Product Matching using TF-IDF Word Vectors of Text Features

ROMIL SAKARIYA, Dublin City University, Ireland

Product matching challenges have been a pertinent challenge for all machine learning aficionados owing to its vast implications in terms of saving effort and time. Good product matches obtained through machine learning algorithms can have significant impacts throughout many industries. One such industry where product matching finds a direct application is the fashion industry where millions of products are sold to customers that are manufactured by thousands of organizations. These manufacturing organizations or brands may choose to list their products on various e-commerce platforms that interact with the end consumer over the internet. One such platform specializing in business-to-consumer model is Zalando that focuses on sales concerning the fashion industry. Zalando has provided a dataset consisting of various products that they have listed on their website and various other products that are listed on their competitor's website. Their proposed challenge is to match products listed on Zalando to similar products that are listed on their competitor's website. To tackle this challenge, I created word vectors using term frequency-inverse document frequency (TF-IDF) from feature engineered text content that describes a listed product. I then used a model that finds the similarity amongst products using cosine similarity metric over these word vectors. The resulting matches obtained from this model are highly efficient and intuitive.

CCS Concepts: • **Product Matching →TFIDF Word Vectorization**

Keywords: product matching, cosine similarity, TFIDF, similarity measure

1 INTRODUCTION

E-commerce platforms work on a business-to-consumer basis where they allow manufacturers, retailers or brands to list their products on their platforms and consumers can purchase these products from the platform. To ensure consumers have a holistic experience, platforms tend to list as many details as possible. Product features listed on these platforms allow consumers to make an informed decision. Domination for platforms in this market is tough owing to the large number of competitors that have entered the market in recent years. All these platforms are competing to ensure customer satisfaction and retention by keeping their prices attractive and their collection thorough. To maximize sales and customer retention, these e-commerce platforms must make sure their product prices and range is better than other competitors.

Zalando is a key competitor in the ecommerce fashion industry, and they need to ensure the products listed on their website are priced competitively to stay ahead of said competition. This competitive pricing strategy can only be achieved through identifying the selling price of similar products listed by their competitors.

On the one hand. product matching should not be a treacherous task since there are many descriptive features available for each product listed on these various platforms. But on the other hand, these large numbers of features prove to pose a challenge since we need to carefully choose the features that go into a machine learning model otherwise it may cause overfitting or requirement of a large amount of computing space.

In my approach, I decided to use textual features in order to compare the various products listed by Zalando and their competitors. Using cosine similarity measure on TF-IDF word vectors, my model was able to find a good amount of excellent true matches.

2 RELATED WORK

Researchers have been working on improving product matching algorithms' accuracy more intensively in recent years. With such great advancements in neural networks, researchers have leaned towards using the images of products along with their description to obtain a better accuracy of matches [1]. Gupte et al. (2021) conclusively proves the benefit of using a weighted multi-modal approach for product matching by using both images and text descriptions. Their model conclusively outperformed other single modal models. Traditionally, TF-IDF is used to analyze text description of products to classify and/or categorize them. But the same principles can be applied to product matching problems as well. Liu et al. (2018) improve upon the basic TF-IDF methodology to introduce a weighted TF-IDF method of analyzing text content for the purpose of classification. Their approach does outperform the traditional approaches thereby improving the accuracy of their classification. Although related work suggests that use of neural networks would produce better results when it comes to classification or matching tasks, the computing cost is significantly larger than simpler text-based approaches.

3 DATASET

The dataset provided by Zalando consists of three main files: offers_training.parquet (consisting of all offers used to generate training matches), offers_test.parquet (consisting of all offers for whom matches need to be generated) and matches_training.parquet. Both the 'offers' dataset consists of many features that can be used to find similarity amongst products. The dataset comprises descriptive text data features, numeric price data features and image URLs that point to the actual images on the products. All these features can be used cumulatively to find similarity amongst products, but such a model would require a tremendous amount of computing power and time.

Along with the dataset, Zalando also provided a sample Jupyter notebook that helped me tremendously in interpreting the dataset. Each offer in either the testing or training dataset consisted of the following columns: {offer_id, shop, lang, brand, color, title, description, price, URL}. The purpose and explanation of each of these is also provided by Zalando. They have also performed some

exploratory research on the training offers to help explain the split by shops of the various offers in the dataset.

I performed similar exploratory tasks on the testing dataset to obtain insights that helped me develop my effective strategy to obtain matches. I explored the various columns and their inputs for various offers and one very evident conclusion was that this dataset required significant text pre-processing before it could be used in any model. There is significant inconsistency in general nomenclature of the text inputs in title, description color and brand columns.

