Product Matching in Online Retail using Natural Language Processing, Doc2Vec and Cosine Similarity

Teenu Prathyush
Master's Scholar
School of Computing
Dublin City University, Ireland
teenu.prathyush2@mail.dcu.ie

ABSTRACT:

The e-commerce industry has grown rapidly in the past decade. One of the major reasons for this growth is the significant amount of research that has been conducted on product matching using Machine Learning. Online retailers are interested in efficiently selling their products at an optimal price and comparing their products to the exact same product from another retailer helps them to accomplish this task. A variety of methods have been proposed, however, a traditional approach to finding a solution for this problem is called as Multimodal Learning, which focuses on implementing joint representations of different modalities such as those from text and images.

In this research paper, it is proposed to use a model that utilizes natural language processing techniques. An unsupervised learning algorithm called as Doc2Vec is used to create numeric representation of documents, tokenization is used for splitting the words into sentence and cosine similarity is used as the metric to measure the similarity between sentences.

Keywords: E-commerce, Product Matching, Multimodal Learning, Doc2Vec, Tokenization, Cosine Similarity.

1. INTRODUCTION

Zalando is a leading European online platform for fashion and lifestyle based in Berlin, Germany [1]. It works with nearly six thousand international brands and offers customers a wide range of clothing and accessories [8]. In the present day, it has more than 48 million active customers and it is vital for Zalando to recommend matching products to its customers to increase the sales [9]. One of the important propositions from Zalando is that they want to offer competitive prices in every one of its dynamic market environments so as to relieve its customers from having to compare prices, this in turn increases its revenue growth. To accomplish this task for its wide range of individual products, Zalando needs to identify exact product matches from its European competitors like About You.

Product matching is a complex task for retail companies as two different images or two different product descriptions of similar products from different retailers could represent the same product or it could represent entirely different products. A similar case scenario exists in stores like Walmart and Amazon where they allow different sellers to offer the same product on their online platforms [10]. However, product matching becomes a complex task as similar products needs to be grouped together, even when the titles, descriptions and images etc. of the products vary. Amazon uses a matching function for multimodal learning that combines different types of modalities or different types of information for improving the performance [10]. In this research study, product matching is conducted on the dataset provided by Zalando by utilizing Natural Language Processing techniques. The model is implemented on Jupyter Notebook, running on an Intel i7 processor laptop having configuration of 16GB internal memory and 4GB of graphics memory.

In the next section, I am going to introduce some of the related work in the field of product matching using text. In section III, I am going to introduce the dataset provided by Zalando and discuss the exploratory data analysis conducted on this dataset. In section IV, I am going to discuss about the methodology of the whole process. In section V, I will discuss the experiments performed in this research study. In the final section, I will conclude this research study and propose future scope for this research.

2. RELATED WORK

The product matching problem could be viewed as a text matching problem. A significant amount of research has been conducted to map full sentences into k-dimensional vectors for the purpose of text matching.

In [2] the authors demonstrated the use of the Word2Vec bag-of-words model that utilizes a duplicate detection algorithm with semantic embedding for short text. They use hamming distance for calculating the distance between vectors and the results obtained show that they achieve higher accuracy and recall rate when compared to the traditional Word2Vec model and TF-IDF method for text matching.

In [3] the authors proposed a multimodal learning approach that utilizes NFNet, Swin_Transformer, and EfficientNet to get image embeddings and Distil-BERT, AL-BERT, Multilingual-BERT and TF-IDF to get text embeddings. They chose the K-Nearest Neighbors algorithm for classification and Cosine Similarity as the similarity metric. The authors then compare the results of the seven models and use different training methods for improving the efficiency of the models.

In [4] the authors use a dynamic approach for word embedding. They utilize DRMM and K-NRM models with an autoregressive pre-trained language model based on Transformer-XL. The models are implemented on two Text Retrieval Conference Collections and the results obtained from the textual embeddings were an improvement on the traditional Word2Vec embedding technique.

