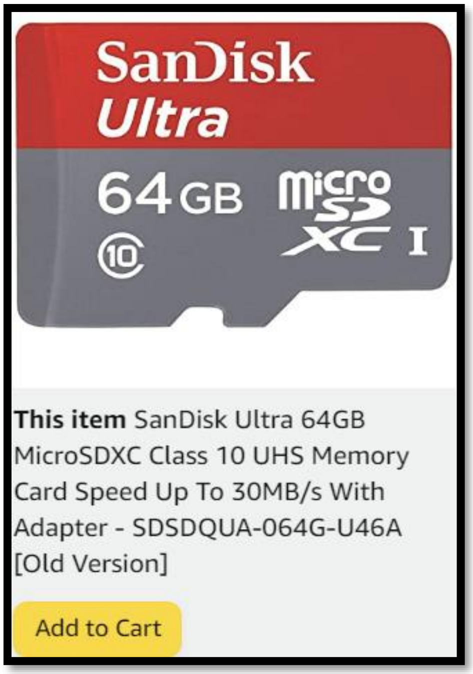


This item SanDisk Ultra 64GB MicroSDXC Class 10 UHS Memory Card Speed Up To 30MB/s With Adapter - SDSDQUA-064G-U46A [Old Version]

Sentiment Analysis of Sandisk Ultra's Amazon Reviews
(Natural Language Processing)

1- Introduction

Sentiment analysis is a powerful technique in natural language processing (NLP) that allows us to extract subjective information from text data. The goal of sentiment analysis is to determine the polarity of a piece of text, i.e., whether it expresses positive or negative sentiment. In recent years, sentiment analysis has become increasingly popular due to its wide range of applications, including customer reviews, social media analysis, and market research.

In this research, sentiment analysis is conducted on Amazon review data about Sandisk Ultra memory card. SanDisk is a reputable digital storage technology company offering a wide range of products such as secure digital cards and memory cards for different electronic equipment, including smartphones, cameras, and computers. They are a leading brand of memory cards, well-regarded by both customers and professionals who are interested in photography and video recording.

Amazon is one of the most popular e-commerce platforms where customers can purchase Sandisk memory cards. Amazon also provides a review system where customers can leave their feedback about the products they have purchased. This review data is publicly available and can be used for sentiment analysis. Analyzing these sentiments not only helps sellers serve customers better, but it can also reveal a lot of customer traits present/hidden in the reviews. In recent years, checking reviews from fellow shoppers helps customers learn more about the product and decide if it is the right product for them or not.

This research presents a sentiment analysis of Amazon review data on Sandisk Ultra memory card 64GB MicroSDXC Class 10 UHS, Speed Up To 30MB/s With Adapter. Specifically, we aim to achieve the following research objectives:

1- To collect Amazon review data about SanDisk memory card from Kaggle.com.
2- To preprocess the collected data by removing irrelevant information, such as product name and reviewer names. Also, to perform text cleaning such as removing stop words, stemming, and lemmatizing the text to prepare the data for analysis.
3- To conduct sentiment analysis on the collected data using a supervised machine learning approach. Train a classification model using a labeled dataset and then use the model to predict the sentiment of the review data.
4- To evaluate the performance of different machine learning algorithms for sentiment analysis on Amazon review data.
5- To provide insights and recommendations based on the sentiment analysis results. Determine the overall sentiment of the data, identify if the customers are satisfied or dissatisfied with the product.

In the following, this report will provide a detailed sentiment analysis that can provide valuable insights into customer satisfaction and product quality of SanDisk memory cards. The results of this research can have practical applications in market research, product development, and customer service.

