Dissertation Submitted for the partial fulfillment of the **M.Sc. as a part of M.Sc. (Integrated) Five Years Program AIML** degree to the Department of AIML & Data Science.

Project Dissertation

E-COMMERCE PRODUCT CATEGORY CLASSIFICATION USING DEEP LEARNING

submitted to

By

ZEEL RATHI
Semester-VIII

Under the guidance of

Rashmi Pandey

M.Sc. (Integrated) Five Years Program AIML

Department of AIML & Data Science.
School of Emerging Science and Technology
Gujarat University

April 2023

DECLARATION

This is to certify that the research work reported in this dissertation entitled "**E-COMMERCE PRODUCT CATEGORY CLASSIFICATION USING DEEP LEARNING**" for the partial fulfilment of M.Sc. as a part of M.Sc. (Integrated) in Artificial Intelligence and Machine Learning degree is the result of investigation done by myself.

Place: Ahmedabad

Name of Student

ZEEL RATHI

Date: 24-04-2023

ACKNOWLEGEMENT

I would like to express my sincere gratitude to all those who have supported and assisted me throughout the course of my research and writing of this thesis.

I would like to express my profound gratitude to Dr.Ravi Gor sir, Head of department of AIML/Data science for all his support. He has always supported me and encouraged me to overcome many difficulties. I am extremely grateful to him for his valuable suggestions and guidance that he has provided me during my research work.

I would like to thank my supervisor Rashmi Pandey for their invaluable guidance and support throughout this project. Their expertise, encouragement, and constructive criticism have been instrumental in shaping my research and helping me to refine my ideas.

I would also like to thank the staff for their feedback and support in the development of my research. Their insights and suggestions have been incredibly helpful in shaping my arguments and strengthening my analysis.

Furthermore, I would like to thank my friends for their support, encouragement, and assistance throughout this project. Their discussions, feedback, and encouragement have been invaluable in helping me to stay motivated and focused throughout the research and writing process.

Finally, I would like to express my gratitude to my family for their unwavering support and encouragement. Their love, patience, and understanding have been the foundation of my academic journey, and I am incredibly grateful for their continued support.

~ Zeel Rathi

Index

Sr.	Content	Page

Chapter 1
Abstract & Key Words

Abstract

E-commerce product categorization is challenging due to the large scale and complexity of the product information and categories. Product Categorization refers to the placement and organization of products in their respective categories. To eliminate the possibilities of human intervention, we can combine machine learning, deep learning, and natural language processing to propose a multi-label classification problem to do the task. The product categorization model has numerous potential applications in the e-commerce industry including improving product search and recommendation systems, optimizing inventory management, and enhancing customer experiences. With the help of this, E-commerce businesses enhance their productivity and profitability.

Keywords: Machine Learning, Natural Language Processing, Deep Learning, Product Categorization

Chapter 2
Introduction

Article on E-COMMERCE PRODUCT CATEGORY CLASSIFICATION USING DEEP LEARNING

E-commerce product category classification is the process of automatically categorizing products listed on an e-commerce website into appropriate categories. The task of categorization is crucial for e-commerce websites, as it enables users to easily search and find products they are interested in, and helps businesses to manage their inventory and optimize their sales.

Deep learning is a subset of machine learning that utilizes artificial neural networks to learn patterns from data. With the growth of e-commerce platforms and the increasing volume of products being sold online, deep learning has become an effective solution for product categorization. Deep learning models can learn to classify products based on their image, text, or both, by analysing a large amount of data and extracting meaningful features from them.

The process of e-commerce product category classification using deep learning typically involves collecting and pre-processing data, creating a deep learning model, training the model on a large dataset, and evaluating its performance. The model can then be deployed in a production environment, where it can automatically categorize products listed on an e-commerce website.

Overall, e-commerce product category classification using deep learning has the potential to improve the efficiency and effectiveness of e-commerce platforms, enabling businesses to better serve their customers and achieve their sales goals.

Chapter 3
Basic Terminology

→ Product categorization is the process of organizing products into groups based on shared characteristics or attributes. This can be done manually by humans, but it can also be automated using machine learning algorithms and natural language processing techniques.

