Master Thesis: Imputation of missing Product Information using Deep Learning

A Use Case on Amazon Product Catalogue

Aamna Najmi, 18.01.2019, Kickoff

Advisor: Ahmed Elnaggar

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

Outline

- Introduction
- Motivation
- Objective
- Approach
- Research Questions
- Dataset
- Methodology
- Timeline of the Project

© sebis

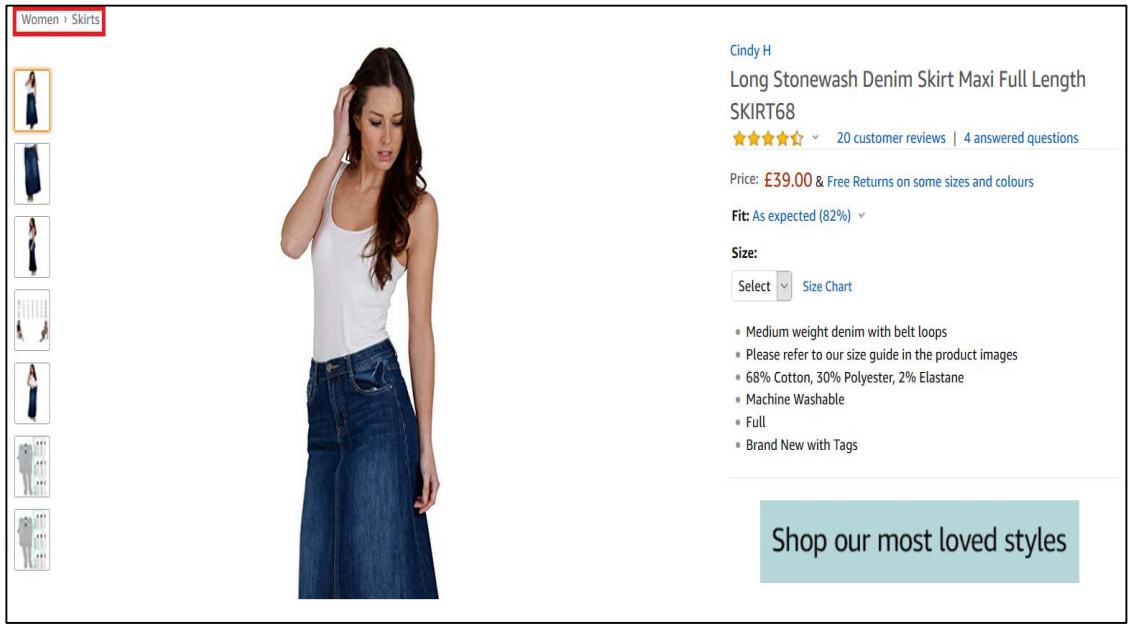
Outline

- **Introduction**
- Motivation
- Objective
- Approach
- Research Questions
- Dataset
- Methodology
- Timeline of the Project

© sebis

Introduction

- Global retail ecommerce sales will reach about $4 trillion in 2020, accounting for 14.6% of total retail spending worldwide [1]
- 20% of purchase failures are potentially a result of missing or unclear product information [2]
- Detailed product information = improved customer experience and company profit