In [5] the authors proposed to use a Fuzzy matching algorithm for approximate string matching. They utilize strings from text to identify similarities using matrices. The authors explain the challenges of using different string matching algorithms and provide a comparison of existing string matching approaches and state that fuzzy matching is better suited for information retrieval if the right similarity metrics are used.

In [6] the authors use the Fuzzy String matching on nine different tools and compare the performance of Fuzzy String matching algorithm to the traditional String matching algorithm. The results obtained show that Fuzzy matching performs better in accurately matching tools when compared to traditional String matching.

3. DATASET AND EXPLORATORY DATA ANALYSIS

DATASET: The dataset was provided by Zalando for the purpose of this research. It consists of two files – training data and testing data, each containing offers of products in parquet file format. An offer is a description of a product by either Zalando or its market competitor - About You. There is a total of 102,884 product offers in the training set and a total of 106,741 product offers in the test set. Both the training set and test set consists of the following fields –

- Offer id - a unique identifier for an offer of a product.
- Shop – contains either Zalando or About You.
- Lang – "de" (German).
- Brand – contains different brands e.g. Guess, Pieces etc.
- Color – contains the color of the product e.g. Gold, Pink, etc.
- Title – contains the title of the product e.g. Polo-Shirt, Sneaker, Pullover etc.
- Description – contains product description that may include cleaning instructions, material composition, etc.
- Price – contains the price of the product in Euro currency.
- URL – contains the URL of the product description page.
- IMAGE_URLs – contains a list of product images that may include close up photos, stock photos, stock photos with model etc.

Another parquet file containing the matches of offers that describe the same products using the "offer id" attribute is also provided. It contains the following fields – zalando (offer id from Zalando shop), aboutyou (offer id from About You shop), and brand (identifier for the brand representing the matched products).

EXPLORATORY DATA ANALYSIS:

Upon exploring the training data it is observed that the number of products from About You (61980) is significantly more than that of Zalando (40904). Figure 1 shows the plot of the total number of 'Pullover' products on offer by Zalando and About You. Figure 2 shows the plot of the total number of 'Pullover' products on offer by Zalando and About You. It is observed that in

both cases the number of products on offer by About You is slightly more than that of Zalando.

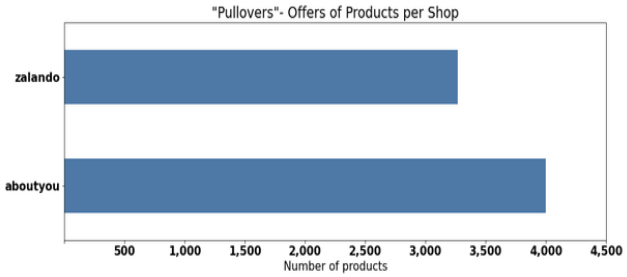

Figure 1: Number of "Pullovers" on offer by Zalando & About You

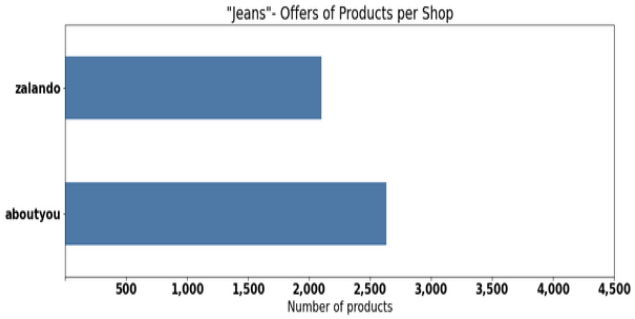

Figure 2: Number of "Jeans" on offer by Zalando & About You

Upon further exploring the data it is observed that there are 145 unique brands in the training set and there are 164 unique brands in the test set. The top 10 brands in terms of the most number of products on offer in the training set is shown in Figure 3.