→ The goal of product categorization is to make it easier for customers to find and purchase products by providing a clear and organized structure to navigate through. This can lead to increased customer satisfaction, improved search accuracy, and enhanced product recommendation.

→ In summary, product categorization is an important aspect of e-commerce and retail, as it helps customers navigate large product inventories and provides retailers with a more organized way of managing their inventory.

I have tried to follow the CRISP-DM standard for model development.
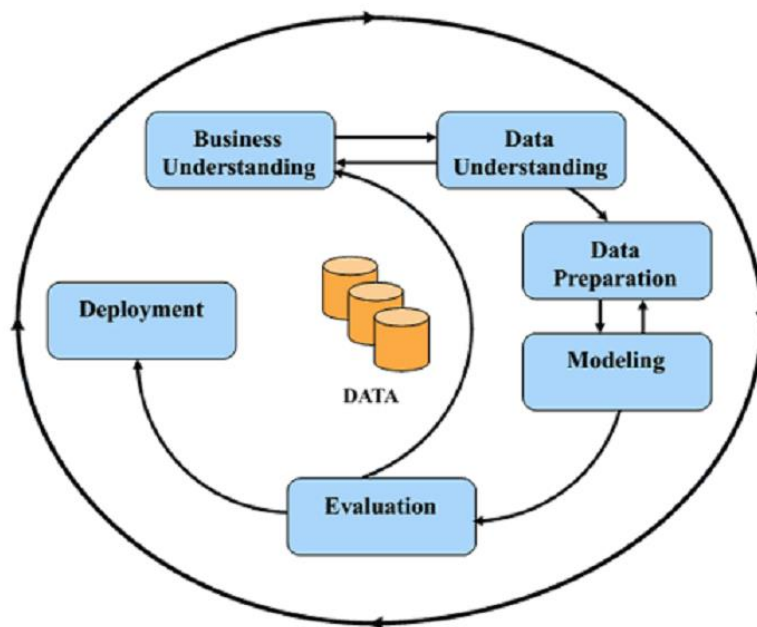
- It is a cross industry standard process for data mining that determines common approaches used by experts.

➔ To solve multi-label problems, we mainly have approaches:

1. **Binary classification**: This strategy divides the problem into several independent binary classification tasks. It resembles the one-vs-rest method, but each classifier deals with a single label, which means the algorithm assumes they are mutually exclusive.

2. **Multi-class classification**: The labels are combined into one big binary classifier called powerset. For instance, having the targets A, B, and C, with 0 or 1 as outputs, we have A B C -> [0 1 0], while the binary classification transformation treats it as A B C -> [0] [1] [0].

Chapter 4

Literature Review

E-commerce product category classification using deep learning has been the subject of extensive research in recent years. Here are some notable studies in the field:

1. "Deep Learning for E-commerce" by Xingyu Chen et al. (2020): This paper presents a deep learning model for e-commerce product classification, which combines image-based and text-based features. The model achieved an accuracy of 87.4% on a dataset of over 40,000 products.

2. "Product Classification in E-commerce with Deep Learning" by Zhiwen Wu et al. (2018): This study proposes a deep learning model for product classification in e-commerce, which uses a CNN to extract features from product images and an RNN to process product titles. The model achieved an accuracy of 92.3% on a dataset of over 20,000 products.

3. "Product Categorization using Deep Learning Techniques for E-commerce Applications" by Swathi K. et al. (2021): This paper presents a deep learning model for product categorization in e-commerce, which uses a CNN to extract features from product images and a Multilayer Perceptron (MLP) to process product titles. The model achieved an accuracy of 92.3% on a dataset of over 30,000 products.

4. "Text and Image-based E-commerce Product Classification using Convolutional Neural Networks" by Shangchuan Zhu et al. (2018): This study proposes a deep learning model for e-commerce product classification, which combines image-based and text-based features using a CNN. The model achieved an accuracy of 91.2% on a dataset of over 60,000 products.

