CTU Journal of Innovation and Sustainable Development
ISSN 2615-9422 | e-ISSN 2815-5602

DOI: 10.22144/ctu.jen.2023.018

Identify and predict incorrect prices by Machine Learning Model

Lam Thanh Toan[1*], Nguyen Xuan Ha Giang[1], and Nguyen Hoang Thuan[2]

[1]*Information Department, Can Tho University of Technology, Viet Nam*

[2]*The Business School, RMIT University, Viet Nam*

**Correspondance: Lam Thanh Toan (email: lttoan@ctuet.edu.vn)*

Article info.

Received 18 Jan 2023
Revised 03 Mar 2023
Accepted 06 Apr 2023

Keywords

Machine Learning, Random Forest, Price supervision

ABSTRACT

Electronic commerce (e-commerce) brings huge advantages to businesses for selling products through multiple online shops. However, companies have difficulties in supervising the prices of products set by different retail shops on e-commerce platforms. Addressing these difficulties, we suggest a method to identify and predict products that sell at incorrect prices using a machine learning model combined price analysis. The study uses four machine learning models: K-nearest Neighbor (KNN), Random Forest (RF), Support Vector Machine (SVM), and Multinomial Naive Bayes (MNB) and two text-based information extraction methods: BoW and TF-IDF to find to the best method. The research results show that the RF model and text-based information extraction method by the BoW provide more average accuracy than other specific models, when experimenting on the filter dataset the average accuracy after 10 runs are RF: 98.06%, SVM: 83.92%, MNB: 92.21%, KNN: 94.06%. Experimental results on the product dataset have an accuracy of RF: 83.02%, SVM: 55%, MNB: 79.33%, KNN: 79.36%.

1. INTRODUCTION

Buying and selling over electronic commerce (e-commerce) has been popular in the world, bringing multiple benefits to customers, such as discounts, free shipping, and promotions (Muljono, 2018). Companies are increasingly promoting product sales in e-markets because of the increasing demand for online shopping by customers and creating competitive advantages in the market (Gielens, 2019; Hamad, 2018; Tan, 2019). Because of the increasing demand for online shopping by customers and creating competitive advantages in the market, companies continue to pour resources into their e-commerce operations. The benefits lead to appear common systems and web-based tools that companies use for faster and more reliable communications. Advanced systems help enhancement of planning, forecasting and replenishment. These e-commerce systems address the needs of organizations, merchants and consumers to cut costs while improving the speed of service delivery and increasing the quality and quantity of commodities and services, assuring customers getting the best deal.

The usage of e-commerce by companies has changed recently. Previously, companies used e-commerce platforms to sell products directly to customers. Recently, companies have multiple distributors and retail shops to sell products to customers. These distributors and retails have different selling tactics, including setting different prices and discounts to attract buyers. Sometimes, distributors compete to offer discount policies to attract customers and create business advantages in the same system. This leads to two problems: (1) customers are not accessible to the full collective

benefits of companies' incentive programs, (2) the potential within controlling e-retailing to create a unique or value-added feature through communities is affected. If there is no policy to control the price of products on the e-market, it will cause price fluctuations and confusion for buyers and reduce confidence buyers in the company's products.

The challenge has also been highlighted from the research perspective. Existing research has addressed this challenge through (Agrawal et al., 2019) predict the price and price trend of stocks by applying optimal Long short-term memory (LSTM) deep learning achieved mean prediction accuracy 59.25%. (Hartford et al., 2017) uses deep neural nets to predict the value of some outcome variable in airline sales underprice variable. (Nassar et al., 2020) compares the price prediction models performance of deep learning (DL) models with statistical as well as standard machine learning (ML) models of price prediction of fresh produce in market to protect retailers from overpriced fresh produce. It is found that the conventional ML models perform less performant as compared with the simple or compound DL models; the simple DL models, LSTM, are outperformed by the compound one, the Convolutional LSTM Recurrent Neural Network (CNN-LSTM).