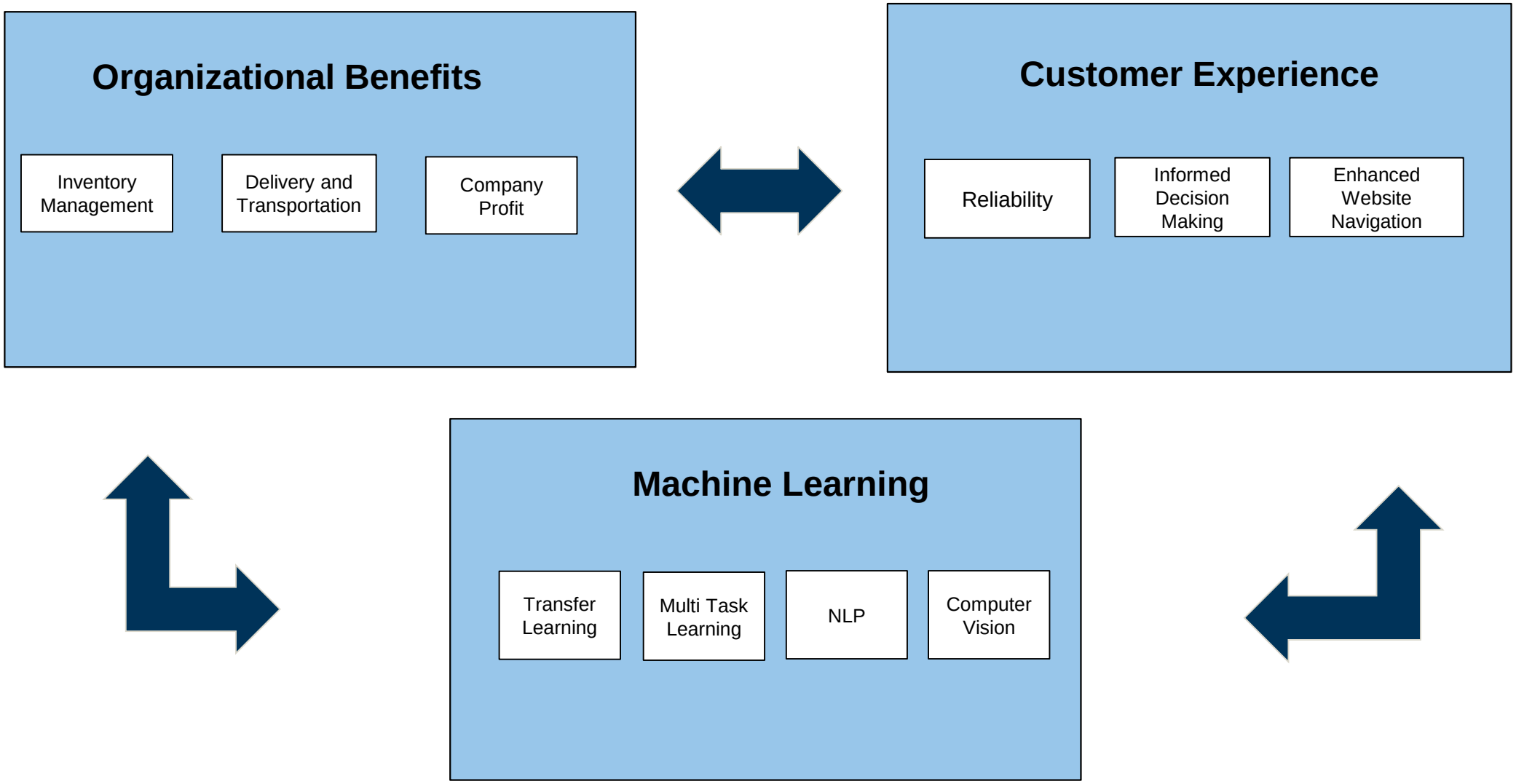
© sebis

Outline

- Introduction
- **Motivation**
- Objective
- Approach
- Research Questions
- Dataset
- Methodology
- Timeline of the Project

© sebis

Motivation



© sebis

Outline

- Introduction
- Motivation
- **Objective**
- Approach
- Research Questions
- Dataset
- Methodology
- Timeline of the Project

© sebis

Objective

Predict missing information (Category, Color and Gender) of Amazon products belonging to the Fashion department in the European market using textual information in five languages, namely English, German, French, Spanish and Italian, and product images as inputs to the model.

