ARTIFICIAL INTELLIGENCE THEORY and APPLICATIONS

ISSN: 2757-9778 || **ISBN :** 978-605-69730-2-4

More information available at **aita.bakircay.edu.tr**

E-commerce Product Categorization Using Big Data Analytics

Sedat USLUOĞLU [1], Deniz KILINÇ [2], Fatma BOZYİĞİT [2,*]

[1] *Boyner Group, R&D Department, İstanbul, Turkey*
[2] *Izmir Bakircay University, Dept. of Computer Engineering, Turkey*

* Corresponding Author: Izmir Bakircay University, Dept. of Computer Engineering, Turkey
Tel.: +90 232 2930000. E-Mail: fatma.bozyigit@bakircay.edu.tr

Publication Information

Keywords :
- E-commerce platforms;
- Product categorization;
- Machine learning;
- Big data analytic tools.

Category : Research article

Received : 26.07.2021
Accepted : 25.09.2021

A B S T R A C T

E-commerce platforms need to have a well-managed online product catalog to make products easily accessible. However, the organization of catalog and categorization of products can be time-consuming due to the large volume of product data in e-commerce. In this direction, our study aims to develop an accurate categorization of product data with the adoption of big data analytics. Accordingly, various machine learning algorithms (Support Vector Machine, Naive Bayes, and Stochastic Gradient Descent) were utilized to organize online catalogs from Spark MLLib. Performed classifiers were trained and tested on product catalog data collected from a fashion retailer in Turkey, Boyner Group, which combines cutting-edge digital services with a vast network of exciting stores.

1. Introduction

With the advanced technological developments in recent years, data production and sharing between the digital platforms have increased consistently. Since the huge amount of electronically exchanged information, the data processing reveals a significant need to get meaningful predictions about future processes. Considering the usage of technology in our daily lives, we can easily see a huge amount of data cycling between the customers and suppliers, such as Business to Business (B2B) e-commerce [1]. B2B marketplaces may use customer information and behaviors to extract semantic patterns, increasing their next sales.

B2B platforms, the intermediate layer for marketing communications, provide many advantages to their customers. Since a B2B electronic commerce must provide a common protocol for data exchange between companies and shoppers, content management is one of the major issues to be handled. Thus, it is common to organize the online catalog to ensure that products are listed in the correct category [2]. However, the manual placement of products in relevant categories may be challenging due to the large volume of data on e-commerce platforms. Moreover, organizing the product catalog may cause expensive errors to be fixed since it affects shopping transactions.

A recent literature review on this topic shows that more researchers have recently focused on the organization of product catalogs for e-commerce. In this study, the product descriptions were categorized using ((Naive Bayes (NB), Stochastic Gradient Descent (SGD), and Support Vector Machines (SVM) ML algorithms using big data frameworks. To the best of our knowledge, the main limitation in the previous researches is the absence of a well-designed and comprehensive dataset. To fill this gap, we created data set including a well-known B2B marketplace (Boyner Group) [3] retail data. It contributes to the literature by experimenting with different ML classifiers with different feature vectorization strategies that categorize product catalogue data written in Turkish. It applies TF-IDF, word2vec and HashTrick text representation methods and then uses the feature reduction technique since the text data produces a high dimensional training space for ML algorithms. Then, it compares ML classifiers using Mean Accuracy and Standard Deviation performance metrics. The remaining parts of the study are organized as follows. Section 2 discusses the related works. Section 3 presents the materials and methods used in our proposed model. Section 4 provides an evaluation of our approach based on real-world data provided by Boyner Group. Section 5 provides conclusions and future directions.