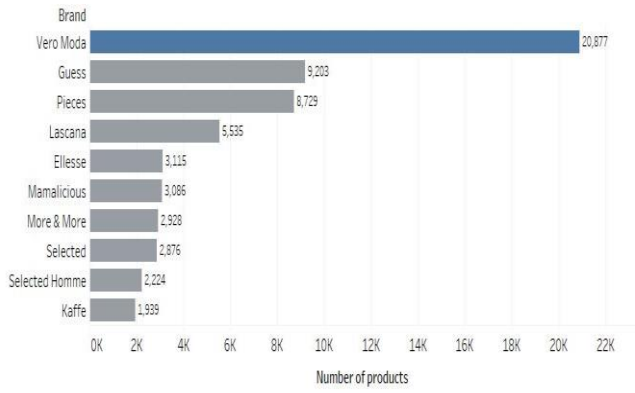

Figure 3: Top 10 brands with most number of products in Training set

Similarly, the top 10 brands in terms of the most number of products on offer in the test set is shown in Figure 4. It is observed that some brands are more popular than other ones in both the training set as well as the test set.

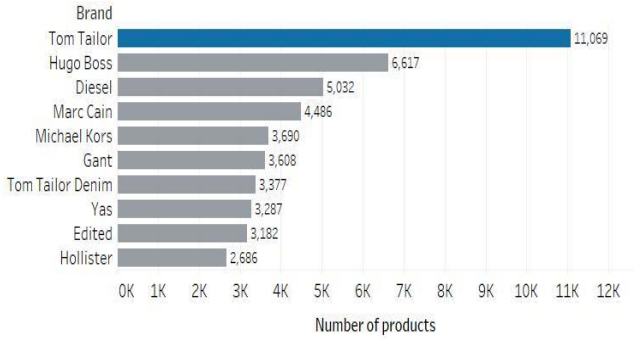

Figure 4: Top 10 brands with most number of products in Test set

Figure 5 shows the frequency of products according to prices. It is observed that the distribution is right-skewed with majority of the products priced in the range of €5 to €150.

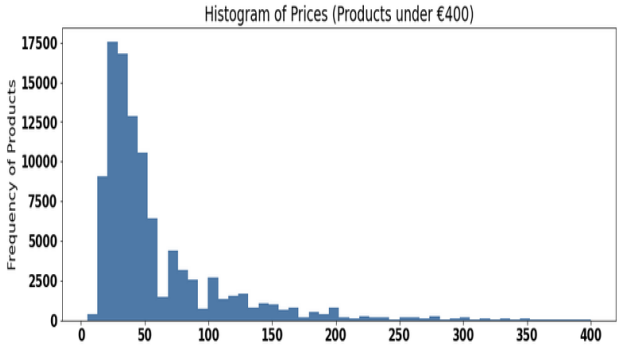

Figure 5: Histogram of prices of products under €400

Using the data from the training set, the top 5 most expensive products along with their brand name from the About You shop is displayed in figure 6. Similarly, the top 5 most expensive products from Zalando is displayed in figure 7. It is observed that the most expensive product offer is from Zalando. However, About You has several products that are more expensive than some of the top Zalando product offers.