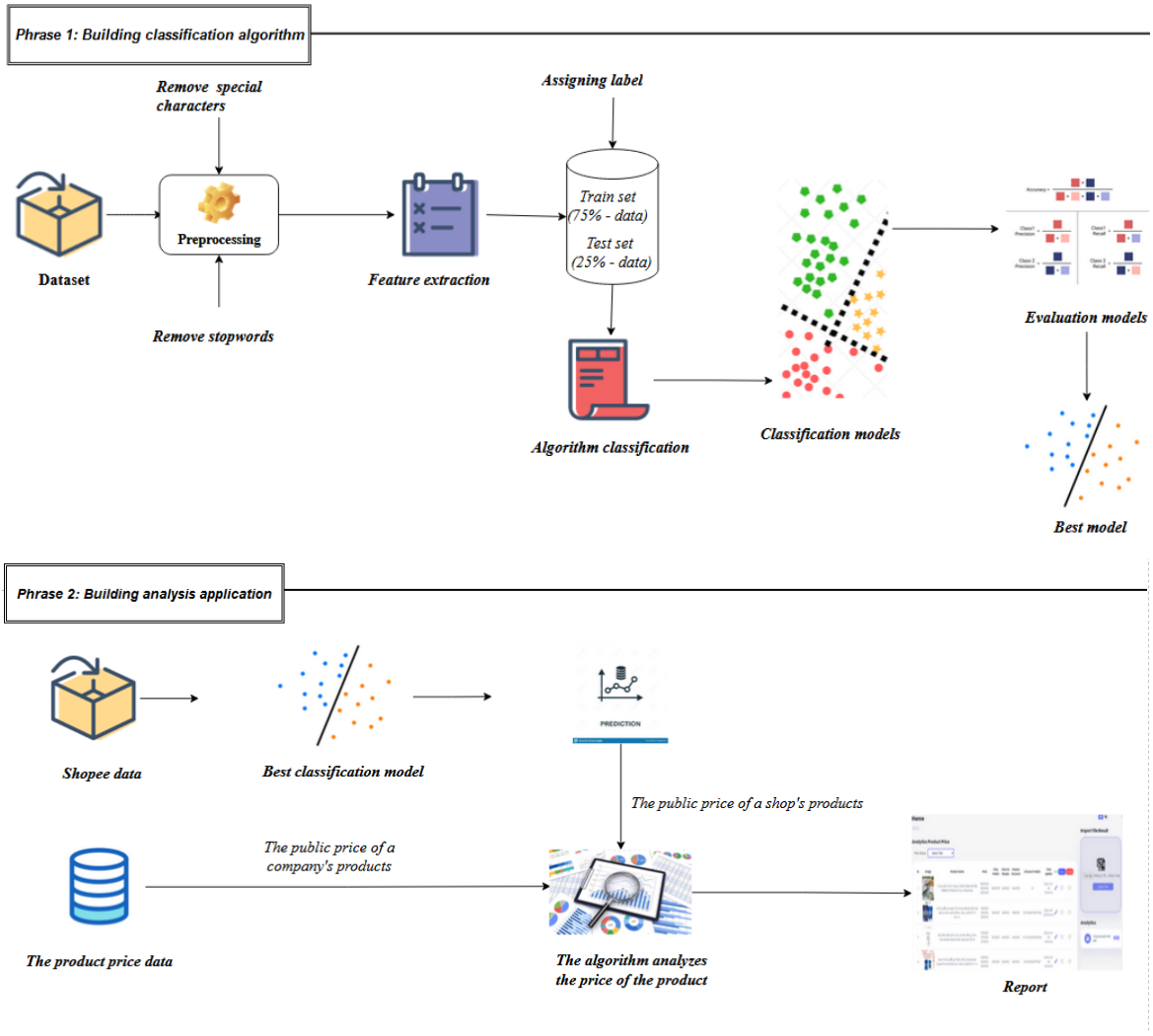


Figure 1. Processing Identification and predicting products that sell at incorrect prices of the company by machine learning model combined price analysis

Recent development of data mining provides a promising approach to address the challenge. Data mining on the e-commerce platform is mainly text mining, such as analyzing the user's emotions

(Shalehanny, 2021; Agustina, 2020). There are two stages in analyzing user emotions: the first stage, processing the text-by-text processing techniques (Joachims, 1998). The second stage is constructing a machine learning model for classification. Machine learning models are commonly used as KNN algorithm (Fix, 1952), Naive Bayes (Good, 1965) decision trees (Quinlan, 1993; Breiman, 1984), Support Vector Machine learning (Vapnik, 1995), model aggregation algorithms, including Boosting (Freund, 1995; Breiman, 1998) and random forests (Breiman, 2001). These models are promising to address the above identified challenge.