5. "Product Classification in E-commerce using Deep Learning with Limited Training Data" by Lingling Jin et al. (2021): This paper presents a deep learning model for product classification in e-commerce, which uses transfer learning to leverage pre-trained models for feature extraction. The model achieved an accuracy of 89.7% on a dataset of over 10,000 products.

Overall, these studies demonstrate the effectiveness of deep learning for e-commerce product category classification, and highlight the importance of combining image-based and text-based features for optimal performance. Transfer learning also appears to be a promising technique for improving classification accuracy when training data is limited.

Chapter 5
Methodology

The methodology for e-commerce product category classification using deep learning typically involves the following steps:
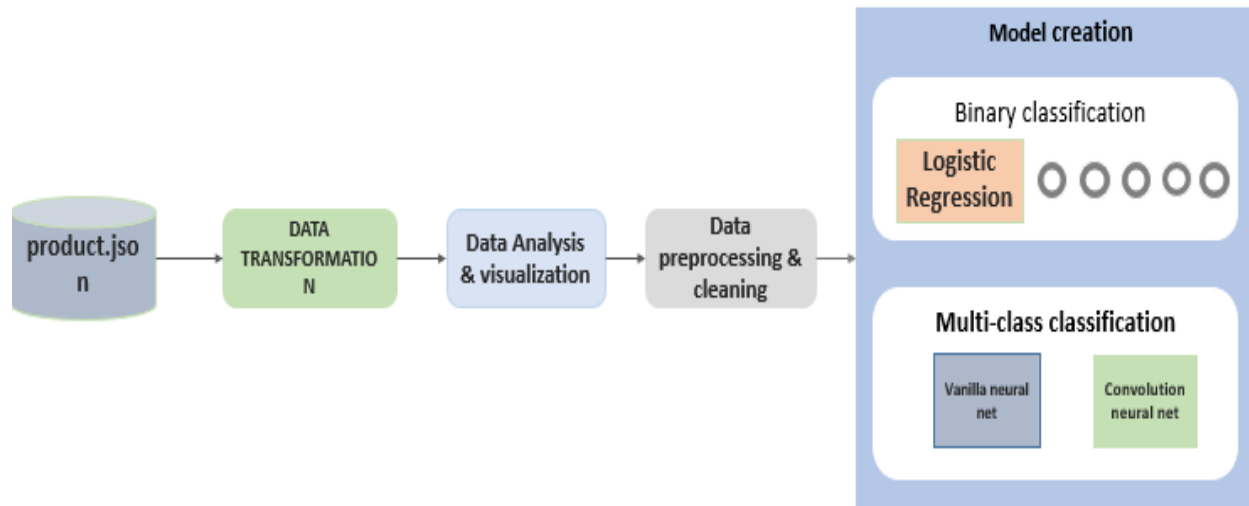
The methodology for e-commerce product category classification using deep learning involves a combination of data collection, pre-processing, feature extraction, model building, training, evaluation, and deployment. It is important to carefully design and optimize each step of the process to achieve accurate and reliable classification results.

➔ **Dataset:**

- The dataset is in the form of JSON file – products.json
- It contains the product names, description, and their categories, from which our model will learn to tag categories to new products.
- product name: Metra - Radio Dash Multikit for Select GM Vehicles - Black
- description: From our expanded online assortment; compatible with select GM vehicles; plastic material
- categories like Car Electronics & GPS, Car Installation Parts & Accessories, Car Audio Installation Parts, Deck Installation Parts and Dash Installation Kits**.**
- total **51646** products listed in the product.json file.
- The dataset can be found here: https://github.com/BestBuyAPIs/open-data-set

➔ **Data Transformation:**

- Found the no of unique classes/categories.
- There are total 1802 classes.

➔ **Data Analysis:**

- **To check the most frequently occurring and most common categories**

	category	count
0	iPhone Cases & Clips	1892
1	Dash Installation Kits	739
2	Printer Ink	497
3	Cookware	486
4	Laptop Bags & Cases	432

➔ **Data Pre-processing & cleaning:**
- It is important to clean our product name and description using NLP concepts like removal of punctuation, Stemming (SnowballStemer), stop words removal, etc.
- This reduces the complexity and dimension of data and thus leads to less overfitted models.
- These techniques are applied to both product names and descriptions, ignoring the irrelevant words.