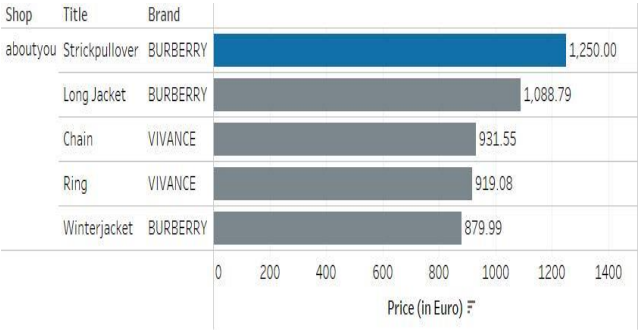

Figure 6: Top 5 most expensive product offers from About You

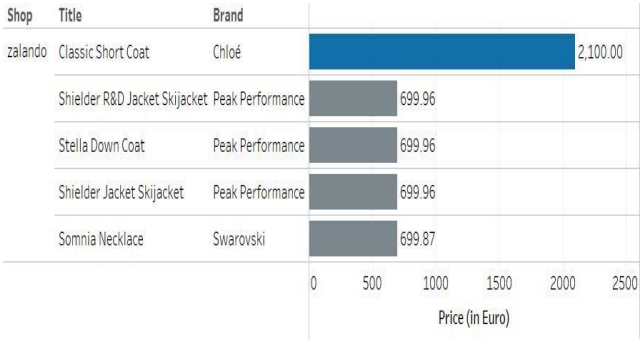

Figure 7: Top 5 most expensive product offers from Zalando

Finally, the 'matches_training' file was explored which contains the offer ids of matching products in the training set. It is observed that the number of ground-truth matches in the training set is 15,170. The goal is to predict all the products that are matching in the test set.

4. METHODOLOGY

CRISP-DM data mining methodology was adopted for performing this research study. The first two phases of the CRISP-DM methodology i.e., Business Understanding and Data Understanding, were discussed in the previous sections. In this section, the next two phases of the CRISP-DM methodology are discussed, which are Data Preparation and Modelling. Evaluation and Deployment will be discussed in the next section.

The dataset provided by Zalando does not contain missing values. However, there were certain columns in the dataset which required further processing, prior to modelling the data. In this research study, three text columns i.e., Brand, Title and Color, were utilized for modelling.

The first step involved translating these columns from German to English. In the second step, the translated columns were concatenated into a single column. In the third step, the sentences from the concatenated column were processed to remove all special characters. Finally, the sentences were converted to lowercase. The sentences does not contain any stop words, therefore, the step of removing stop words from sentences was avoided.

The next step in processing involved tokenization of sentences into words. From the Gensim library, TaggedDocument and Doc2Vec modules were imported. A tagged sentence corpus is created from the tokenized sentences. A tagged corpus contains sentences as tagged documents which has a list of words and a tag associated with it. Finally, the tagged sentences were trained using the Doc2Vec model. The Doc2Vec model converts the sentences into fixed dimension vectors and the trained vectors are then used to find the similarity between words and phrases of test data by calculating the distance using the Cosine Similarity metric. The Cosine Similarity measure the cosine of the angle between two or more vectors and it returns a value that is bounded in the range of 0 and 1. If the Cosine Similarity measure is closer to 1, that means the products are more similar and if it is closer to 0 then the products are dissimilar.

5. EVALUATION

The model is implemented on Jupyter Notebook. The training time is significantly less when compared to other models such as Fuzzy logic for text matching and TF-IDF (Term Frequency Inverse Document Frequency) for text matching. The model was trained using optimal hyper-parameters with vector_size = 20, window = 2, and min_count = 1 over 100 epochs.

As previously stated, the Cosine Similarity metric used for calculating the similarity between two or more vectors by determining whether the vectors are pointing in the same direction. The formula for measuring Cosine Similarity is given by:

$$\text{Similarity} = \cos(\theta) = \frac{A.B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Figure 8: Formula for calculating Cosine Similarity from [7].

To evaluate the efficiency of the model, metrics such as Precision, Recall and F1-Score are to be used. The main goal of the model was to predict the similarity between products from Zalando and Aboutyou and to maximize the F1-Score.

Precision is the ratio between True Positives and all of the positives in the dataset. In this scenario, it is the total number of products that were correctly matched out of all the products that were matched in the dataset. The formula for Precision is given by:

$$Precision = \frac{TP}{TP + FP}$$

Figure 9: Formula for calculating Precision Score from [7].