2- Dataset

The dataset is retrieved from Kaggle.com (https://www.kaggle.com/code/halimedogan). Data contains reviews about only one product: SanDisk memory card 64GB MicroSDXC Class 10 UHS, Speed Up To 30MB/s With Adapter. There are 4915 unique customers and their reviews about the memory card between 01/09/2012 and 12/07/2014. In addition to ‘review text’, the dataset contains the following columns:

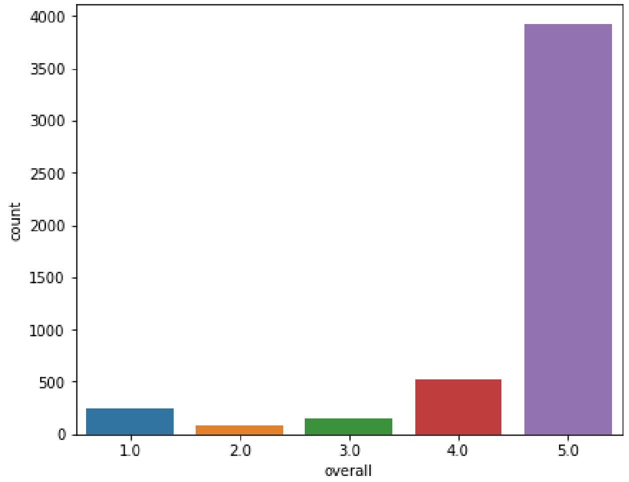
1. Reviewer ID
2. Asin (product ID)
3. Reviewer name
4. Helpful (whether the review was found to be helpful by other customers or not)
5. Review Text
6. Overall (rating)
7. Summary
8. Unix Review Time (the sign-up day of customers)
9. Review time
10. Day difference (between Unix review and review time)
11. Total vote

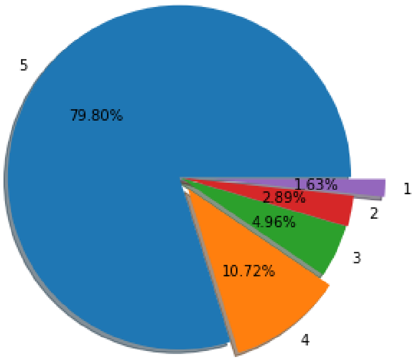
3- Methodology

Sentiment Analysis (emotion AI or opinion mining) is a branch of NLP that aims to recognize and extract emotions, opinions, and attitudes from unstructured text like social media posts, news, and reviews. It uses open-source tools and NLP techniques to turn unstructured text into structured data. Social media platforms such as Facebook, Twitter are excellent sources of sentiment data as users share their views and reactions on a wide range of subjects.

4- Data Wrangling

The analysis begins with loading and exploring the dataset, which includes customer reviews on SanDisk memory card between 01/09/2012 and 12/07/2014. The raw dataset consisted of 4915 rows and 11 columns. The size of the dataset was normal, and there was no need to reduce it. In the data preparation stage, the review time and unix review columns are converted to a datetime object. Overall ratings are visualized using value counts function. Results revealed that around 80% of the ratings are highly positive (5/5) and we will potentially have an imbalance issue even after cleaning the data.





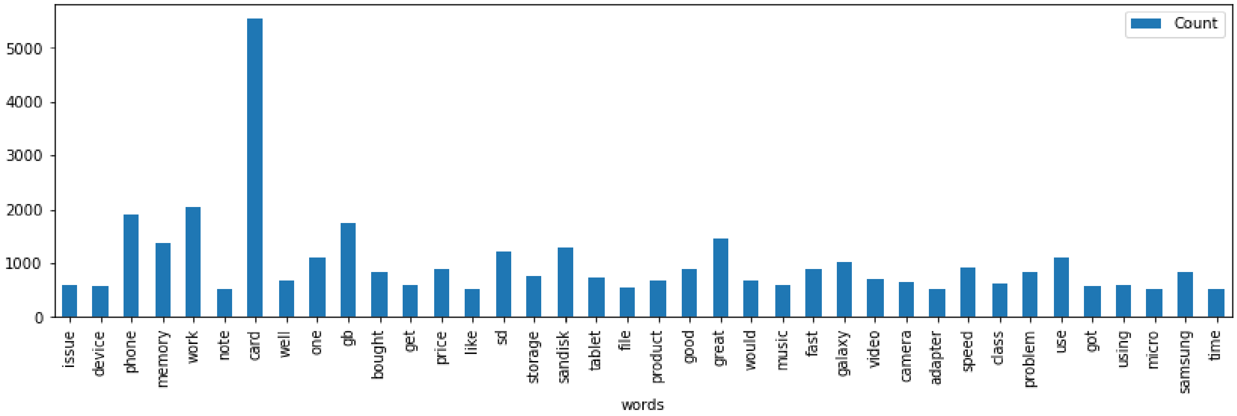
Since we will focus on the sentiment of reviews, ‘unix review time’, ‘reviewer names’, and ‘asin (product id)’ columns are dropped from the dataset. Since ‘helpful’ and ‘helpful_yes’ columns included same information, ‘helpful’ column also is dropped from the dataset.