© sebis

Outline

- Introduction
- Motivation
- Objective
- **Approach**
- Research Questions
- Dataset
- Methodology
- Timeline of the Project

© sebis

Approach

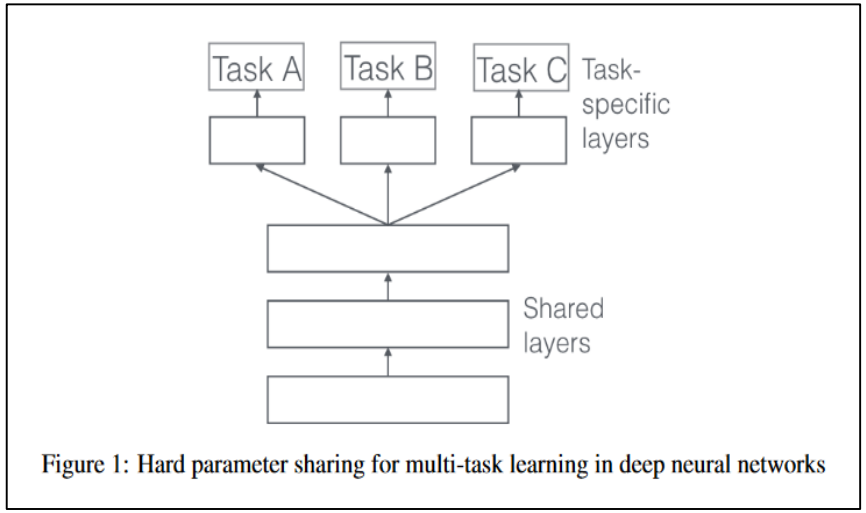
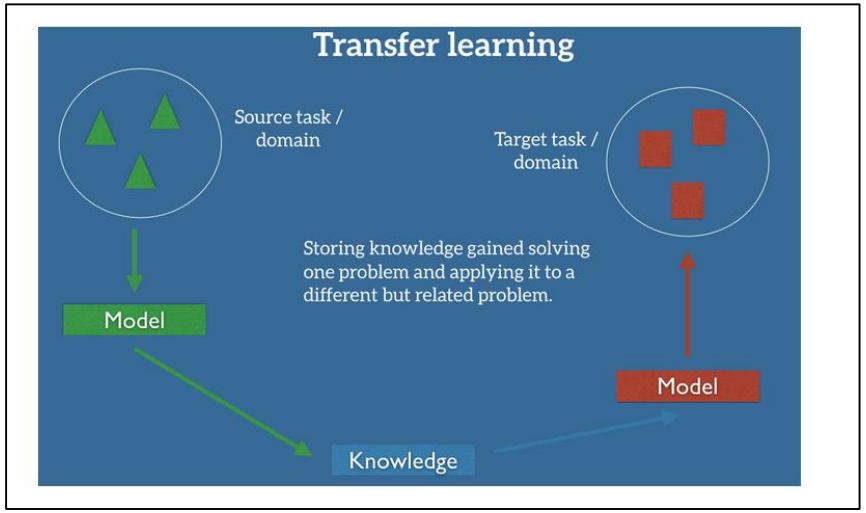


Figure 1: Hard parameter sharing for multi-task learning in deep neural networks



- Multi-task learning : Train one model to perform multiple tasks concurrently [3]
- Transfer Learning : Use pre-trained network to apply its weights and biases to a different domain [4]

© sebis

Outline

- Introduction
- Motivation
- Approach
- **Research Questions**
- Dataset
- Methodology
- Timeline of the Project

© sebis

Research Questions

1. ***Could training on multiple tasks and transfer learning perform better than training on one task independently using the Amazon Product Catalog Dataset?***

2. ***What architecture choices and hyperparameters shall we use in both multi-task and transfer learning to obtain better performance?***

3. ***Can multi-task learning and transfer learning be useful in the ecommerce domain to enhance user-experience?***

© sebis

Outline

- Introduction
- Motivation
- Approach
- Research Questions
- **Dataset**
- Methodology
- Timeline of the Project