In this paper, we suggest a method to identify and predict products that sell uncorrected price policy by machine learning models in combination with price analysis. The method comprises two stages. In stage 1, we train two text classification models: the noise filtering model and identifying the product model. In stage 2, the models trained in stage 1 are used to conduct the experiment according to the procedure shown in Figure 1. In which, the experimental dataset is crawled directly from Shopee e-commerce. This research contributes to exploring the design and adoption of product classification from the Shopee e-commerce website. It also supports to fulfilling the need for digital transformation for business. We further design an application that allows end users to automate product classification and analyze prices. Then, the final stage of this progress shows product lists infringe policy prices that the company announced to customers. Our application helps companies more active in supervising product price of the distributors.

The research makes both theoretical and practical contributions. From a theoretical perspective, our study aims at comparing performance of various classifiers in machine learning algorithms and developing more accurate mis-price prediction model for the real electronic market. From a practical perspective, producers and companies can employ a machine learning based mis-price prediction model for better real incorrect price appraisal, and identifying produce in supply chain.

The rest of this paper is organized as follows. Part 2 reviews some present related work. Part 3 presents the method of collecting and processing text data, and the model building method. Part 4 and 5 presents methods of analyzing and predicting products sold at wrong prices. Part 6 presents experimental results, and finally, conclusions and future works.

2. RELATED WORK

Text feature extraction is one step important for text classification problem. According to the characteristics of each dataset, there are different text feature extraction methods. Some popular techniques are used to extract text features such as Term Frequency-Inverse document Frequency (TF-IDF), Word2Vec, Global Vectors for word representation (GloVe). Choosing a classification model for the best accuracy classification result is

the most important step in the text classification model. There are some machine learning models used to activities such as Logistic Regression model (LR), K-nearest Neighbor (KNN) (Fix, 1952), Support Vector Machine (SVM) (Vapnik, 1995). Random Forest (Breiman, Random forests, 2001), Naïve Bayes (Good, 1965).

Currently, in the field of the Vietnamese text data classification problem, a number of studies have been researched with very feasible results. Neural network (Van, 2017) that data collected from electronic news sites such as Vnexpress.net, Tuoitre.vn, Thanhnien.vn, and Nld.com.vn with classification accuracy of 99.75%. Specifically, it is mentioned that Vietnamese text classification with SVM (Nguyen, 2019) using data collected from the electronic page Vnexpress.net, Tuoitre.vn, Thanhnien.vn, Teleport-pro.softonic.com, and Nld.com.vn with classification accuracy 94%.

Related research for prices supervisions have been seen in current research. Lamon (Lamon et al., 2017) analyzes the ability of news and social media data to predict price fluctuations basing on Logistic Regression, Linear Support Vector Machine, multinomial Naive Bayes, and Bernoulli Naïve Bayes. (Park et al., 2015) experimented in C4.5, Ripper, Naïve Bayesian, AdaBoost for predicting housing price prediction. (Kohli et al., 2015) results the best performance from the AdaBoost algorithm as compared to other techniques: Gradient boosting, SVM, Random forest. The all four machine learning algorithms determined influence the stock trend and predict the behavior of stock exchange. (An et al., 2019) suggested linear regression-based machine learning algorithm to predict oil prices.

However, little research has been performed on developing a better produce mis-price prediction model cooperating price analysis through performance evaluations of several machine

learning algorithms. Our study is experimented on several text extraction feature methods and machine learning algorithms combined price analysis to determine the best solution in identifying and predicting products that sell at incorrect prices of the company.

3. DATA COLLECTION

3.1. Crawling

In this research, we use a dataset of product names that users post on the Shopee. The data were directly collected by the Shopee-crawler library. There are two methods of collecting data from the Shopee-crawler library: keywords, and shop link. Figure 3 shows the visualization of the data_product dataset, which contains 33 products displayed on two dimensions.

