CA684 Machine Learning Assignment Product-matching using Machine Learning (Zalando challenge)

Arfat Shaikh #21264880
School of Computing, Dublin City University, Ireland
Email: arfat.shaikh3@mail.dcu.ie

***Abstract*—Consumers may now buy things from thousands of e-commerce sites. To assist consumers in selecting the correct store to purchase a product, a basic approach to product matching and classification is required. The completeness of the product requirements and the taxonomies utilized to organize the products, on the other hand, vary amongst e-commerce sites. Different product integration approaches on the Web are required to improve the user experience, for example, by enabling easy comparison of offers from different vendors. This report discusses a machine learning approach to product matching that blends deep learning techniques with conventional artificial neural networks (ANNs). The following report describes a method for product matching and categorization that combines neural language models and deep learning techniques with classic classification approaches. The model below compares products using unstructured descriptions as well as picture matching.**

KEYWORDS: zalando; e-commerce; machine learning; product matching; tensorflow.

I. INTRODUCTION

With the advancement of the Internet and mobile technology, online shopping is becoming increasingly popular. Massive numbers of transactions have been transferred online, resulting in success for online platforms such as Amazon, eBay, Tmall, and JD. Zalando is one of such well-established e-commerce platforms based in Europe. Customers in 23 European markets can purchase fashion and lifestyle products through the company's portal. The growth of the e-commerce business raises issues related to buyer decision-making from various online stores. Hence, consumers get the option of acquiring the product from many online stores, making the product purchasing decision more complex. [4]

Product matching is a critical issue for e-commerce platforms, their consumers, and the e-commerce sector as a whole. A method for matching e-commerce products based on their descriptions is required to improve the consumer purchasing experience.This report presents a machine learning model for matching e-commerce products within a single platform. [3] The goal is to improve product search and comparison.

Zalando aspires for ”trustworthy” prices as a customer offer. That is, the company aims to offer competitive prices in each of its dynamic market situations, removing the need for clients to evaluate costs and driving revenue growth. EAN barcode systems enable each product to be uniquely identified. However, reliable EAN data is not always available. Zalando solves the challenge using multi-modal data, relying on graphics and text.

II. RELATED WORK

Several product-matching approaches identify comparable products that different brands can create, while others find identical products from different online sites. Although there is no mature framework for discovering duplicate products across multiple online stores, several proposed methods in the literature seek to find duplicate texts. Several strategies for text duplicate detection in databases and information networks may be discovered in the literature.

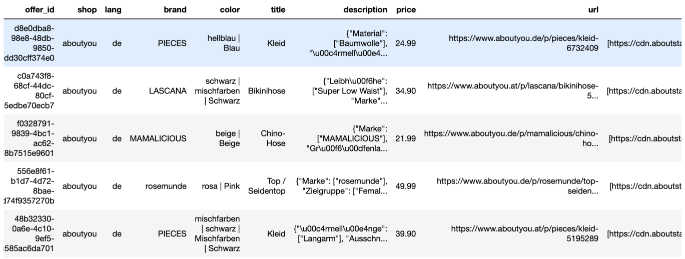
Bakker et al [2] suggest two methods for detecting duplicated products on the internet, focusing on pair-wise matching. Based on the product descriptions, the key step in both systems is to assess if the two products under evaluation are duplicates or not. To locate related product names, the first approach use a model-words algorithm. If the products match, the distance measure is used to determine the degree to which their attributes are comparable. The second method is extended model-words, which employs the modelwords algorithm to assess the similarity of product names and attributes at the same time. Precision, recall, and F1-measure performance metrics were utilized for evaluation, with an average of 63.7 percent, 59.7 percent, and a value of 0.607, respectively. When comparing the extended technique to each methodology independently (once with the model-words algorithm and once with the attribute distance algorithm), the p-value was utilized to compute the significance level. Although the extended model-words technique performed well, it can be enhanced by applying optimization and feature extraction procedures to the descriptions prior to the process of locating duplicate products. [1]

Ristoski et al. [5] used ANNs for product matching and classification. The attributes used in their work are the short name and description of the product. The following algorithms are used to extract the product's features: a. Dictionary-Based: a dictionary of the product qualities and values described in structured product descriptions. b. Conditional Random Field (CRF): a CRF model with discrete features. c. CRF with Text Embeddings: an enhanced CRF model that incorporates text embedding features and is used to handle different variants of a term, such as synonyms, found in product descriptions. d. Image Feature Extraction Model: In addition to textual features, the authors of [5] developed an image embedding model with convolutional neural networks (CNNs). Random forest, support vector machines (SVM), Nave Bayes, and logistic regression models were used to assess the performance of matching products for the four models mentioned above. An examination of electronic datasets yielded encouraging results. However, only if the product name exists in the text can the approach provided by [5] combine unstructured product descriptions.

