

**SENTIMENT ANALYSIS USING AMAZON REVIEW DATASET WITH ML AND DL**

SHARANYAK PODDER

ROLL NO. 19287

Department of Data Science & Engineering

INTRODUCTION

In this project I will use Natural Language Processing (NLP) techniques to find broad trends in the written thoughts of the customers. The goal in this project is to predict whether customers recommend the product they purchased using the information in their review text.

One of the challenges in this project is to extract useful information from the **"Review Text"** variable using text mining techniques. The other challenge is that we need to convert text files into numeric feature vectors to run machine learning algorithms.

DATA

The data is a collection of 22641 Rows and 10 column variables. Each row includes a written comment as well as additional customer information. Also each row corresponds to a customer review, and includes the variables.

FEATURES OF THE DATA

**1) Clothing ID:** Integer Categorical variable that refers to the specific piece being reviewed.

**2) Age:** Positive Integer variable of the reviewers age.

**3) Title:** String variable for the title of the review.

**4) Review Text:** String variable for the review body.

**5) Rating:** Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst, to 5 Best.

**6) Recommended IND:** Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.

**7) Positive Feedback Count:** Positive Integer documenting the number of other customers who found this review positive.

**8) Division Name:** Categorical name of the product high level division.

**9) Department Name:** Categorical name of the product department name.

**10) Class Name:** Categorical name of the product class name.

EXPLORATORY DATA ANALYSIS AND VISUALIZATION

I have done a quick exploratory data analysis using pandas profiling which is an open-source software. I have done a detailed EDA later also. I have visualized the datasets and tried to understand the relations between features. I have dropped values of unimportant variables.

MAKING ML AND DL MODELS

Before modelling, as data pre-processing steps I will need to **train-test split** and need to perform **vectorization.** But I will perform the vectorization for the first time.

Machine learning algorithms most often take numeric feature vectors as input. Thus, when working with text documents, we need a way to convert each document into a numeric vector. This process is known as text vectorization. Commonly used vectorization approach that we will use here is to represent each text as a vector of word counts.

At this moment, I have our review text column as a token (which has no punctuations and stop-words). I can use Scikit-learns Count-Vectorizer to convert the text collection into a matrix of token counts. I can imagine this resulting matrix as a 2-D matrix, where each row is a unique word, and each column is a review.

**For Deep learning model, I will implement embedding layer for all words.**

After performing data pre-processing, I will build your models using following classification algorithms:

* Logistic Regression,
* Support Vector Machine,
* Random Forest,
* Ada Boosting
* Deep Learning Model.

CONCLUSION

**In this project we have used sentiment analysis to determine whether the product is recommended or not. I have used different machine learning algorithms to get more accurate predictions and deep learning algorithm for comparing it with machine learning models.**

REFERENCE

* [https://pandas.pydata.org/docs/reference/api/pandas.read\_csv.html](about:blank)
* [https://www.researchgate.net/publication/49814836\_Problematic\_standard\_errors\_and\_confidence\_intervals\_for\_skewness\_and\_kurtosis](about:blank)
* [https://www.researchgate.net/publication/304577646\_Young\_consumers'\_intention\_towards\_buying\_green\_products\_in\_a\_developing\_nation\_Extending\_the\_theory\_of\_planned\_behavior](about:blank)
* [https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/Simon](about:blank)
* [https://www.researchgate.net/publication/263372601\_Resistance\_motivations\_trust\_and\_intention\_to\_use\_mobile\_financial\_services](about:blank)
* [https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html](about:blank)
* [https://machinelearningmastery.com/power-transforms-with-scikit-learn/](about:blank)
* [https://en.wikipedia.org/wiki/Dummy\_variable\_(statistics](about:blank))
* [https://www.displayr.com/what-are-dummy-variables/](about:blank)
* [https://stattrek.com/multiple-regression/dummy-variables.aspx](about:blank)
* [https://www.statisticshowto.com/dummy-variables/](about:blank)
* [https://en.wikipedia.org/wiki/Feature\_scaling](about:blank)
* Neural Networks from Scratch in Python (by Kinsley § Kukiela)
* Practical Statistics for Data Scientists (by Bruce & Gedeck)
* Applications of Deep Neural Networks (by Jeff Heaton)
* Applied Predictive Modelling (by Kuhn & Johnson)
* Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (by Aurélien Géron)