2. Related Works

Increasing competition among businesses within developing technology forces retailers to manage e-commerce product catalogues. Thus, they give customers the confidence to buy. More researchers have recently addressed mapping the products into corresponding places in a website, an integral part of a practical online user experience. For instance, Shen et al. [4] proposed a technique based on extracting domain-specific features to increase the classification accuracy of item categorization in the e-commerce domain. They have used deep domain information and linguistic patterns to categorize product groups using SVM classification algorithm. They have proved the effectiveness of their proposed approach using a real-world commercial platform (eBay). The other study utilized by Mathivanan et al. [5] compares performance results of five frequently used ML methods such as NB, k-Nearest Neighbour (kNN), Decision Tree (DT), SVM, and Random Forest (RF) to determine the most efficient method for product title classification. Experimental results demonstrate that kNN has better performance scores than the other experimented classification models. Another study was utilized by Zahavy et al. [6]. They suggested a decision-level fusion approach to classifying multi-modal products by using Deep Neural Networks. The experimental results of the study show that the proposed method improves the top-1 accuracy on a real data set, including both text and image inputs. In another study, Manchuska et al. [7] introduced a recursive product catalogue pattern matching and learning scheme to categorize products in the e-commercial platform. They have used an expectation-maximization-based NB classifier. The experimental results of the study present that recursive product learning improves the accuracy of product category identification.

In recent years, big data has emerged in the e-commerce industry in which the data is heterogeneous, distributed, and private. With the help of big data analytic, e-commerce data is processed efficiently at a higher conversion rate to enhance decision-making [8]. Moreover, big data analytics have various advantages on e-commerce platforms: better communication between research and development, improving customer service, and improving pricing strategies [8]. Akter and Wamba [9] published a study to identify various conceptual dimensions of big data in e-commerce by focusing on big data analytics.

3. Materials and Methods

3.1. Dataset

When the studies covered in the literature are examined, it is observed that they have some limitations in the dataset used. Thus, we constructed a comprehensive dataset with the use of product catalogues of Boyner Group e-commerce platform.

The dataset includes 1.5 million product information to be mapped regarding five different classes such as gender (man, woman), colour (blue, red, etc.), type (swim wear, pants, and sneaker), and group (sport, kid, baby, accessories, bag and shoes, clothing, cosmetics, etc.).

3.2. Big data technologies

3.2.1. Apache Spark

Apache Spark is a framework that can instantly perform processing tasks on extensive data sets [10]. It supports many operations which are essential for data analysis (i.e., data transformations and actions). Moreover, it also offers distributed data processing tasks across multiple computers and cluster computing for its analytics power as well as its storage. Thus, a Spark application can be developed using a computing cluster's CPU, memory, and storage resources.

Running a Spark application includes five key entities shown in Figure 1: a driver program, a cluster manager, workers, executors, and tasks. The Driver is the process that converts a user application to tasks and then coordinates them with a cluster manager on executors of the Spark program. Along with executing the app, SparkSession which represents a connection to a computing cluster is created. Spark Executors are the processes at the worker nodes which are responsible for running the assigned tasks.

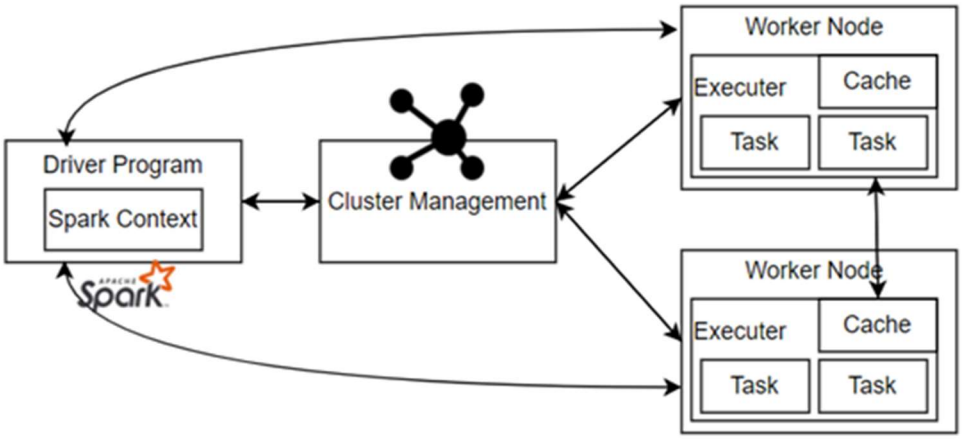

Figure 1. Key entities for running a Spark application

Real-time analytics has many benefits for developing industries such as tourism, health, and marketing. For example, a user's behaviour at a particular time can be analysed rather than on historical behaviour to realize some predictions. In this direction, Apache spark having four main components (Spark SQL, Streaming, MLlib, and Graphx) one of the widely used tool by the researchers. Spark Streaming facilitates to process

real-time streaming data. It integrates a wide variety of many sources, including Hadoop Distributed File System (HDFS), Kafka, and Twitter. Then the data can be processed using complex algorithms and sent to file systems, databases, and live dashboards. MLlib library aims to utilize practical machine learning algorithms built on top-level DataFrame-based APIs [11]. It provides baseline machine learning methods such as classification, regression, and clustering. Graph Calculation (GraphX) allows users to create interactive graph-based calculations.