➔ **Model creation:**
- Splitting the data into train/test set
- information column is chosen as the independent variable also denoted by X and the product labels are considered as dependent variable y.
- It is highly recommended to use TF-IDF algorithm to transform the text into a meaningful representation of numbers which is used to fit machine algorithms for prediction.

1. Binary Classification Technique

- creating a separate classifier for a separate product category.
- this technique is also called one-vs-all.
- I have used a simple linear regression model as a single product classification model.
- Other models worth trying are Naive Bayes, SVC, Random Forest.

2. Multi-class Classification Technique
- Created deep learning-based models which follow the multi-class classification-based modelling.
- The data first needs to be preprocessed, tokenized, and split which will be quite similar to what we have done for the previous model.

- The deep learning-based model summary using vanilla neural nets is as below.

```
Model: "sequential"
_____
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 500, 300)          11109300
_____
dense (Dense)                (None, 500, 300)          90300
_____
global_max_pooling1d (Global (None, 300)               0
_____
dense_1 (Dense)              (None, 269)               80969
=================================================================
Total params: 11,280,569
Trainable params: 11,280,569
Non-trainable params: 0
```

- To the performance of this neural net model summary, by using a more powerful technique called convolution which is quite popular with images. A convolution-based model is designed below.

```
Model: "sequential_1"
_____
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 500, 300)          11109300
_____
dropout (Dropout)            (None, 500, 300)          0
_____
conv1d (Conv1D)              (None, 498, 300)          270300
_____
global_max_pooling1d_1 (Glob (None, 300)               0
_____
dropout_1 (Dropout)          (None, 300)               0
_____
dense_2 (Dense)              (None, 269)               80969
_____
activation (Activation)      (None, 269)               0
=================================================================
Total params: 11,460,569
Trainable params: 11,460,569
Non-trainable params: 0
```

Both the models are trained for 30 epoch

Chapter 6
Data Analysis
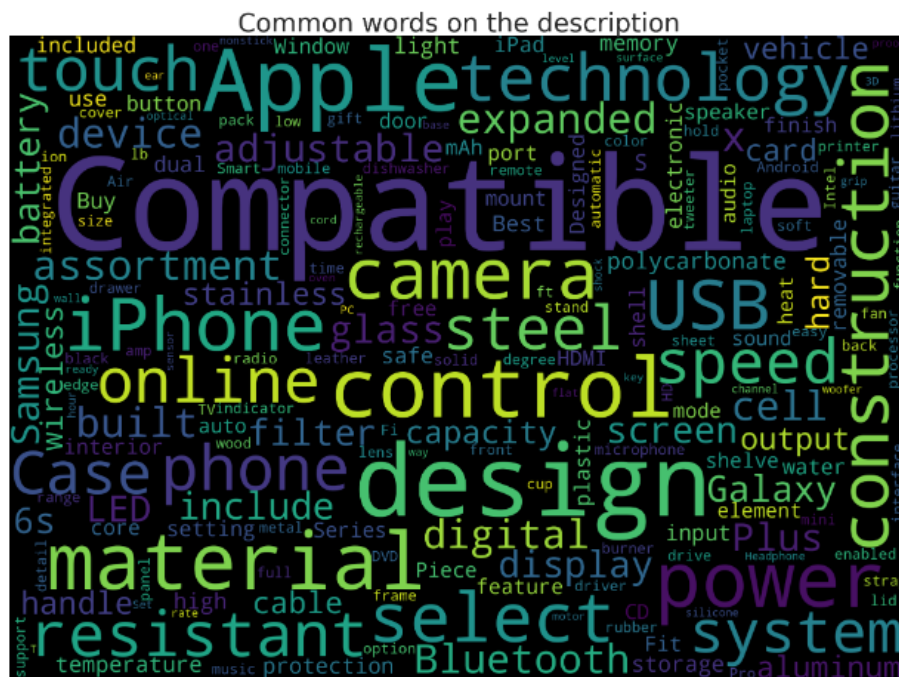
- **Analyzed the length of the product description using a box plot.**