© sebis

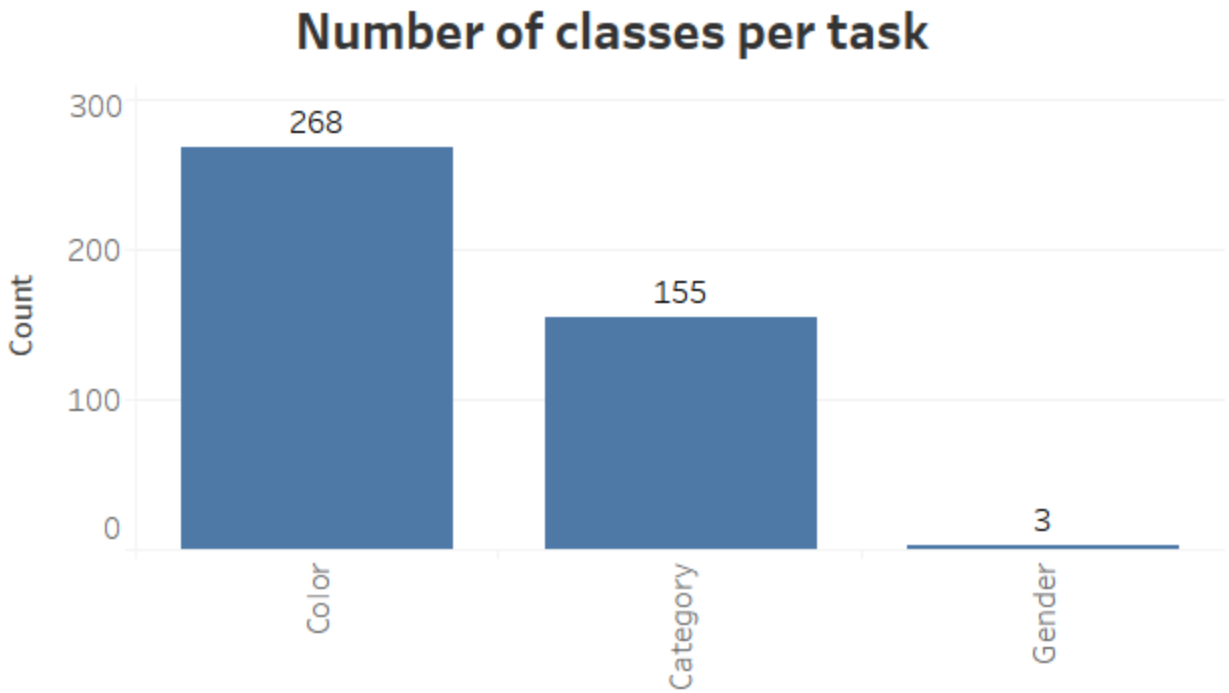
Dataset

- **Source**: Scraped data from five regional Amazon websites (UK, DE, FR, IT,ES)
- **Records**: 200k from each website, 1 million in total
- **Attributes**: Product ID, Product Title, Product Description, Color, Category, Product Summary, Product Specifications, Product Image
- **Sample**:

title	color	category	content	summary	description	specification	product_id	
b'Kono 24 Inch Medium Luggage Super Lightweigh...	Hair-brushed Black	Suitcases & Travel Bags>Suitcases	['Product Dimensions: \n \n 42 x 26 x 65...	✈ Size(including wheels):24"-(H)65cm, (L)42cm, ...	Package include: 1pc 24inch suitcase Product D...	Height:65 centimetres, Length:42 centimetres, ...	B07HGYWR4H	
b'\xe2\x9c\x85 Titan PICKPROOF Combination Pad...	Black Padlock Search Alert (2 Pack)	Accessories>Travel Accessories>Luggage Locks	['Product Dimensions: \n \n 3 x 1.2 x 0....	✅ THE MOST SECURE TRAVEL LUGGAGE LOCKS with IN...	✅ LIFETIME WARRANTY. TOTAL SATISFACTION GUARAN...	Height:7 millimetres, Length:30 millimetres, W...	B078TL4TZM	
b'Aerolite 4-Wheel Spinner Hardshell Cabin Han...	Black	Cooking & Dining>Kitchen Storage & Organisatio...	['Product Dimensions: \n \n 55 x 35.5 x ...	Dimensions: 55x35.5x20cm (All Parts) / 33L Cap...	These Hardshell luggage bags are the perfect t...	Height:20 centimetres, Length:55 centimetres, ...	B00FANS57C	