3.3. Pre-processing and feature representation model

The first step of product categorization is pre-processing, which is performed before the feature representation. Pre-processing involves a series of methods, as shown in Figure 2. We first performed tokenization to split phrases into tokens such as words, numbers, and punctuation marks. Then, the stop words, numbers, and punctuation marks with no specific meaning are excluded from being ignored during indexing terms. In the last step of pre-processing, we utilized stemming to reduce inflectional morphemes in words. Figure 2 illustrates the pre-processing steps applied in this study.

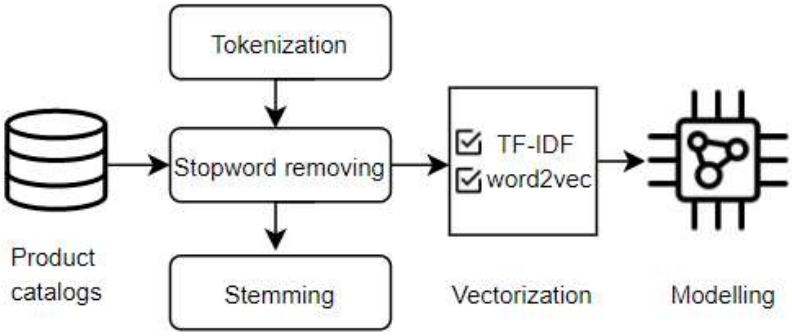

Figure 2. Steps in the preprocessing task

After pre-processing step, product descriptions need to be converted to an appropriate form of feature representation so that the classifiers can perform better. The vector space model (VSM) is one of the most used text representation models. There are two sorts of research commonly applied to generate VSM: indexing and term weighting. This study used the Term Frequency Inverse Document Frequency (TF-IDF) term weighting and word2vec methods to vectorize product descriptions.

The high dimensionality of textual data causes high costs on model training and execution, and extra efforts may be necessary to optimize the performance of ML algorithms. Therefore, we also performed Spark's feature hashing technique for vectorization. The method maps feature values to indices in the feature vector called “Hashing Trick” [12].

3.4. Machine learning algorithms

In this section, we provided a brief overview of classification algorithms: NB, SGD, and SVM.

Naïve Bayes (NB): NB classifier is one of the mostly preferred ML approach in the text classification because of being fast, easy to implement, and effective [13]. It is a simple probabilistic classifier based on

Bayes' theorem given in Equation 1 with the assumption of independence among instances. This means that a NB classifier presupposes that a particular instance in a class is unrelated to the presence of any other instance.

$$P(c \mid D) = \frac{P(c)P(D|c)}{P(D)} \quad (1)$$

where *P(c)* is the prior probability of the appearing in the class *c*, *P(D)* is the prior probability of a given document, and *P(D|c)* is the conditional probability of the document *D* given class *c*.

NB classifier has been commonly used for document classification and has been shown to produce remarkable performance. One advantage of NB classifier is that a small amount of training data is enough to estimate classifier parameters. On the other hand, a main disadvantage of NB is the conditional independence presumption, which is infrequently valid in real-world applications.

Gradient Descent Algorithm (GDA): GD, a first-order optimization algorithm, considers the first derivative when the updates occur on the parameters. It is one of the most popular algorithms to perform optimization. However, it has some disadvantages. The first, it requires many steps for converging the local optimum and second, it is not guaranteed to converge the global minimum if there are several local minimums in the error surface. To overcome these difficulties, one common variation of gradient descent is Stochastic GD (SGD). The SGD algorithm is obtained by Equation 2.

$$w_i \leftarrow w_i + v\,(t - o)x_i \quad (2)$$

where *t*, *o*, and x_i are target value, unit output, and i_{th} input for the training example.

