the best results considering the above points.

Calculating Dominant Topic and Total Important Topics for a document

After running our LDA model with k=7 as the number of topics, we get the final result as the probability (hence, values will be between 0 and 1) of occurrence of each of the 7 topics, for each individual document. From here, we can calculate 2 more features, namely Dominant Topic and Total Important Topics, to further help us to give more inferences about the dataset.

- **Dominant Topic -** This can simply be defined as the topic having the highest probability out of all the 7 topics. The dataset can further be grouped by this feature and average of the target variable, i.e. votes, in this case, can be calculated, along with the count of people giving the reviews for that dominant topic, to recognize and get the final conclusions on the most important topic in the corpus.

- **Total Important Topics -** For each document, we will be having the probability of occurrence of each of the 7 topics. Now, we can decide a threshold value, say 0.05, and topics having probability more than this threshold value will only be considered for the total count of important topics. This feature tells us about the various different topics a particular document might be talking about, along with its count, thereby letting us get the inference if writing about more number of topics in the same review is efficient or not. The groupby function on this feature, along with the count of people, as previously discussed, can be applied again to get conclusions on the optimal number of documents to be discussed.

Conclusion – Review breadth has a **curvilinear relationship** with helpful votes count. Reviews that cover too few topics may be seen as lacking depth, while those that discuss too many topics might become unfocused and overwhelming. A moderate number of topics likely strikes the right balance, providing enough detail without losing clarity, making the review more useful to readers.

6. Verified Purchase

An 'Amazon Verified Purchase' review means that Amazon has verified that the person writing the review purchased the product from Amazon, and didn't receive the product at a discount.

We measured this as a dummy variable, with a value of one if the verified purchase was ‘TRUE”, and zero if “FALSE.

Conclusion – Verified Purchase status has a **positive linear relationship** with helpful votes count. Reviews from verified purchasers tend to be perceived as more trustworthy and credible since they confirm that the reviewer actually bought the product. As a result, these reviews are more likely to be marked as helpful compared to unverified ones.

7. Number of Entities

Named Entity Recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc. The spaCy library in Python was used to calculate the various entities in a review and the total count of these entities was returned.

Conclusion – The relationship between the total count of named entities and helpful votes count is **postive**. Reviews that mention more named entities—such as product features, brand names, specific dates, or quantities—tend to be more detailed and informative, which makes them more valuable to readers. As a result, these reviews may be perceived as more helpful and receive higher helpful votes.

8. Review with Image

Incorporating media elements, like images, videos, and audio messages, in addition to the written description of a product can improve the customer's user experience by adding more sensory detail and help them make informed purchasing decisions. Our dataset, too, had reviews with image URLs attached with them. We measured the presence of user-provided images of a product in the review as a dummy variable, with a value of one if images were attached alongside a review and zero without them.

Conclusion – The presence of images in reviews has a **high positive relationship** with helpful votes count. Reviews that include images tend to be perceived as more authentic and useful, as they provide visual proof of the product's quality or features. Images help to enhance the overall user experience by adding sensory detail and clarity, which can make the review more convincing and informative. Therefore, reviews with images are more likely to receive higher helpful votes. However, if images are irrelevant or low-quality, the relationship could be weakened or neutral.

Notes –

Content Richness –

1. Content richness refers to information on social media being adequate, clear, and analytical for people to understand and process (Xu & Zhang, 2018). It can be assessed by message cues comprising the amount of information, presence of media elements and writing styles.

2. Richness is an umbrella term to include a set of message cues reflecting whether information is adequate, specific and analytical. Includes some peripheral cues –

i. amount of information – below review with image thing

ii. wrirting style – analytic 2nd point, starting could be “audiences can form impressions based on the structure and style of online content”

Analytic –

1. Third, recent research has suggested that writing style on social media (analytical vs. narrative) is another possible cue to reflect content richness (Pennebaker et al., 2015). Different

writing styles may evoke different impressions, which in turn induce varying engagement behavior (Choi & Stvilia, 2015). In a risky situation, people tend to seek for information to reduce uncertainty and anxiety (Zheng et al., 2021). They prefer information written in an analytical style which is logical and consistent, avoiding chaotic and noisy information. F. Liu et al. (2014) studied rumor retransmission in disasters and showed that ambiguous information is less shared by online users.

2. Formal, logically ordered, specific, and consistent writing gives a convincing impression. the public is compelled to seek clarity and certainty. In doing so, they are wired to avoid noisy and chaotic information (Allport & Postman, 1947), and become more receptive to information that is presented in a formal, consistent, and specific fashion.

3. Richness was measured firstly by the LIWC category Analytic. The category reveals the degree of analytical, logical and consistent thinking, as opposed to more intuitive, narrative writing. This category is derived from prior studies linking the use of articles, prepositions and conjunctions to logical and analytical thinking.

LIWC Analytic score, ranging from 0 to 100, was computed. A higher Analytic score refers to a higher degree of formal, logical, and analytical thinking in the text, whereas a lower score means a more narrative, intuitive writing style.

Review Balance (Sentiment) –

A pos review will give the consumer the affirmation and the confidence to buy the product. At the same time, a neg review will let the consumer know to not buy the product and look for better options instead, thereby helping only with his/her final decisions. THEREFORE, Message cues indicative of emotionality (both negative and positive) predict perceived helpfulnes or votes.

Review with image –

Social media messages are typically short, but additional information can be packed into multiple media elements such as video clips, images, or URLs. These cues enhance the “telepresence”, “media richness” and “vividness” of a message in that they create more direct sensory experience (Liu, Ji, North, & Yang, 2017). Prior studies show that content with more multimedia cues predicts a higher chance of retweeting.

Verified Purchase (Authority) – not sure of the whole thing tho

The construct of Authority first refers to the existing influence of a message source. Audiences analyze source characteristics to infer whether a message is trustworthy. Online opinion leadership is indicated by having a large social following andcertain status symbols. A large follower count indicates source influence after social vetting and can be used to influence an audience’s judgment of source credibility.

Bibliography

1. https://www.analyticsvidhya.com/blog/2022/01/text-cleaning-methods-in-nlp/#:~:text=work%20for%20me',Removing%20Numbers,them%20than%20to%20keep%20them
2. https://www.analyticsvidhya.com/blog/2021/05/topic-modelling-in-natural-language-processing/
3. https://dzone.com/articles/topic-modelling-techniques-and-ai-models
4. https://stackoverflow.com/questions/17421887/how-to-determine-the-number-of-topics-for-lda
5. https://neptune.ai/blog/pyldavis-topic-modelling-exploration-tool-that-every-nlp-data-scientist-should-know

Citation

Please cite the following paper if you use the data in any way:

Ni, J., Li, J., & McAuley, J. (2019, November). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 188-197).