An Autonomous Approach for Type Categorization of Product Data Based on Numerous Text and Attribute Factorizations

Sudesna Baruah[1], *Bagya Lakshmi V.*[2], *Navneeth Devaraj*[3],
Gnanaprakash A.[4], *and Deepak Kumar Jayaram*[5]

Abstract

Product categorization or Product classification is an area of Natural Language Processing. This special area happens to have a lot of challenges, especially for e-commerce organizations. Scientists and researchers have actively worked on this area over the years using various machine learning & artificial intelligence techniques. Product categorization is the placement and organization of products into their respective categories.

Keywords : **Context Builder, Continuous Bag of Words (CBOW), Convolutional Neural Network (CNN), Deep Neural Network, Natural Language Generation (NLG), Natural Language Processing (NLP), Product Categorization, Shallow Neural Network, Skip Gram, Vocabulary Builder, Word2Vec**

I. INTRODUCTION

Product categorization is placing and organizing products into their respective categories or hierarchies. This sounds very simple like choosing a correct category or department for a product. However, the process is highly complicated as there is a huge volume of products in a retail website. Additionally, there are some products that belong to multiple categories as well.

Product Categorization in E-Commerce Platform

Now, we will try to understand the concept of product categorization. An item in a retail website falls under

Manuscript Received : February 22, 2022 ; Revised : March 15, 2022 ; Accepted : March 17, 2022. Date of Publication : April 5, 2022.

S. Baruah[1], *Data Scientist,* Retail Cluster, AI & Machine Learning, Tata Consultancy Services Pvt. Ltd. Magnum SEZ IT Park, Pallikaranai Marshland, Kotivakkam, Chennai - 600 096, Tamil Nadu, India. Email : baruahsudesna@gmail.com ;
ORCID iD : https://orcid.org/0000-0001-7298-1344
B. Lakshmi V.[2], *Principal Scientist,* Retail Cluster, AI & Machine Learning, Tata Consultancy Services Pvt. Ltd., Magnum SEZ IT Park, Pallikaranai Marshland, Kotivakkam, Chennai - 600 096, Tamil Nadu, India.
Email : bagyalakshmi_vasudevan@yahoo.com ; ORCID iD : https://orcid.org/0000-0002-3969-771X
N. Devaraj[3], *Functional Consultant,* Retail UK, Tata Consultancy Services Limited, Magnum SEZ IT Park, Pallikaranai Marshland, Kotivakkam, Chennai - 600 096, Tamil Nadu, India. Email : navneeth.d@tcs.com ;
ORCID iD : https://orcid.org/0000-0002-3826-1615
G. Arumugam[4], *Functional Consultant*, Retail Cluster, Tata Consultancy Services Limited, Magnum SEZ IT Park, Pallikaranai Marshland, Kotivakkam, Chennai - 600 096, Tamil Nadu, India. Email : gnanaprakash.a@tcs.com ;
ORCID iD : https://orcid.org/0000-0001-6850-2195
D. K. Jayaram[5], *Functional Consultant*, Retail Digital & CPG, Tata Consultancy Services Limited, Magnum SEZ IT Park, Pallikaranai Marshland, Kotivakkam, Chennai - 600 096, Tamil Nadu, India. Email : d.jayaram@tcs.com ;
ORCID iD : https://orcid.org/0000-0001-9373-2523

DOI : https://doi.org/10.17010/ijcs/2022/v7/i2/169680

some category. Also, there are situations where it could fall under multiple categories. Below are examples of some product hierarchies under which a product is found in retail web page. These show an instance of how a product might appear under multiple categories or hierarchies.

For example, "Men's Sports Sneakers" could use the following path and categories:

Shoes & Accessories > Men's Shoes > Men's Athletic Shoes

It follows 3 levels of product hierarchy.

However, it could also follow this path:

Sporting Goods > Team Sports >Sports Equipment > Shoes > Men's Sports Shoes

It follows 5 levels of product hierarchy.

Deciding on the right path will totally depend on how the customers look for the desired products on the website. It also depends on the categories. In addition to choosing the correct category, product categorization also deals with the organization of those categories.

There are several reasons why the concept of product categorization is important for retailers. Through the accurate classification of their products in a website, they can increase the conversion rates, strengthen their search engine, and improve their site's Google ranking as it leads to the following:

- Improves the user experience
- Improves search relevance
- Help customers find the website

A. Statement of the Problem

Retailers are nowadays looking for solutions that could ideally help them in placing a new product under certain sections of their retail website. Retailers very often have to reshuffle products or even introduce new products for sale and they have to involve manual efforts for this. This proposed solution helps in placing a product, new or existing, under a category or end level of a product hierarchy or breadcrumbs.