Support Vector Machine (SVM): SVM is a widely used supervised machine learning technique in text categorization tasks [14]. The method determines an optimal hyperplane that can separate the two classes of given samples by a maximal margin. The optimal hyperplane refers to the shortest distance between the nearest data points and any point on the hyperplane (see Figure 3). The subset of data that lies on the margin is called support vectors. New unlabelled data is obtained to a class based on its geometric position relative to the classifier function.

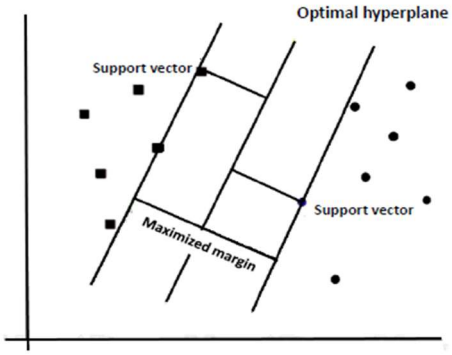

Figure 3. Decision boundary and margin of SVM

4. Experimental Study

In this study, we presented the implementation of product categorization on real-world e-commerce catalog data. Figure 4 shows the general framework of the proposed model. After the pre-processing steps were utilized, some baseline ML algorithms from Spark MLLib, which are commonly used to classify the textual data.

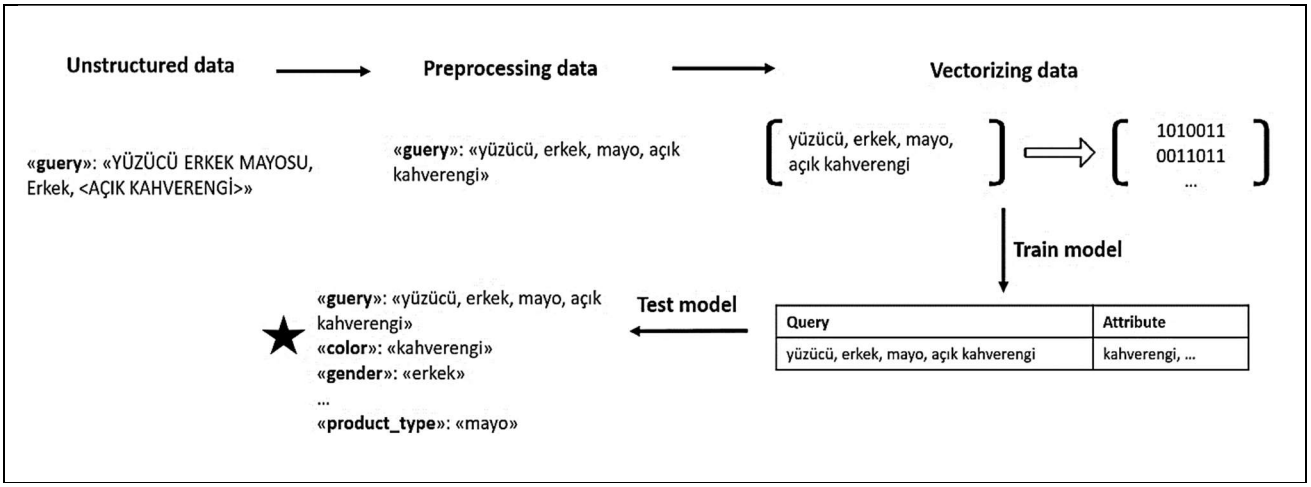

Figure 4. The general illustration of our proposed workflow

The evaluation results of each machine learning method were obtained with the use of 10-fold cross validation. The performance scores of the experimented classifiers in terms of Mean Accuracy and Standard Deviation are presented in Table 1.

