**Unstructured Data Analytics to Improve Digital Eligibility**

Suneet Abraham  
Krannert School of Management  
Purdue UniversityWest Lafayette, USA  
abraha46@purdue.edu

Sharan Shirodkar  
Krannert School ofManagement  
Purdue UniversityWest Lafayette, USA  
sshirodk@purdue.eduAkshay Deshmukh  
Krannert School of Management  
Purdue UniversityWest Lafayette, USA  
deshmu25@purdue.edu

Nikhitha Siddi  
Krannert School ofManagement  
Purdue UniversityWest Lafayette, USA  
nsiddi@purdue.eduAnish Jasti  
Krannert School of Management  
Purdue UniversityWest Lafayette, USA  
jastia@purdue.edu

Matthew A. Lanham   
Krannert School ofManagement  
Purdue UniversityWest Lafayette, USA  
lanhamm@purdue.edu

*Abstract— With the businesses currently scaling up at a rapid rate, automation is a key component for sustainability. It ensures a quick turnaround time and fewer errors caused due to human interaction. This leads to an improved customer satisfaction. According to Our retail client, only 33% of their active products are eligible to be purchased. A lack of digital eligibility restricts Our retail client to sell their products in their store. The generation of product description is currently handled by vendors, and multiple products have descriptions which are either missing or not a good fit. This lowers the digital eligibility and in turn the number of products that could be listed on the website. We work on two different objectives using unstructured data, which help improve the digital eligibility of the products. One, we score the existing product descriptions using an algorithm which we developed. This informs us about the product descriptions which need to be updated. Two, we create a model to generate product description based on the product images. This automation will reduce the effort invested by vendors who manually write the product descriptions.*

Keywords—digital eligibility, unstructured data, automation

# Introduction

The advancements in technology over the past few years have led to a drastic change in the customer experience (CX), making it possible for customers to go through the entire purchase process, from finding a product to making a purchase, without ever having to step into a physical store or speak with a live person [1]. The COVID-19 pandemic has only served to further accelerate this digitization of customer interactions [2].

In order to provide a successful online shopping experience, it is essential for websites to have clear, well-written product descriptions that not only increase visibility through optimized use of SEO keywords but also assist customers in making informed purchasing decisions. Retail giant has seen a noticeable increase in the number of customers choosing to purchase products through its website in recent months. However, there is a challenge that Our retail client has faced in ensuring that all of its products are digitally available for purchase.

Our retail client's policy requires vendors to provide updated information about the product, including the product's name, description, features, and at least one relevant photograph, for a product to be sold on its website. Unfortunately, many vendors are failing to adhere to these requirements, resulting in Our retail client missing out on potential sales, as only 33% of its products are currently sold online. The company has been investing in employees who manually correct these descriptions, but this is a time-consuming and costly solution.

To address this issue, we propose the development of an algorithm that scores product descriptions and identifies which ones need to be updated. The algorithm takes into account the length and number of token words in the description, as well as the product title and image, to create a new description that meets Our retail client's requirements.

Diagram

Description automatically generated

The first step in this process involves using Azure Cloud Computer Vision to scrape the text from the product images listed by vendors. This information, along with the product title, is then provided to a language model, such as ChatGPT, to create a product description with a specified word count limit, based on the product's hierarchy. The hierarchy is determined by the category the product belongs to, which can range from bakery goods and fresh produce to apparel and technology.

The rest of the paper is divided into several sections, including Data, Methodology, Model, Results, Conclusion, and References. The Data section serves as a data dictionary and provides detailed information on each variable. The Methodology section describes the experimental design, while the Model section elaborates on the design and evaluation parameters of our algorithm. The Results section compares the performance of our algorithm with a baseline model. Finally, the Conclusion highlights the potential applications of our research and avenues for future work. The References section contains the research we referred to in our study.

It is important to note that while Our retail client serves as a case study in this research, the algorithm we propose has the potential to be applied to other retailers facing similar challenges in ensuring that all of their products are digitally available for purchase. By streamlining the process of updating product descriptions and reducing the need for manual correction, retailers can increase their online sales and provide a better customer experience. Furthermore, by using language models like ChatGPT, retailers can ensure that the product descriptions are well-written, informative, and optimized for search engines, further improving the customer experience.

In conclusion, the digitization of customer interactions has resulted in a major shift in the customer experience, making it increasingly important for retailers to have clear, well-written product descriptions. By developing an algorithm that scores product descriptions and identifies which ones need to be updated, retailers can streamline the process and increase their online sales, ultimately leading to a better customer experience. Our research serves as a starting point for further exploration into this area and highlights the potential.

# DATA

In this study, we used two data tables provided by the client - product description and product image data. The product description table consists of 800,368 unique SKUs present on client’s website. The table consists of product information such as the SKU (stock keeping unit code), UPC (universal product code) type, name of the product, product category and subcategory, product description, product features etc.