B. Objective

The aim of the proposed solution is to categorize a given product under a category within a retail website. The solution involves footwear products for explanation, but it is applicable across all sections of retail like apparels, accessories, grocery etc. provided we have similar kind of data structure.

C. Project Goal and Scope

This project aims at achieving auto categorization of retail footwear products. It targets towards minimizing human efforts and also aids an existing machine in achieving appropriate data categorization.

D. System Overview

This proposed work involves a Word2Vec algorithm for the purpose of procuring all the footwear products, given a category based on semantic similarity based relation. The model is made contextually aware of the data fed to it through a dictionary containing category and footwear types/sub-types as key-value pair. This dictionary data is assigned weights for setting the importance measure in order to derive a suitable probability score against each product given a product category. This score will aid in understanding how closely related a particular product is with respect to a category. For example, if we consider the category 'casual', the model should yield results like 'Laced tie-up Sneakers' (0.95), 'Criss-Cross Slippers' (0.89), and 'Wedged-heel Loafers' (0.87).

Here, each footwear item is a casual wear and the associated probability score helps in understanding how closely each item is related to the category 'casual'.

E. NLP Techniques Involved

The techniques involved here are:

(a) Contextual Dictionary Building (manual)

(b) Weights assignment

(c) Word2Vec Algorithm

II. LITERATURE REVIEW

Here are some relevant papers and articles that have been read before working this technology out for generating copies for a given product.

Relevant Works

1) Systems and Methods for Multi-modal Automated Categorization : The proposed solution in [1] can perform classification based on a plurality of different signals, including, for example, webpage text, images, and website structure or category organization. For text classification, the system can use one or more classifiers. For example, in the form of a Bag-of-Words (BoW) based word representation and a word vector embedding (e.g., WORD2VEC) based representation. The inputs used for text classification are product titles, descriptions and breadcrumbs of webpages. For image classification, the system can use an image classifier. For example, an 8-layer Convolution Neural Network (CNN) receives images of the items from the webpages as input. A classifier fusion strategy can be used to combine the results text classification and the image classification results and generate a content likelihood of the item belonging to a specific category (for example, the item belongs to women's hats). To exploit latent category organization provided by website operators or owners (for example, merchants for product webpages), the systems can also use crawl graph properties of webpages to estimate a probability distribution for item categories associated with the webpages. To address issues associated with scarcity of labeled data, an unsupervised as well as a semi-supervised model can be used to compute this prior probability distribution. The probability distributions can be combined with content likelihood to yield a holistic categorization output.

2) Everyone Likes Shopping! Multi-class Product Categorization for e-Commerce : The objective in [2] is to build a multi-class classifier which can predict the category of new product that is to be launched in a retail website. The objective is fulfilled through a set of titles describing products and a product taxonomy. The model takes a product title as an input and returns the whole product category hierarchy as output.

The product categorization task is modeled as a classification problem, where for a given collection of labelled training examples, the objective is to learn a classification function.

For our experiments, we used two multiclassification algorithms from the large scale machine learning toolkit Vowpal Wabbit (Beygelzimer et al., 2009): one-against-all (OAA) and error correction tournament (ECT). OAA tries to reduce the Kway multi-classification problem into multiple binary classification tasks. This is done by iteratively classifying each product title for category ‘x’ and comparing it against all other categories.

3) A Machine Learning Approach for Product Matching and Categorization **[3] :** The given approach deals with product matching and categorization. This proposed solution consists of three main steps, that is, feature extraction, calculating similarity feature vectors, and finally classification.

The approach can be elaborated through a workflow that runs in two phases: training and application. The training phase starts with preprocessing both the structured and the unstructured web product descriptions. Then, we build four feature extraction models as follows:

(i) Dictionary-Based : This is built using the product attributes and the values.

(ii) Conditional Random Field (CRF) : This is built from the discrete product features.

(iii) CRF with Text Embeddings : This is built for handling the dynamic pattern of the product descriptions and for enhancing the training in the model ii.

(iv) Image Feature Extraction Model : This has been built for handling images and image embeddings.

4) Managing Product Feeds : Classifying Items using Word2Vec : In [4] we employ both image and text classification techniques. We focus on how we have applied a specific Natural Language Processing technique to boost the performance of our text classifiers.