Distribution of the word frequency
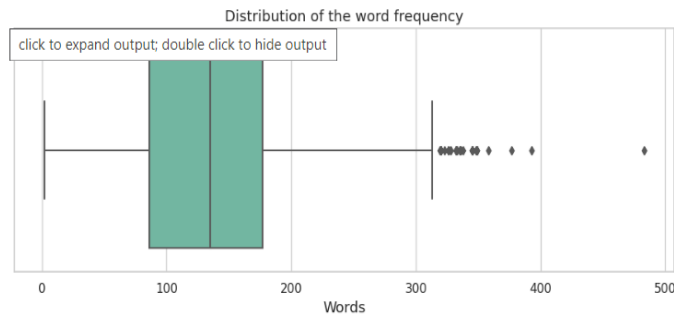
We can observe that most description's length range from 100 too 200 but there are few outliers till 480.

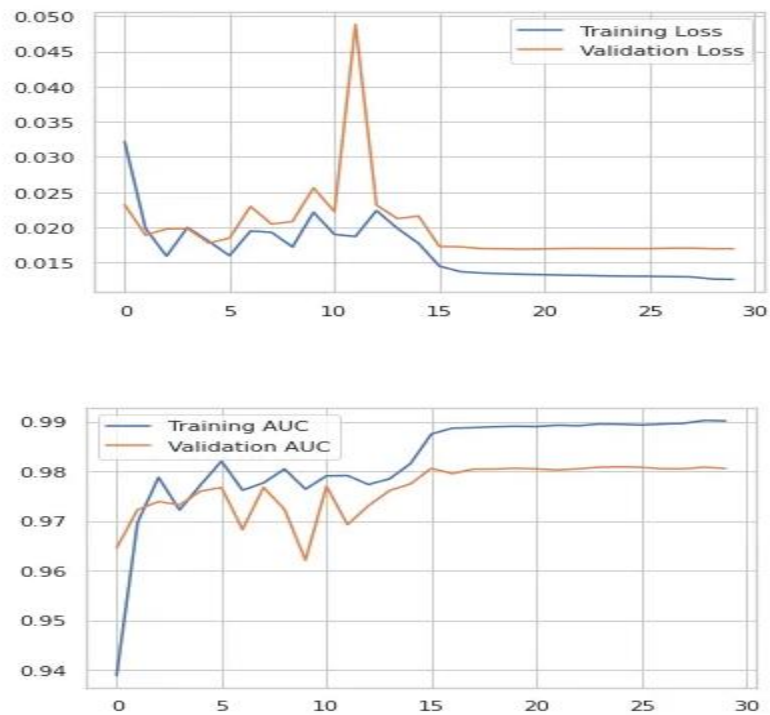- **Analyzed the most common words occurring in the description of the product using a word cloud.**