The product image table consists of the links of the images of the products listed on client’s website. It can contain multiple images for a given product. It also contains information such as the side from which the image was taken, image approval date, image expiration date, image upload date etc. for 188,163 unique SKUs. Both these data sets can be joined on the SKU information.

Table 1 provides a brief description of the different tables used in this study.

TABLE 1. DATA DESCRIPTION

| Table No. | Table Name | Description |
| --- | --- | --- |
| 1 | Product Description Data | 800,368 SKUs with their UPC, name, category, description, features listed |
| 2 | Product Image data | 188,163 SKUs with the image link, and information about the image |

# METHODOLOGY

The objective of this research is to evaluate the quality of product descriptions and identify those that do not meet the digital eligibility standards. To achieve this goal, we developed a scoring algorithm that considers the length and readability of the product description as well as whether it contains the product name.

We started by preprocessing the data, clipping the outliers at the 95th and 5th percentiles to eliminate extreme values. We then assigned varied weightage to the different metrics based on their relative importance.

To calculate the final score, we normalized the sum of weighted metrics score to a scale of 0 to 1. Products with a score of 0.5 or below were deemed ineligible for digital use and required new product descriptions that meet the digital eligibility standards.

In order to generate new product descriptions, we utilized product images and fed them into Azure Cloud Computer Vision for optical character recognition. The extracted text was then fed to a language model, such as ChatGPT, to generate product descriptions with optimum character length and readability score.

Overall, our methodology provides a comprehensive and effective approach for evaluating and improving the quality of product descriptions to meet the digital eligibility standards.

The methodology used to answer the above questions is extensively described in Fig. 1.

FIGURE 1. METHODOLOGY

A screenshot of a computer

Description automatically generated with medium confidence

## Explanatory Data Analysis (EDA)

The first step towards solving any data problem is visualizing the data and finding insights which might be useful in the modeling process. Some interesting insights that we got from the EDA are:

## Data Preprocessing

## Modeling

## Validation

# MODELS

## Model 1

## Model 2

# RESULTS

# CONCLUSION

# Literature Review

Optical Character Recognition (OCR)

Optical Character Recognition (OCR) is a technology used to convert scanned images or scanned documents into editable and searchable text. It involves analyzing an image or document and recognizing characters, symbols, or words within it, and converting the information into a machine-readable format. OCR enables text data to be extracted from images and documents for purposes such as text search, document indexing, and text analysis. This technology can be leveraged in scraping the text from the product images. The below mentioned methods are different ways to perform the same.

OCR using Artificial Neural Networks

The process of recognizing characters involves the following steps: loading an image with a character from the hard disk and eliminating noise in the preprocessing step, selecting a class of characters from a set of classes, recognizing the character by comparing it to characters in the selected class, training the network to improve recognition efficiency if the result is correct, and correcting the result if it is incorrect by entering the correct character and training the network. The network's training can also be reset if it has been wrongly trained.

Conditional random field (CRF) to predict the best labeling for an image

The process described is a method for labeling an image using a conditional random field (CRF). The CRF consists of nodes representing objects, attributes, and prepositions in the image. The objects and stuff in the image are detected using object and stuff detectors and grouped into object nodes. The appearance of the objects and stuff are then classified using attribute classifiers and represented as modifier nodes. Preposition nodes are created for each pair of object and stuff detections based on their spatial relationship. The label for each node is selected from a domain that is specific to the node type. An energy function is minimized over the labeling to determine the best labeling for the image. The energy function consists of unary potential functions based on image models and pairwise and trinary potential functions based on text models.

OCR using Tesseract

Tesseract is an OCR (Optical Character Recognition) software that analyzes input images and converts them into text. It can handle both black and white text and performs analysis on connected components to form blobs of text lines. The lines are then broken into words based on character spacing, and the software performs two passes of recognition to accurately identify the words. Finally, it resolves fuzzy spaces and checks alternative hypothesis for x-height to locate small and capital text.

Text Generation

These are the various techniques that can be implemented as stand alone models or in combination of two or more to derive a final model that successfully generates product descriptions using data such as product titles and information from product image mining.

Coordinate Encoder

The process described involves using a combination of coordinate encoders (Transformers and Gated-CNNs) to improve the accuracy and diversity of the generated product description. The use of a Transformer encoder is believed to improve the accuracy of the output, while the use of a Gated-CNN encoder helps to enhance the diversity of the description by capturing local correlations. The model also includes Gated Linear Units and residual connections to improve its overall performance.