Table 1. Performance results of performed classifiers

Model	Mean Accuracy	Standard Deviation
SVM with TF-IDF	**0.93**	**0.036**
SVM with HashTrick	0.96	0.4
SVM with word2vec	**0.97**	**0.042**
SGD with TF-IDF	0.92	0.051
SGD with HashTrick	0.94	0.057
SGD with word2vec	0.94	0.50
MNB with TF-IDF	0.88	0.087
MNB with HashTrick	0.89	0.094
MNB with word2vec	0.91	0.067

In the experimental studies, the real-world data, including product descriptions, were used to perform categorization tasks. Table 1 compares the Mean Accuracy and Standard Deviation obtained by performing ML algorithms such as NB, SGD, and SVM. The results show that the best performing model is word2vec based SVM with a mean accuracy of 97%. The closest criteria result to SVM was achieved by SGD with a mean accuracy of the 94%. On the other hand, word2vec based MNB was the worst-performing algorithm with an 88% mean accuracy score among classifiers.

Another point to be noticed is that the ML algorithms with the word2vec method performed better than ones with the TF-IDF encoding method. The good performance of word2vec representation can probably be based on the amount of training data.

5. Conclusion

In e-commerce platforms, it is common to organize the online catalog to ensure that products are listed in the correct category. However, the manual placement of products in relevant categories may be challenging due to the large volume of data on e-commerce platforms. Moreover, organizing the product catalog may cause expensive errors to be fixed since it affects shopping transactions. We experimented with three ML algorithms (NB, SVM, and SGD) using big data analytics to organize product catalogs in a well-known e-commerce website, boyner.com. Consequently, the products were tagged through gender (man, woman), color, type (swimwear, pants, and sneaker), and group (sport, accessories, clothing, cosmetics, etc.) class labels. Experimental results showed that the best-performing method is SVM, and it achieved a 97% value of mean accuracy. Considering the results, it can be concluded that this study appears promising for future studies on product categorization systems.

Acknowledgment

Funding for this work was partially supported by Research and Development Center of Boyner Group accredited on Turkey - Ministry of Science.

References

[1] Sila, I. (2013). Factors affecting the adoption of B2B e-commerce technologies. Electronic commerce research, 13(2), 199-236.
[2] Umaashankar, V., & Prakash, A. (2019). Atlas: A Dataset and Benchmark for E-commerce Clothing Product Categorization. arXiv preprint arXiv:1908.08984.
[3] Boyner. (2021, May, 15). Internetin boyner'i online alışverisin adresi [Online]. Available: https://www.boynergrup.com/en
[4] Shen, D., Ruvini, J. D., Somaiya, M., & Sundaresan, N. (2011, October). Item categorization in the e-commerce domain. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 1921-1924).
[5] Mathivanan, N. M. N., MdGhani, N. A., & Janor, R. M. (2019). Performance analysis of supervised learning models for product title classification. IAES International Journal of Artificial Intelligence, 8(3), 228.
[6] Zahavy, T., Magnani, A., Krishnan, A., & Mannor, S. (2016). Is a picture worth a thousand words? a deep multi-modal fusion architecture for product classification in e-commerce. arXiv preprint arXiv:1611.09534.
[7] Manchusha, K. N. R., & Renukadevi, P. Recursive Product Catalog Pattern Matching and Learning for Categorization of Products in Commercial Portal.
[8] Minelli, M., Chambers, M., & Dhiraj, A. (2013). Big data, big analytics: emerging business intelligence and analytic trends for today's businesses (Vol. 578). John Wiley & Sons.
[9] Akter, S., & Wamba, S. F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. Electronic Markets, 26(2), 173-194.

[10] Salloum, S., Dautov, R., Chen, X., Peng, P. X., & Huang, J. Z. (2016). Big data analytics on Apache Spark. International Journal of Data Science and Analytics, 1(3), 145-164.
[11] Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., ... & Talwalkar, A. (2016). Mllib: Machine learning in apache spark. The Journal of Machine Learning Research, 17(1), 1235-1241.
[12] Kılınç, D. (2019). A spark-based big data analysis framework for real-time sentiment prediction on streaming data. Software: Practice and Experience, 49(9), 1352-1364.
[13] Bozyiğit, A., Utku, S., & Nasibov, E. (2021). Cyberbullying detection: Utilizing social media features. Expert Systems with Applications, 179, 115001.
[14] Özçift, A., Kilinc, D., & Bozyigit, F. (2019). Application of grid search parameter optimized Bayesian logistic regression algorithm to detect cyberbullying in Turkish microblog data. Academic Platform Journal of Engineering and Science, 7(3), 355-361.