In a *Bag-of-Words* model, a document is represented in a vector format, where each dimension corresponds to a single word. For each product category and subcategory of our retail product catalogue, such as shoes or smart phone, we have a manually curetted dictionary that contains words that represent a category.

For the task of classifying whether a given item belongs to a given category or not, a feature vector is built from the item name and description that consists of the count of occurrences of each dictionary word in the item text.

However, the model has some known limitations:

➥ words are equally distant; and

➥ vectors are often quite sparse, which is not optimal for any machine learning algorithm.

Hence, neural word embeddings are the most sought-after models. This approach represents single words using low dimensional continuous vectors that can be learnt using a shallow neural network. One such implementation is Word2Vec, which comes in two flavors:

(1) *Skip-gram*, where the task is to predict the surrounding words given an input word; and

(2) *Continuous Bag of Words*, where the task is to predict a word given the surrounding words.

In comparison to bag of words models Word2Vec offers the following advantages:

➥ Vectors are continuous

➥ Similar words are close in vector space; and

➥ Learning the embedding vectors is unsupervised (no need for dictionaries).

5) A Study of Consumer Perception towards Online Grocery Shopping : Challenges and Prospects : Various aspects of online shopping in all dimensions are studied in [5]. The study covers the aspects: characteristics, processes, present consumer perception towards online grocery shopping, Indian players in the market, and the different variables around which the study revolves. The paper provides the concept and ideas behind expanding the market of online grocery shopping, the challenges as well as the future business prospects in Tier II. The study highlights essential points such as wide variety of products, offers and discounts, free home delivery, time saving and convenience, cost effective, easy terms and conditions, user friendliness, authenticity and genuineness, easy to order and cash on delivery. The author highlights these factors to be important for strengthening the pillars of online shopping.

6) State of the Art Artificial Neural Network, Deep Learning, and the Future Generation : In [6] the authors highlight the use of neural networks or artificial intelligence or deep learning in its broadest way. It throws light on how the area of deep learning has developed over time and has been successful in areas like pattern recognition. The authors say that machine learning can be utilized as an auxiliary component of applications to enhance or enable new types of computation such as approximate computing or automatic parallelization.

III. FEASIBILITY STUDY AND REQUIREMENT ANALYSIS

A. Feasibility Study

With rapid development in the fields of NLP & NLG, the feat of achieving in the areas that are more complex and challenging is in vogue. Scientists are ready to experiment these challenging areas and are progressively achieving and improving.

Using a Word2Vec algorithm for achieving the task of product categorization had been equally challenging. Of course, there are better Deep Learning models for achieving the given target but enabling a Word2Vec (shallow neural network) to perform in a similar way with decent performance metrics had been an enriching and delighting experience. Building a context aware dictionary, assigning weights to each category and product pair were some innovative ideas included in the approach. This contextual data fed to the Word2vec algorithm made the task of product categorization a success.

B. Requirement Analysis

After extensive analysis of the problems, we realized the requirements for handling product categorization problems. The model needs to cover these points:

➥ The system should be able to understand the data and the association of product categories with each product in the data.

➥ The system should be able to output a list of products given a product category, at a time. It should recall all the relevant data fed to it during the training phase.

C. Dataset Understanding

The dataset is a collection of fashion footwear: ID, title, brand, price, heel type, heel height, toe type, closure type, pattern, shoe height etc. We are essentially interested in footwear types and subtypes. To this, we add a footwear category list that is essentially adopted by almost all footwear selling retailers. Each footwear present in the data will be mapped to each category from the list by means of semantics relation.

D. System Design and Architecture

1) ***Architecture :*** We consider the Word2Vec model as it is advantageous over Machine Learning techniques and comprises of neural embeddings to handle the limitations of machine learning models and bag-of-words model. The essential components of the proposed model are as follows:

Fig. 1 shows system architecture and design of Word2Vec Model. It depicts the units & blocks that make up the model.

2) ***Vocabulary Builder :*** This module is the basic building block of Word2Vec model. Since we are dealing with fashion footwear data, we create a custom vocabulary or dictionary. This comprises of the product and product-category as the key-value pair. Thus, the model is forced to refer this, instead of the in-built vocabulary.

3) ***Context Builder :*** This is the next level to convert words to vector. We initialize weights to each key-value pair for assigning importance to each product and category pair. This aids the algorithm to understand the data and the importance of each vector.