Seq2seq

The process described involves the use of an LSTM encoder to produce a sequence of hidden states from input tokens. The decoder receives the word embeddings of the previous words and computes the attention distribution using learnable parameters, which is used to produce a weighted sum of the encoder hidden states known as the context vector. The context vector and decoder state are then fed through linear layers to obtain the vocabulary distribution, and the network is trained using the negative log-likelihood of the target word at each time step.

eBert

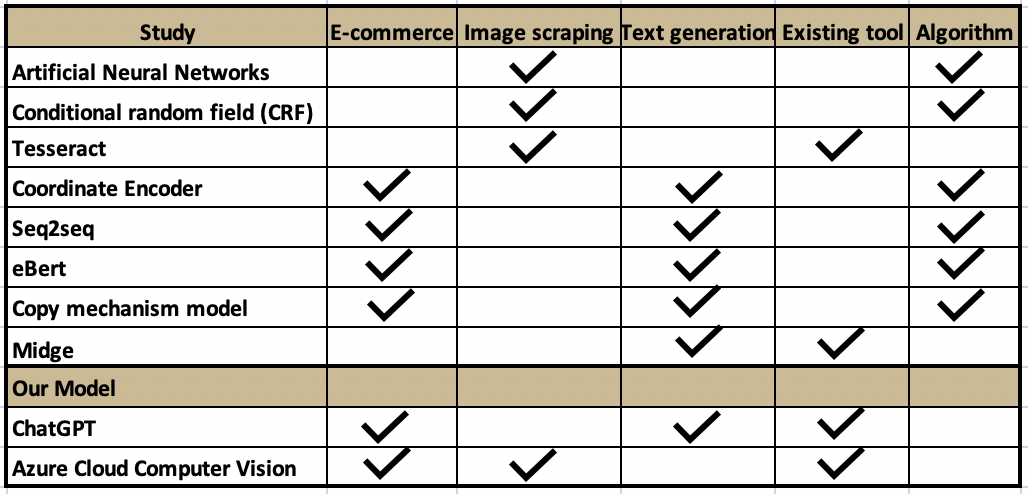
The process described involves using BERT embeddings, which are pretrained on a large corpus of text, as the basis for a language model specifically designed for e-commerce. The embeddings are further pretrained on product descriptions from an e-commerce website to enhance their representation of the language used in e-commerce. This updated BERT model, referred to as eBERT, is then used as the token embeddings for a summarization task. The use of eBERT embeddings is expected to speed up training time and improve the accuracy of the summarization model. The learning rate and warm-up steps for the encoder are decreased to prevent overfitting

Copy mechanism model

The copy mechanism model is a type of pointer-generator network that combines both the seq2seq network and the pointer network. This model has a sequence-to-sequence architecture and uses the attention distribution and context vector to calculate the generation probability of a word, pgen, which is the probability of choosing between generating a word from the vocabulary or copying a word from the source sentence. The model calculates the probability distribution over the extended vocabulary, which includes both the vocabulary and the source sentence words, by combining pgen and the vocabulary distribution. This copy mechanism helps to deal with out-of-vocabulary words and improves accuracy.

Midge: Generating Descriptions of Images

The Midge system uses output from vision detections to generate language. The vision detections provide information about objects, attributes, and their relationships in an image. The language generation process in Midge is based on a lexicalized derivation, where nouns from the object detections form the basis of the generated output. Syntactic trees are used to gather likely adjectives, determiners, prepositions, and verbs to create present-tense declarative sentences.

TABLE 1. SUMMARY OF LITERATURE REVIEW AND STUDY COMPARISON

# Appendix

##### **References**

Kedia, S., Mantha, A., Gupta, S., Guo, S., & Achan, K. (2021). Generating Rich Product Descriptions for Conversational E-commerce Systems. *Companion Proceedings of the Web Conference 2021*. <https://doi.org/10.1145/3442442.3451893>

Zhang, T., Zhang, J., Huo, C., & Ren, W. (2019). Automatic Generation of Pattern-controlled Product Description in E-commerce. *The World Wide Web Conference*. <https://doi.org/10.1145/3308558.3313407>

Mitchell, M., Han, X., & Hayes, J. (2012). Midge: Generating Descriptions of Images. *International Conference on Natural Language Generation*, 131–133. <https://aclanthology.org/W12-1523/>

Kulkarni, G. S., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2011). Baby talk: Understanding and generating simple image descriptions. *Computer Vision and Pattern Recognition*. <https://doi.org/10.1109/cvpr.2011.5995466>

He, Y. (2020). Research on Text Detection and Recognition Based on OCR Recognition Technology. *International Conference on Information Systems*. <https://doi.org/10.1109/iciscae51034.2020.9236870>

Mithe, R., Indalkar, S., & Divekar, N. (2013). Optical Character Recognition. *International Journal of Recent Technology and Engineering*, *Volume-2*(Issue-1). <https://www.ijrte.org/wp-content/uploads/papers/v2i1/A0504032113.pdf>