© sebis

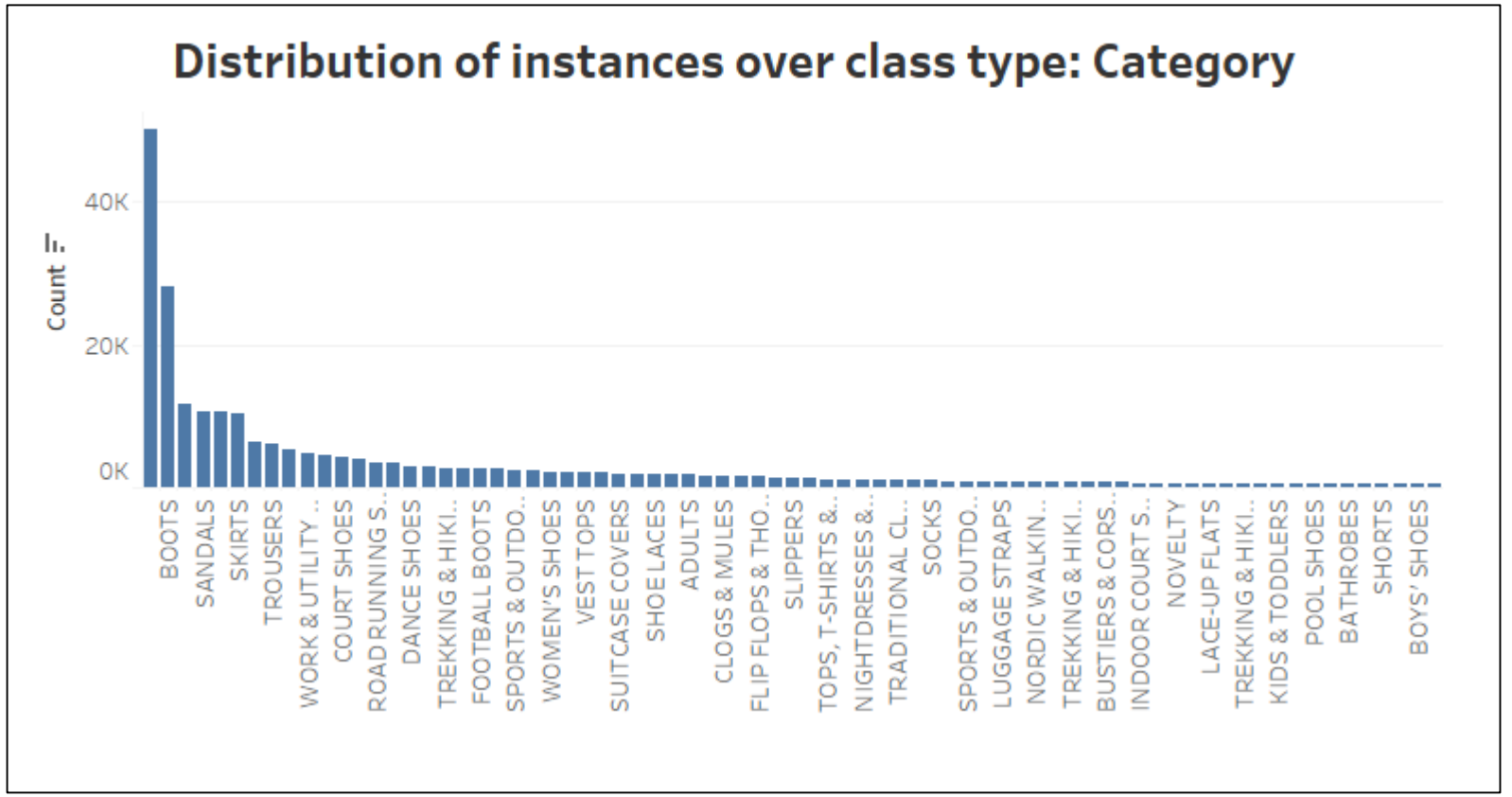
Dataset (contd.)