4) ***Neural Network With Two Layers :*** Word2vec is a shallow neural network. It comprises of the following layers:

(i) One input layer which has as many neurons as there are words in the vocabulary for training.

(ii) The second layer is the hidden layer, layer size in terms of neurons is the dimensionality of the resulting word vectors.

(iii) The third and final layer is the output layer which has the same number of neurons as the input layer.

5) ***Modelling :*** The model processes the context aware dictionary data and the list of footwear and footwear categories. It tries to build a semantic relation around the footwear types and the categories, and pulls out relevant footwear data, given a category. Each output comes along with a probability score, which aids in understanding the degree of relevance of a footwear with respect to a category.

6) Implementation on Dataset : All the footwear information was picked from the dataset, which broadly comprised of all boots, sandals, slippers, and sneakers. A category list was chosen, which is widely accepted by almost all footwear retailers, and which had the following: 'Casual', 'Formal', 'Outdoor' & 'Sports'.

A dictionary was built using this information with category and footwear items comprised of the key-value

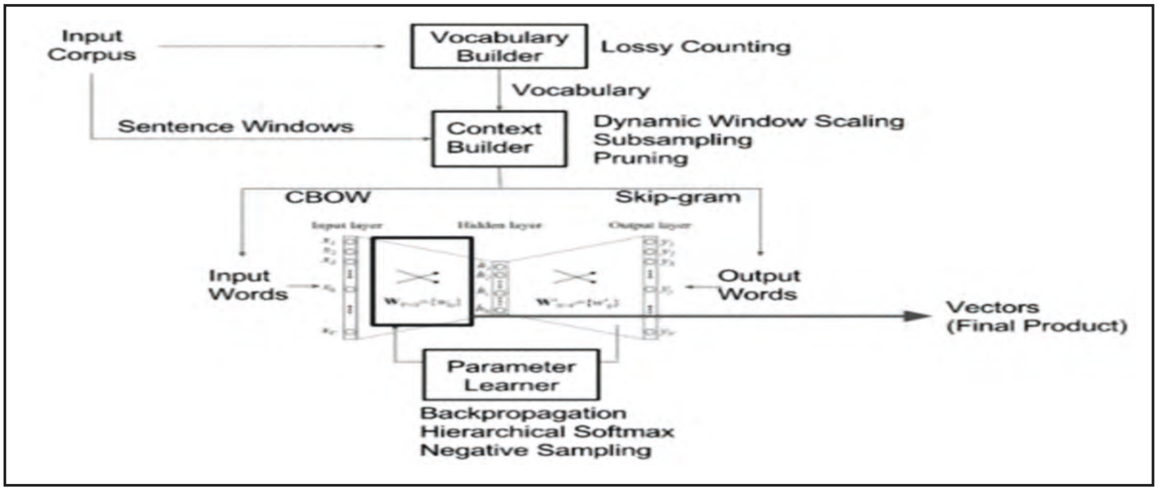


Fig. 1. System Architecture and Design of Word2VecModel

pair. Each key and value was assigned an importance or weight to arrive at an appropriate probability score or semantic score. The dictionary data was ingested by a parametric layer that further enriches each data point. The parameters involved were: minimum count or *min_count* (= 1), context window size or *size* (= 500), worker nodes or *workers* (= 4), skip gram or sg (= 0 implies CBOW & = 1 implies skip gram), hierarchical softmax or *hs* (= 1) & seed (=1).

Then the model was saved and called, given a category, it outputs semantic similar footwear and along with it a similarity score.

Fig. 2. shows sample output for sports category. Given the ***'sports'*** category as input, the model derives the list of sports footwear from the footwear data. Fig. 3 shows sample output for safety category. Given '*safety*' category as input, the model derives the list of safety wear from the footwear data.

IV. LIMITATIONS AND FUTURE ENHANCEMENTS

A. Limitations

Although the Word2vec algorithm is a very efficient solution to most of our problems, there are situations when it might not be a good approach to opt for. It is very easy to develop but equally difficult to debug. It does not handle ambiguities. This means that if a word has multiple meanings, the algorithm takes an average of these in a vector space, leading to undesired consequences.