Not to have a measurement problem in our analysis, we removed punctuations, numbers, stop words using regular expression from review tests and normalized the case folding by converting upper cases in review text into lower cases. We then, removed the words that make no sense (rare words). In the last step, we tokenized the reviews and then reduced the words into their roots (lemmitization).

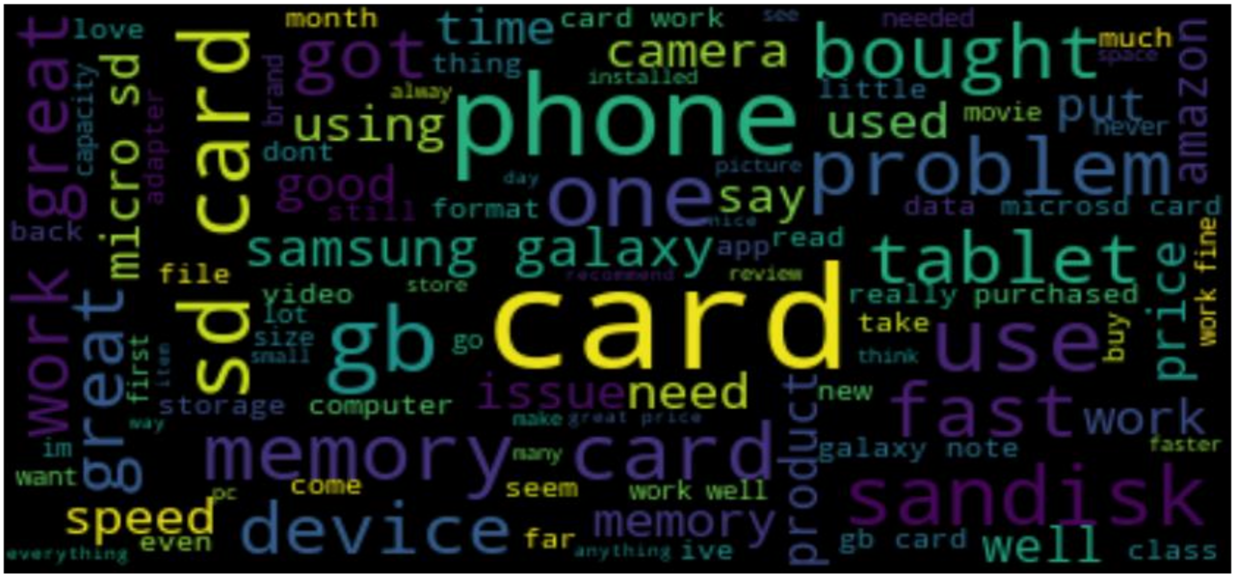
5- Exploratory Data Analysis (EDA)

We started to make some visualizations to observe the dataset. Word counts were calculated using term frequencies and we visualized the most frequent words using bar plot and word cloud.