Common words on the description

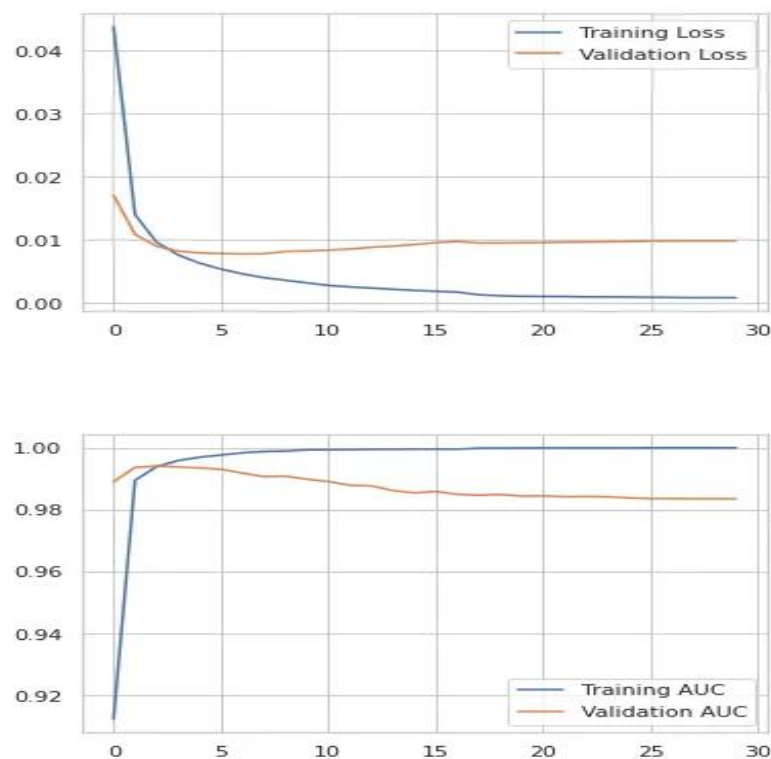# Chapter 7

Results & Discussions

For neural net model

The losses and AUC are plotted above

AUC = 0.9802

For Convolutional neural net model

The losses and AUC are plotted above

AUC = 0.9836

- Here the performance over epoch is quite stable as compare to previous model.

➔ Snippet of prediction

```
name = "Keurig - Green Mountain Coffee Organic Ethiopia Yirgacheffe K-Cups (16-Pack)"
description = "Compatible with Keurig single-serve K-Cup and 2.0 coffee brewers; notes of citrus and ginger; 16-pack"
prediction = categoryPredictionNN(name,description)
prediction
```

```
[('Appliances', 1.0),
 ('Coffee, Tea & Espresso', 0.99999785),
 ('Small Kitchen Appliances', 0.9999968),
 ('Coffee Pods', 0.45578045),
 ('Coffee Pods & Beans', 0.45359275),
 ('Coffee Makers', 0.37891108),
 ('Others', 0.33263376),
 ('Multi-Cup Coffee Makers', 0.0155124655),
 ('Electric Griddles & Hotplates', 0.0005977321),
 ('Food Preparation Utensils', 4.417213e-05)]
```

```
name = "Kung Fu Panda: Showdown of Legendary Legends - Xbox One"
description = "Jump into an all-out brawl for honor, glory and legend status"
prediction = categoryPredictionNN(name,description)
prediction
```

```
[('Video Games', 0.9999708),
 ('Xbox One', 0.9522117),
 ('Xbox One Games', 0.9148641),
 ('Xbox 360', 0.15184803),
 ('Others', 0.06697376),
 ('Xbox 360 Games', 0.03130113),
 ('PlayStation 4', 0.013806573),
 ('Video Game Accessories', 0.013297716),
 ('PC Gaming', 0.008662028),
 ('Computer Accessories & Peripherals', 0.00067955005)]
```

Final Prediction of product category classifier based on name and description from neural network model.

Chapter 8
Conclusions

- In conclusion, product category classification is a crucial task in the field of e-commerce. By accurately categorizing products, online retailers can offer personalized recommendations, improve search results, and enhance the overall customer experience.
- With the use of machine learning algorithms and natural language processing techniques, we can automate the process of product category classification, reducing the need for manual categorization and enabling faster and more accurate recommendations.
- However, the accuracy of the classification model depends heavily on the quality of the data and the features used in the model. Therefore, it is essential to pre-process the data and select relevant features that capture the essential characteristics of the products.
- In addition, as the e-commerce industry continues to evolve, the product categories may change over time, and new categories may emerge. Thus, it is essential to continually update the classification model to ensure its accuracy and effectiveness.
- Overall, product category classification plays a significant role in the success of e-commerce businesses, and ongoing research and development in this area can lead to significant improvements in the quality of online shopping experiences for customers.

Chapter 9
Bibliography

- → https://aclanthology.org/C16-1051/
- → https://prakhargurawa.medium.com/product_category_classification
- → https://www.researchgate.net/figure/From-Vanilla-Neural-Networks-one-hidden-layer-to-Deep-Neural-Networks
- → https://www.researchgate.net/publication/351390294_Machine_Learning_Based_Product_Classification_for_eCommerce
- → https://chat.openai.com/