Note: The above image is representative of the data scraped from www.amazon.co.uk

© sebis

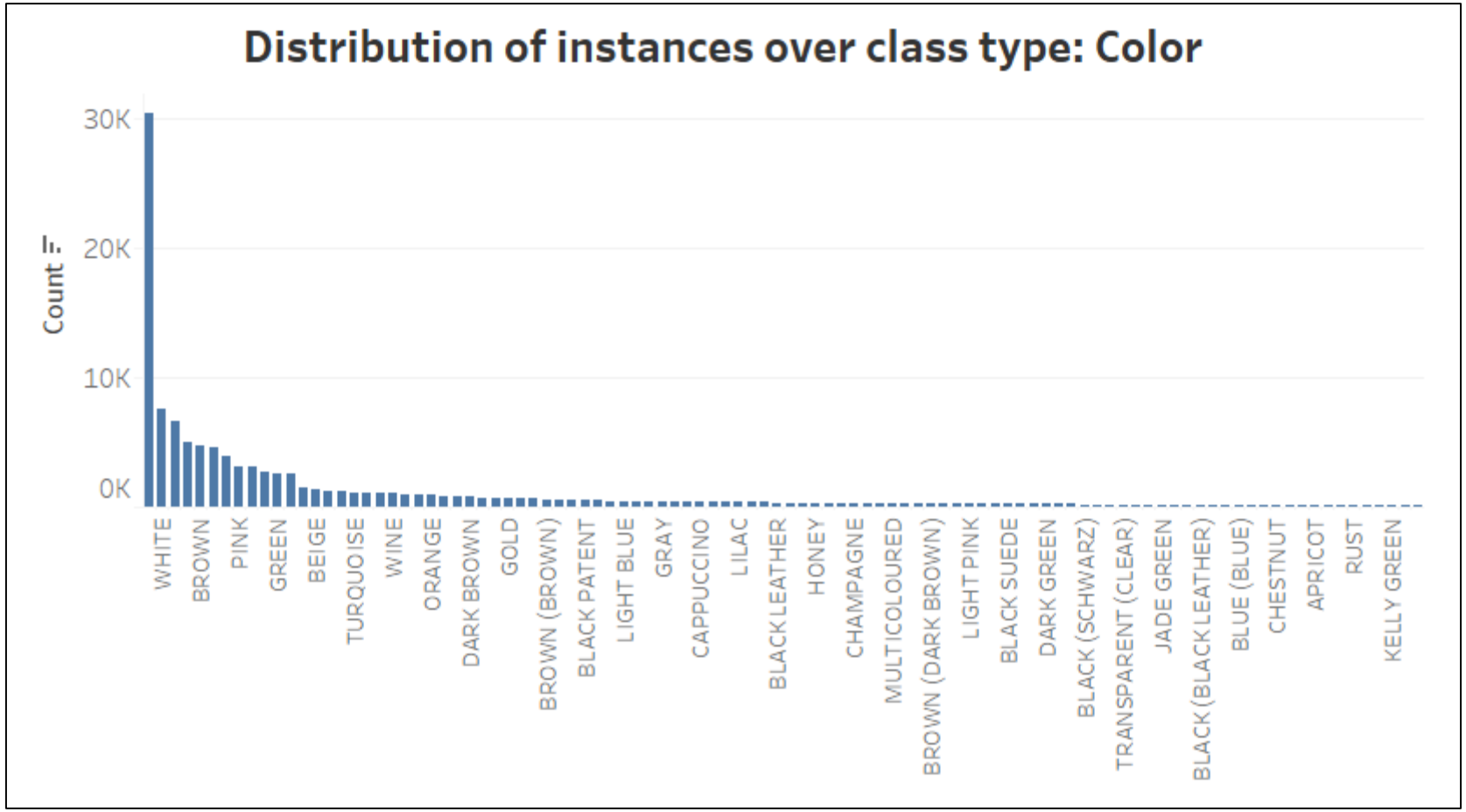
Dataset (contd.)