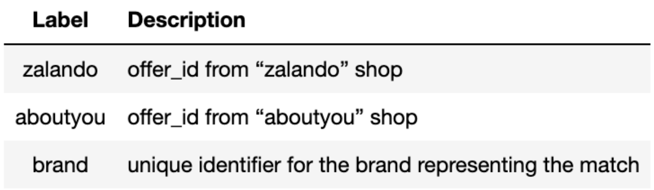
III. DATASET AND EXPLORATORY ANALYSIS

A. DATASET

The dataset contains two files offers- test and offers-training containing product offerings for training and testing, with the following fields: There is one more file named matches-training containing the matches between those offers that use the same offer id to describe the same products which is only available for training offers.



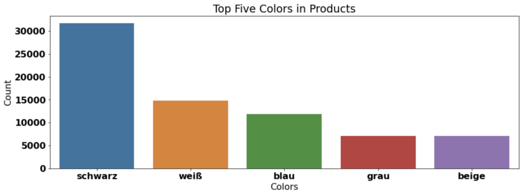
offer_id	shop	lang	brand	color	title	description	price	url	
d8e0dba8-98e8-48db-9850-dd30cff374e0	aboutyou	de	PIECES	hellblau | Blau	Kleid	{"Material": ["Baumwolle"], "\u00c4rmell\u00e4...	24.99	https://www.aboutyou.de/p/pieces/kleid-6732409	[https://cdn.aboutsta
c0a743f8-68cf-44dc-80cf-5edbe70ecb7	aboutyou	de	LASCANA	schwarz | mischfarben | Schwarz	Bikinihose	{"Leibh\u00f6he": ["Super Low Waist"], "Marke"...	34.90	https://www.aboutyou.at/p/lascana/bikinihose-5...	[https://cdn.aboutsta
f0328791-9839-4bc1-ac62-8b7515e9601	aboutyou	de	MAMALICIOUS	beige | Beige	Chino-Hose	{"Marke": ["MAMALICIOUS"], "Gr\u00f6\u00dfenla...	21.99	https://www.aboutyou.de/p/mamalicious/chino-ho...	[https://cdn.aboutsta
556e8f61-b1d7-4d72-8bae-174f9357270b	aboutyou	de	rosemunde	rosa | Pink	Top / Seidentop	{"Marke": ["rosemunde"], "Zielgruppe": ["Femal...	49.99	https://www.aboutyou.de/p/rosemunde/top-seiden...	[https://cdn.aboutst
48b32330-0a6e-4c10-9ef5-585ac6da701	aboutyou	de	PIECES	mischfarben | schwarz | Mischfarben | Schwarz	Kleid	{"\u00c4rmell\u00e4nge": ["Langarm"], "Ausschn...	39.90	https://www.aboutyou.at/p/pieces/kleid-5195289	[https://cdn.aboutsta



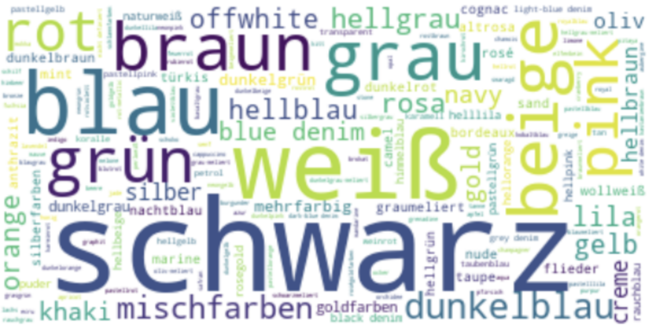
Label	Description
zalando	offer_id from “zalando” shop
aboutyou	offer_id from “aboutyou” shop
brand	unique identifier for the brand representing the match

B. EXPLORATORY DATA ANALYSIS

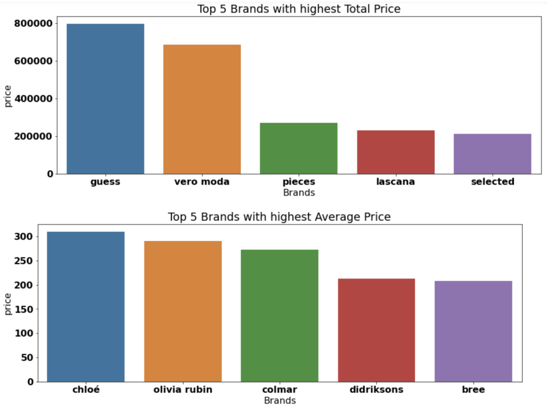
I used Python to explore the dataset and then visualized the data into some interesting graphs and wordclouds using matplotlib.