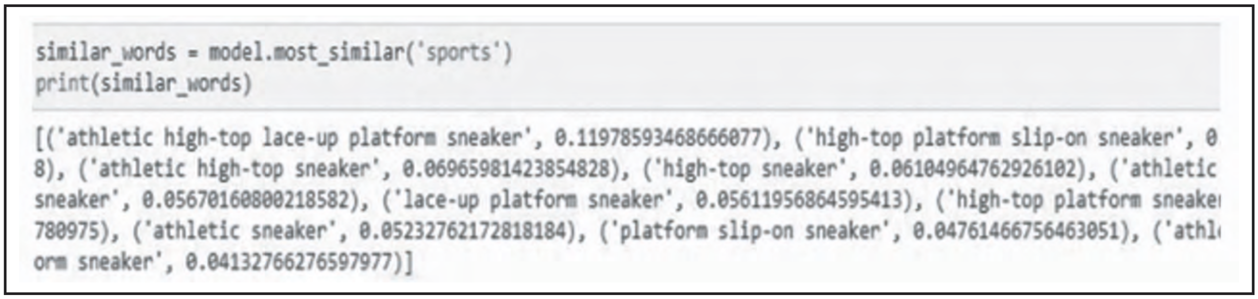
Moreover, it is a shallow neural network (network with few perceptron layers). Hence, a deep learning model could easily outperform its performance at times.

B. Future Enhancements

The task of categorizing products could be further enhanced by the use of deep neural network based models. This could enhance product categorization problems with better performance metrics values. Of course, with more modifications and parameter tunings, even the existing model can be aided for performance improvement.

V. CONCLUSION

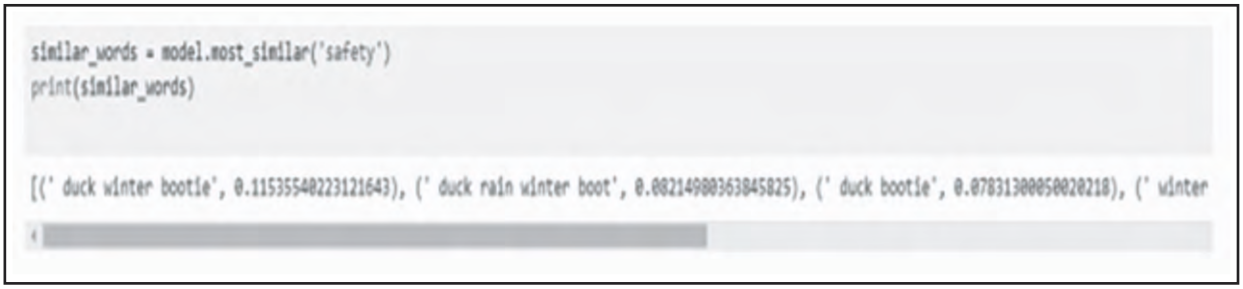
This work aims at categorizing products under a given category and is applicable across all retail products. The model chosen for the task is a Word2vec algorithm which



```
similar_words = model.most_similar('sports')
print(similar_words)
```

```
[('athletic high-top lace-up platform sneaker', 0.11978593468666077), ('high-top platform slip-on sneaker', 0
8), ('athletic high-top sneaker', 0.06965981423854828), ('high-top sneaker', 0.06104964762926102), ('athletic
sneaker', 0.05670160800218582), ('lace-up platform sneaker', 0.05611956864595413), ('high-top platform sneake
780975), ('athletic sneaker', 0.05232762172818184), ('platform slip-on sneaker', 0.04761466756463051), ('athl
orm sneaker', 0.04132766276597977)]
```

Fig. 2. Sample Output for Sports Category



```
similar_words = model.most_similar('safety')
print(similar_words)
```

```
[(' duck winter bootie', 0.11535540223121643), (' duck rain winter boot', 0.08214980363845825), (' duck bootie', 0.07831300050020218), (' winter
```

Fig. 3. Sample Output for Safety Category

is a pre-trained model but for the purpose of product categorization it would need additional training to perform the desired task. The shallow neural network-based model is capable of fetching semantic similar words around a given keyword. It is capable of doing this with the help of a vocabulary list which aids the semantic build.

For the proposed solution, we are concerned about categorization; hence, we build a contextual dictionary and allow the model to understand that instead of an English vocabulary list. Each dictionary data item is assigned with some weights that further guide the algorithm to pull out the most relevant item, given a category. Also, it helps the algorithm to assign a suitable similarity or probability score.

Further, the product categorization performance can be enriched by using Deep Learning based language models.

ACKNOWLEDGMENT

The author Sudesna is immensely grateful to her mentor Bagya Lakshmi for actively mentoring throughout the entire research process and for synergistically helping in producing the desired outcome of the presented work. She is highly indebted to Navneeth Devaraj and Gnanaprakash A. for their immense contribution towards making her understand the data and the relevant details around it. Also, she is immensely grateful to Deepak Kumar Jayaram for assuring all the support and guidance from the domain front seamlessly.