The total number of offers in the test dataset were 10674, of which Zalando up approximately 30 percent. There were some null values in the dataset, but they are only missing at random. All the products belonged to a total of 164 unique brands. The following are a couple of interesting visual representations of the testing dataset:

Fig. 1. Number of offers in Zalando and Aboutyou that have 'jeans' in their title in test dataset

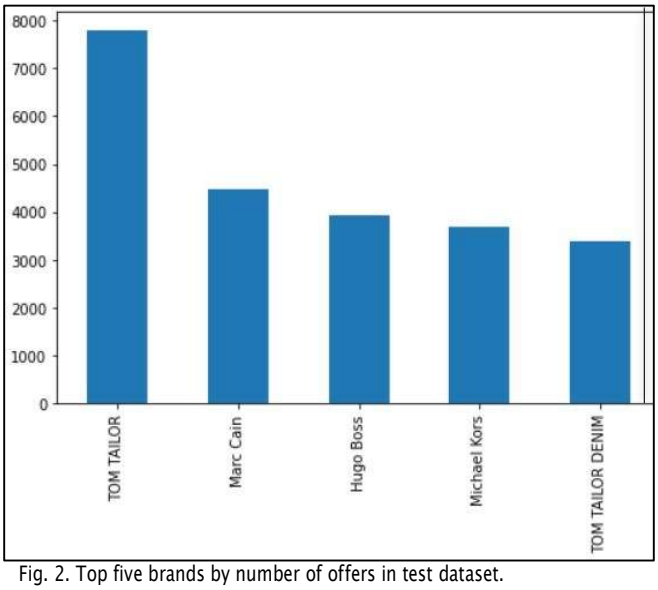

Fig. 2. Top five brands by number of offers in test dataset.

4 METHODOLOGY

Before I could work on the text content in the various columns of the test dataset, I had to perform some text pre-processing steps to set up good quality data for processing. I performed the following text pre-processing steps in the very order listed below:

- Removing punctuations
- Lowercasing
- Removing Stop Words
- Stemming

Once the data was clean and processed, I created a text feature based on the title, brand, and color of a product as the main comparison feature to match products. Product title, brand and color may not be a unique factor if used individually since there may be many products of the same brand, or of the color or title. A combined feature provides a better approach as it makes the comparisons much more unique and yield very accurate matches.

I used this new feature and created word vectors based on their TF-IDF weights in a vector space. The TF-IDF weight is a weight often utilized in information retrieval and text mining. This weight is a statistical measure used to assess how vital a phrase is to a report in a group or corpus. These vectors each represent a product offer in terms of their title, brand and color. In this vector space

I used the cosine similarity measure to determine the similarity between two vectors and set a threshold value of similarity based on experimentation in order to fetch the most similar match for every product. I had to reverse-engineer the matches obtained in terms of the offer ids on the corresponding offers to comply with the requirements of the challenge.

The detailed code for this implementation can be found here: https://github.com/romilskr3/CA684_Machine_Learning_Assignment/blob/main/CA684%20ML.ipynb

5 EVALUATION

Using the methodology described above, I was able to retrieve a total of 1810 offer matches between Zalando and their competitors at a minimum cosine similarity value of 0.5. Upon analyzing some of these matches manually, a vast majority of them are correct true matches. Upon lowering the minimum cosine similarity value, I could have obtained a higher number of matches, but the accuracy and F1 score keeps decreasing in that case.

6 FUTURE WORK

The following are some prospectus future improvements that can be made to this challenge in order to obtain a more efficient outcome:

- Using cleaned and pre-processed descriptions to calculate TF-IDF and subsequent vectors in the vector space.
- Feeding the images of offers into a form of CNN to factor in data from images of the offers.
- Identifying and using keywords in the text description of the offers to create word vectors that can be used by a similarity measure such as Jaro similarity or Levenshtein distance.

7 REFERENCES

[1] Gupte, K., Pang, L., Vuyyuri, H. and Pasumarty, S., 2021, December. Multimodal Product Matching and Category Mapping: Text+ Image based Deep Neural Network. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 4500-4505). IEEE.

[2] Liu, C.Z., Sheng, Y.X., Wei, Z.Q. and Yang, Y.Q., 2018, August. Research of text classification based on improved TF-IDF algorithm. In 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE) (pp. 218-222). IEEE