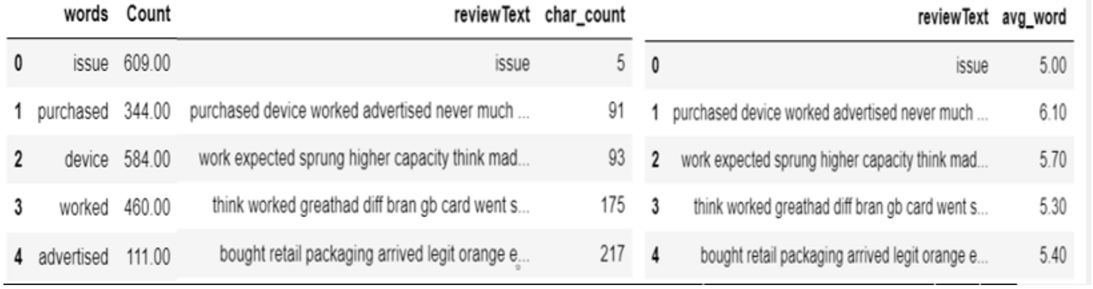
Word Count



Word Cloud



Then, to analyze the sentiment of the reviews, we first converted review texts into upper letter case and calculated the character counts just to have an idea about the length of each review. A function to calculate the average review length is also created.



	words	Count
0	issue	609.00
1	purchased	344.00
2	device	584.00
3	worked	460.00
4	advertised	111.00

	reviewText	char_count
0	issue	5
1	purchased device worked advertised never much ...	91
2	work expected sprung higher capacity think mad...	93
3	think worked greathad diff bran gb card went s...	175
4	bought retail packaging arrived legit orange e...	217

	reviewText	avg_word
0	issue	5.00
1	purchased device worked advertised never much ...	6.10
2	work expected sprung higher capacity think mad...	5.70
3	think worked greathad diff bran gb card went s...	5.30
4	bought retail packaging arrived legit orange e...	5.40

6- Sentiment Analysis

The goal in the sentiment analysis is to mathematically capture the emotional tone of a content about a particular topic, product, or service. This's why it is called emotion AI or opinion mining. Every word in every review text has a valance of positive or negative and they are labeled as negative or positive after a careful sentiment score analysis.

In order to do this task for our dataset, we utilized a sentiment intensity analyzer, called VADER, which is a pre-trained sentiment analyzer that comes built-in with NLTK. After that, polarity scores are calculated. Polarity scores uses 'pretty rate model' which is a pre-trained model in NLTK to determine the meanings of words and assign them a positive or negative value based on their context. This is important because certain wors may have both positive and negative connotations depending on the situation. The focus of the polarity scores is on the compound score which ranges between -1 and 1. If the score is less than 0, the text is labeled as negative, while scores above 0 are labeled positive. Then, we do the followings:

- Polarity scores are created for every word in every review and a compound value is created for each review. Then, a sentiment label column was created based on these polarity scores.
- After checking the value counts of sentiment distribution, we confirmed that we have imbalanced data because we have 3946 positive (80.2%) and 969 negative (19.8) reviews. Therefore, we should pay extra attention to the metrics we use in modeling. Precision score is a better metric to evaluate false positives and the performance of the model in general.
- Reviews are grouped by overall ratings; we get almost equal average scores of positive and negative comments (4.09/4.71).
- Since the labels (neg/pos) has string values we then, transformed the sentiments to numbers using Label encoder and created a new column to list all these integer values.

7- Algorithms and Machine Learning Models

This is a sentiment analysis study; therefore, the final data contains only review text and their sentiment labels. After splitting the data as X (review texts) and y (sentiment labels), we need to perform word vectorizers and convert the text data into a numerical/measurable format to be able to perform machine learning models.

In order to do this, we utilized the pipeline classes using skit-learn library. Using CountVectors, first, ngrams are created to break words into smaller chunks (potential features consist of word phrases) so we can count ngrams, words, and characters. Ngrams consist of both unigrams (single word) and bigrams (combination of two words). After that, Tf-IDF (word vector generation method) is used to standardize/normalize the data and to eliminate potential biases created by count vectorizer method.