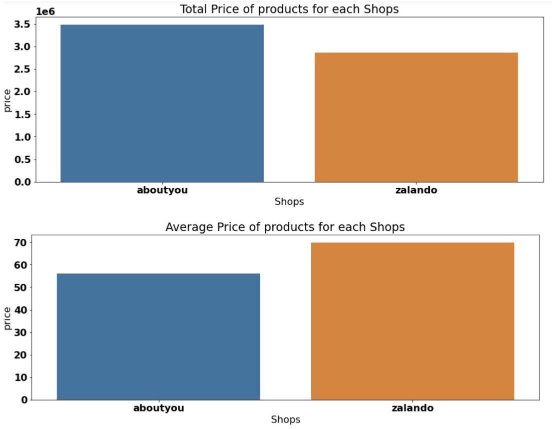
The above graph shows the top 5 Product Colors consumers buy. Shwarz (black) color is the most popular color among the consumers followed by White, Blue, Grey and Beige.



The word cloud above represents the most used titles on the marketplace for the products.



The above graphs show us the comparison between top 5 brands on Zalando and Aboutyou. It is seen that guess is the top brand on Zalando whereas, the top spot in Aboutyou is taken by chloe.



The above graph shows the difference between the total prices of the products and average prices of the products on Zalando VS Aboutyou.

IV. RESULTS

Figure below shows example of similar products from zalando and aboutyou.

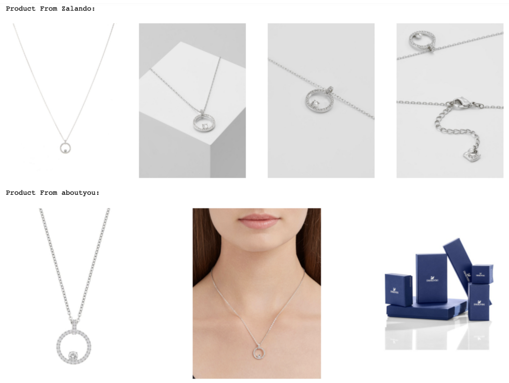
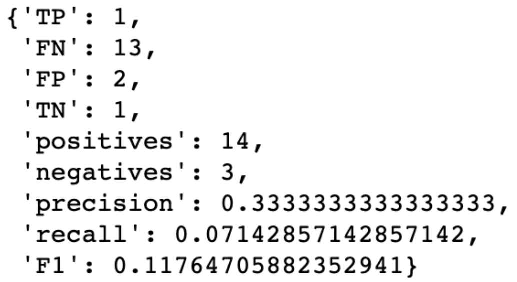


Figure below displays the results of the similarity measure examination. According to the similarity data, the model gives 11 percent average F1-measure.



```
{'TP': 1,
 'FN': 13,
 'FP': 2,
 'TN': 1,
 'positives': 14,
 'negatives': 3,
 'precision': 0.3333333333333333,
 'recall': 0.07142857142857142,
 'F1': 0.11764705882352941}
```

V. CONCLUSION

This study demonstrates an e-commerce product-matching strategy that provides a means for matching products from various online firms. This was accomplished using deep learning techniques such as nlp and TensorFlow. However, with a low average F1 score of 11 percent, the existing model still has a lot of opportunity for development.

REFERENCES

[1] Aisha Alabdullatif and Monira Aloud. Araprodmatch: A machine learning approach for product matching in e-commerce. *International Journal of Computer Science & Network Security*, 21(4):214–222, 2021.

[2] Marnix de Bakker, Damir Vandic, Flavius Frasincar, and Uzay Kaymak. Model words-driven approaches for duplicate detection on the web. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 717–723, 2013.

[3] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 4654–4662, 2019.

[4] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. Text matching as image recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[5] Petar Ristoski, Petar Petrovski, Peter Mika, and Heiko Paulheim. A machine learning approach for product matching and categorization. *Semantic web*, 9(5):707–728, 2018.