Recall is a measure of the model correctly predicting the True Positives. In this scenario, it is the total number of products correctly matched. The formula for Recall is given by: Recall = (True Positives) / (True Positives + False Negatives).

$$Recall = \frac{TP}{TP + FN}$$

Figure 10: Formula for calculating Recall Score from [7].

It is difficult to balance Precision and Recall score, therefore, F1-Score is calculated as it is the harmonic mean of both Precision and Recall. In this scenario of product matching, F1-Score is the best measure for evaluating the model. The formula for F1-Score is given by: F1-Score = 2 * (Precision * Recall) / (Precision + Recall).

$$F1 - Score = 2\left[\frac{Precision * Recall}{Precision + Recall}\right]$$

Figure 11: Formula for calculating F1-Score from [7].

Upon evaluating the model on the training set, it was observed that the results of product matching was satisfactory. However, the model could be improved to obtain a better F1-Score in the future.

6. CONCLUSION

In the past, the field of product matching has been researched extensively using a variety of algorithms and machine learning models. In this research paper, Natural Language Processing techniques were used to get word embeddings from sentences of mostly three columns – title, brand and color. The vectorised text was then trained using the Doc2Vec model with the goal of maximizing the overall F1-Score. The Cosine Similarity metric was used to calculate the distance between products. The model was then evaluated using metrics such as Precision, Recall, and F1-Score.

To conclude, product matching using Natural Language Processing Techniques has been successfully conducted in this research study. There is a lot of scope for product matching and in the future, a multimodal learning approach could be implemented where the model takes both text and images as input, transforms it into a large combined vector and calculates the distance between products. Utilizing this approach will help to achieve an overall F1-Score that could be better than the existing model.

7. REFERENCES

[1] A. Vera-Baquero, O. Phelan, P. Slowinski, J. Hannon. 2021. Open Source Software as the Main Driver for Evolving Software Systems toward a Distributed and Performant E-Commerce Platform: A Zalando Fashion Store Case Study, *IT Professional (Volume: 23, Issue: 1, Jan.-Feb. 1 2021)*, pp. 34-41, DOI: 10.1109/MITP.2020. 2994993.

[2] J. Gao; Y. He, X. Zhang, Y. Xia. 2017. Duplicate short text detection based on Word2vec, *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pp. 33-37, DOI: 10.1109/ICSESS.2017.8342858.

[3] Yaxuan Fang, Junhan Wang, Lei Jia, Fung Wai Kin. 2021. Shopee Price Match Guarantee Algorithm based on multimodal learning, *2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE)*, pp. 84-87, DOI: 10.1109/CSAIEE54046.2021.9543217.

[4] H. Yu, X. Chen, Y. Zhou. 2020. Utilizing Contextualized Word Embeddings for Text Matching. *2020 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*, pp. 54-59, DOI: 10.1109/ICWAPR51924.2020.9494608.

[5] K. P. Kalyanathaya, Dr. D. Akila, Dr. G. Suseendren. 2019. A Fuzzy Approach to Approximate String Matching for Text Retrieval in NLP. *Journal of Computational Information Systems 15: 3 (2019)*, pp. 26-32.

[6] Wen-Yen Wu. 2016. A Method for Fuzzy String Matching, *2016 International Computer Symposium (ICS)*, pp. 380-383, DOI: 10.1109/ICS.2016.0083.

[7] P.P Gokul, B.K Akhil, K. Kumar, and M. Shiva. 2017. Sentence similarity detection in Malayalam language using cosine similarity, *in 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pp. 36-44, DOI: 10.1109/RTEICT.2017.8256590.

[8] Zalando. (2022, March). Our Business Fields. https://corporate.zalando.com/en/company/our-business-fields.

[9] Pleuni. (2022, March 1). Zalando revenue €10.4 billion in 2021. https://ecommercenews.eu/zalando-revenue-e10-4-billion-in-2021/.

[10] Moreno Roxana and Richard Mayer. 2007. Interactive multimodal learning environments, *Educational psychology review 19.3*, pp. 309-326.