Note: The above image is representative of the data scraped from www.amazon.co.uk

© sebis

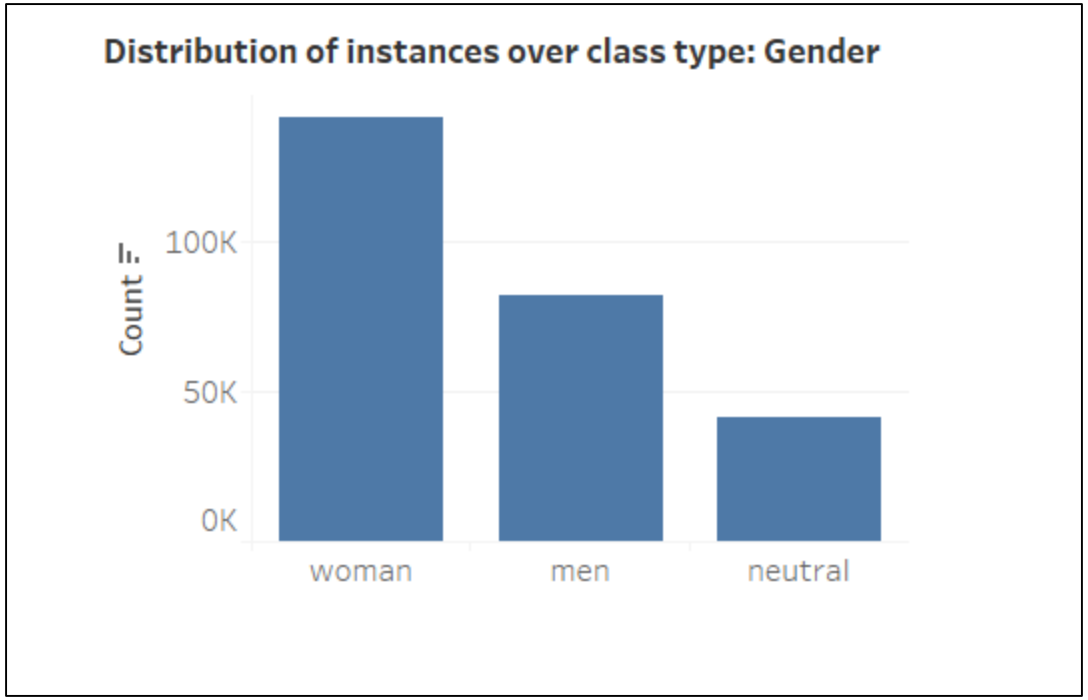
Dataset (contd.)



Note: The above image is representative of the data scraped from www.amazon.co.uk

© sebis

Dataset (contd.)



Note: The above image is representative of the data scraped from www.amazon.co.uk

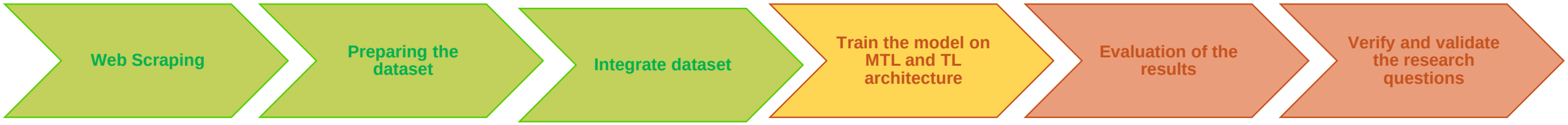
© sebis

Outline

- Introduction
- Motivation
- Approach
- Research Questions
- Dataset
- **Methodology**
- Timeline of the Project