AUTHORS' CONTRIBUTION

Sudesna Baruah worked on creating the whole product categorization of fashion footwear. She read the mentioned references and worked on the data. She built the model, fine tuned, and processed the outputs generated from the model. Bagya Lakshmi V. closely monitored each step in this work and synergistically helped as a mentor, in bringing the desired outcome. Navneeth Devaraj and Gnanaprakash A. assisted in understanding the data and in validating the outputs manually. Deepak Kumar Jayaram mentored Navneeth Devaraj and Gnanaprakash A. and assured all timely support from functional front.

CONFLICT OF INTEREST

The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in the manuscript.

FUNDING ACKNOWLEDGEMENT

The authors received no financial support for the research, authorship, and/or for the publication of the paper.

REFERENCES

[1] H. Wen, Y. Li, N. Naole, E. Chuk, C.-F. A. Deguzman, A. Jain, X. Zhou, R. Mirchandaney, A. Bhardwaj, E. Kobe, and S. Iyer, *"Systems and methods for multi-modal automated categorization,"* [Online]. Available: https://patents.google.com/patent/WO2017165774A1/n

[2] Z. Kozareva, "Everyone likes shopping! Multi-class product categorization for e-commerce," In *Proc. 2015 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technologies,* pp. 1329–1333, 2015. Denver, Colarado, Assoc. Comput. Linguistics. [Online]. Available: https://aclanthology.org/N15-1147.pdf

[3] P. Ristoski, P. Petrovski, P. Mika, and H. Paulheim, "A Machine Learning approach for product matching and categorization," in *Proc. 2015 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technologies,* 2015. [Online]. Available:http://www.semantic-web-J..net/content/machine-learning-approach-product-matching-and-categorization-0

[4] *"Managing product feeds: Classifying items using Word2Vec by Ross Turner,"* [Online]. Available: https://visual-meta.com/2017/03/21/managing-product-feeds-classifying-items-using-word2vec/

[5] S. M. Shukla, "A study of consumer perception towards online grocery shopping: challenges and prospects," *Indian J. Comput. Sci*., vol. 2, no. 4, 2017, doi: 10.17010/ijcs/2017/v2/i4/117852

[6] M. A. Ahmed and S. Trausan-Matu, "State of the art Artificial Neural Network, Deep Learning, and the future generation," *Indian J. Comput. Sci.* vol. 2, no. 6, 2017, doi: 10.17010/ijcs/2017/v2i6/120440

About the Authors

Sudesna Baruah has been working as a Data Scientist at Tata Consultancy Services, Bengaluru and then Chennai for over 5 years. She completed B.Tech. degree in Electronics and Communication Engineering from Manipal Institute of Technology, Majhitar Campus, Sikkim. She also has a Post Graduate Diploma in Data Science from Manipal Academy of Higher Education, Bengaluru Campus, Karnataka. She is a Data Science enthusiast with many research activities and publications in the field of Artificial Intelligence to her credit.

Bagya Lakshmi V. has been working as a Principal Scientist at Tata Consultancy Services, Chennai for over 15 years. She completed B.Tech. degree in Electrical and Electronics Engineering from Government College of Technology, Anna University, Coimbatore. She is also a Postgraduate in Robotics allied with her Under Graduation. She is a pronounced Data Science enthusiast with many research activities and publications in the field of Artificial Intelligence to her credit.

Navneeth Arjun Devaraj has been working as a Functional Consultant (Retail e-commerce) at Tata Consultancy Services, with an overall experience of over 20 years. He has an Engineering degree in Computer Science from MGR Engineering College, Chennai. He is a Business Analytics practitioner with expertise in Business Process Transformation, Process Re-Engineering, and Change Management.

Gnanaprakash Arumugam has been working as a Functional Consultant (Retail e-commerce) at Tata Consultancy Services with an overall experience of over 13 years. He completed MBA from CMS Engineering College, Coimbatore, affiliated to Bharathiar University. He is a Business Analytics practitioner with expertise in Site Merchandiser and Content Management Process.

Deepak Kumar Jayaram has been working as a Senior Functional Consultant (Retail e-commerce) at Tata Consultancy Services with an overall experience of more than 16 years. He has a BBA from Guru Nanak College, Chennai. He is a Business Analytics practitioner with expertise in Business Process Transformation, Process Re-Engineering, and Change Management.