The data has been tested using following models:

Stochastic Gradient Descent (SGD) is an optimization algorithm is to find the values of the weights and biases that minimize the loss function by iteratively updating parameters using random subset of the training data. This allows the classifier to make more accurate predictions in new data. It is commonly used in ML studies for training linear classifiers, including binary classifiers.

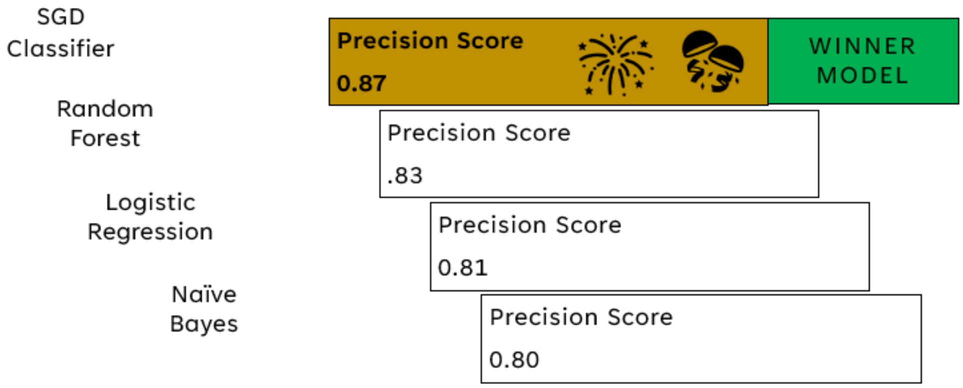
Random forest is an ensemble machine learning algorithm that combines multiple decision trees to make predictions and can handle non-linear relationships and interactions between features.

Logistic regression is a type of linear regression used for binary classification, modeling the probability of an output variable based on input variables.

Naïve Bayes is a probabilistic classification algorithm based on Bayes' theorem and used in natural language processing for tasks such as sentiment analysis and spam detection.

8- Evaluation Metrics

Precision is the major evaluation metric to evaluate the model performance and select the winning model because we have unbalanced data. 80% of the sentiments are positive and the cost of false positives is high for sellers specifically in online retail stores. According to the analysis, we got the following model performances:



In summary, our analysis of four machine learning models revealed that the stochastic gradient descent classifier achieved the best sentiment classification performance with a precision of 0.87. Consequently, we selected this classifier as the top-performing model and proceeded to fine-tune its hyperparameters using the GridSearchCV library. After exhaustive hyperparameter tuning, we obtained an improved precision score of 0.92 based on the best cross-validation score.

9- Conclusion and Future Works

As anticipated, we have successfully developed a sentiment analysis model that can accurately predict the sentiment of Amazon reviews about SanDisk memory cards. Our analysis indicates that the SGD Classifier has performed well in classifying the reviews into their respective sentiment classes. The precision achieved by our sentiment classification model is very satisfactory for end-users.

Our analysis has revealed a predominantly positive sentiment towards SanDisk memory cards, providing valuable insights into customer satisfaction. The sentiment analysis model developed in this study can be used in customer service to enhance customer satisfaction.

Future research can focus on exploring additional vectorization techniques (such as word2vec, GloVe, and Bert embedding) to further improve model performance. Additionally, incorporating higher-order n-gram methods, such as trigrams, can deepen our understanding of the review context and open up new research opportunities.

10- References

https://www.kaggle.com/code/halimedogan/sentiment-analysis-with-nlp-for-amazon-reviews/notebook#5.-Modeling

https://www.analyticsvidhya.com/blog/2021/05/natural-language-processing-step-by-step-guide/#h2_5

https://erleem.medium.com/nlp-complete-sentiment-analysis-on-amazon-reviews-374e4fea9976

https://towardsdatascience.com/sentiment-analysis-on-amazon-reviews-45cd169447ac

https://www.kaggle.com/code/furkannakdagg/nlp-sentiment-analysis-tutorial

https://www.kaggle.com/code/furkannakdagg/sentiment-analysis-with-lstm