© sebis

Methodology

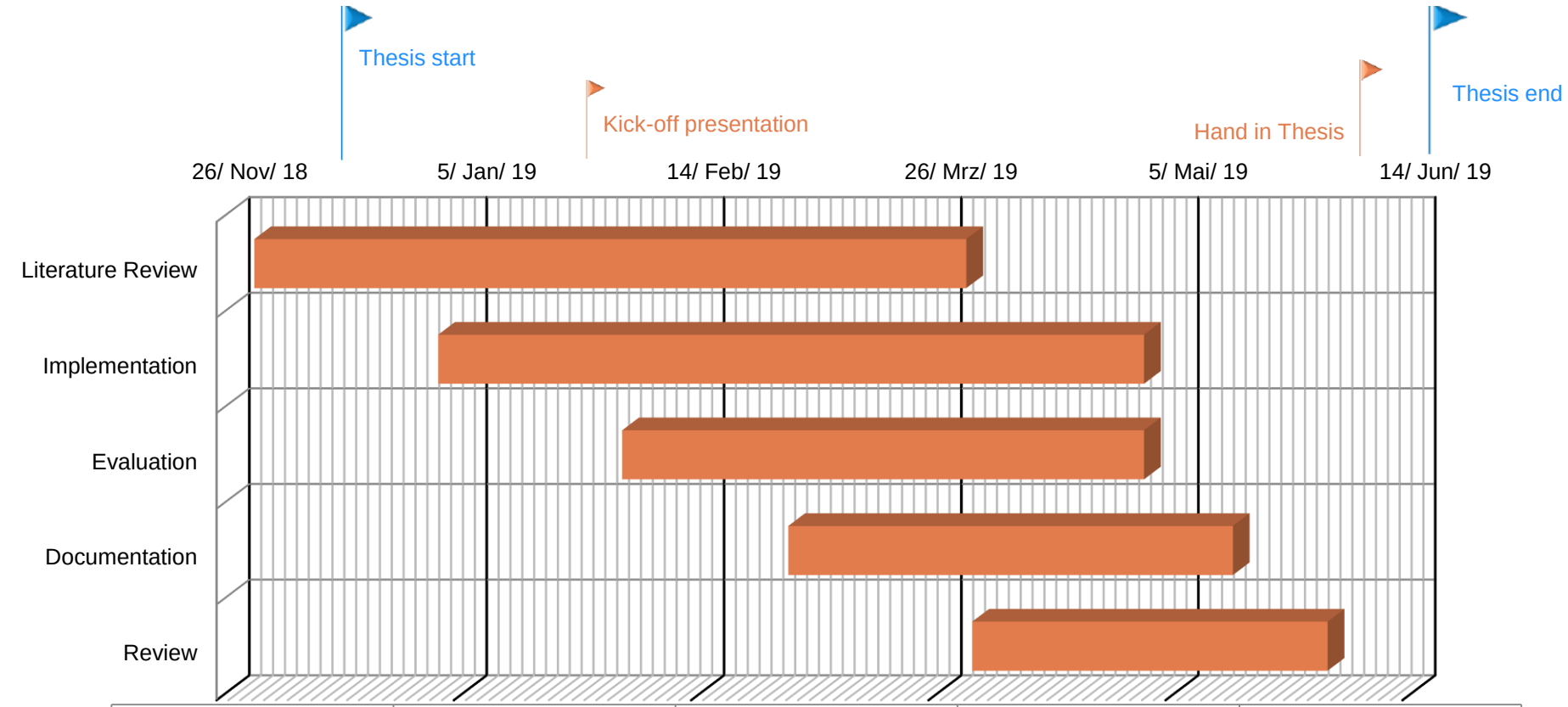


Outline

- Introduction
- Motivation
- Approach
- Research Questions
- Dataset
- Methodology
- **Timeline of the Project**

© sebis

Timeline



	Review	Documentation	Evaluation	Implementation	Literature Review
Start Date	April 1, 2019	March 1, 2019	February 1, 2019	January 1, 2019	December 1, 2018
Duration	60	75	88	119	120
End Date	May 31, 2019	May 15, 2019	April 30, 2019	April 30, 2019	March 31, 2019

© sebis

References

[1] https://www.emarketer.com/Article/Worldwide-Retail-Ecommerce-Sales-Will-Reach-1915-Trillion-This-Year/1014369
[2] https://www.nngroup.com/reports/ecommerce-user-experience/
[3] Sebastian Ruder. An overview of Multi-Task Learning in Deep Neural Networks, 2017
[4] Sebastian Ruder. http://ruder.io/transfer-learning/

Imputation of missing Product Information using Deep Learning

Aamna Najmi

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business Information Systems

Boltzmannstraße 3
85748 Garching bei München

aamna.najmi@mytum.de