In this study, we get data from Shopee by keyword. In which, the keywords we used to collect data were the names of three cosmetic brands: Drceutics (https://drceutics.vn/), Cosmeheal (http://cosmeheal.vn/), Mibitiprudente and (http://www.mibitiprudente.vn/). There are five attributes achieved from the Shopee electronic market: product name, price, discount, product link, and ink product image. In which, we use the product name to build the product identification model and the remaining information is used to check the information in the price analysis stage.

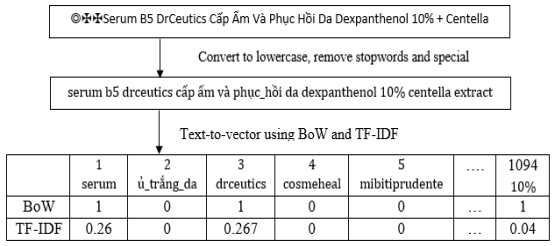
Our experiments assign the label and split data into two datasets: The first one is used to train the product filtering model in the products list basing on brand, and the dataset name is set as data_filter. The second dataset, named data_product, is used to train the product recognition model. There were 33 products belonging to three mentioned brands. Each product was a label. Both datasets were collected directly from Shopee by Shopee-crawler library in 2022. Data details are shown in Table 1.

Table 1. Detail datasets

Dataset	Number of data	Number of classes
Data_filter	2078	2
Data_product	2449	33

3.2. Preprocessing

Data preprocessing is the stage of data cleaning before constructing the training model, the activity supports construct the training model to get the best accuracy.



	1 serum	2 ù_trắng_da	3 drceutics	4 cosmeheal	5 mibitiprudente	....	1094 10%
BoW	1	0	1	0	0	...	1
TF-IDF	0.26	0	0.267	0	0	...	0.04

Figure 2. Data preprocessing period

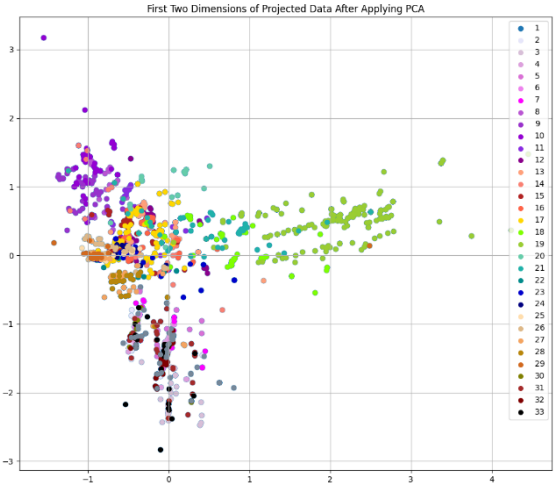


Figure 3. Visualization dataset data_product 33 products (33 labels)

The period consists of three steps: First, the data are converted to lowercase. Next, stop words, animations, and common special characters such as !*&^%$()-#@,.';=-?/+{}[] are removed. The list of Vietnamese stop words included 1,942 words as suggested by (Van-Duyet, 2017) are imported. Finally, the outcomes from separating the text fragments into words and representing the words into vectors (Vector Space Model) are input of training models (KNN, SVM, RF). In the training models, Vietnamese word separation technique performed by Pyvi tool with F1 score of 0.985.

3.3. Text vectorization

3.3.1. Bag of word model (BoW)

This model refers to a concept of a collection containing text types, including words, sentences, and plain documents. BoW model used in Natural Language Processing (NLP) and Information Retrieval (IR) for representing text. It is only interested in words duplicates and not focus on grammar, semantics and order of words appearance in sentences. Document classification methods

basing on words occurrence or rate mainly used the dataset model to implement the features training classifiers. Setting n is the number of words of the dictionary, mapping each text document into n-dimensional space is concluded at the end of the dataset. A n-dimension space was produced in which the appearance frequency of the word in the text document is the weight of each vector.

Vectorizing the text by BoW technique on the data_filter dataset, we retrieve a dataset training of size 2078 x 1561. Specifically, 2078 is the number of data in the dataset, 1561 is the number of words in the dictionary. Similarly, when applying the technique on the data_product dataset, we get a training set of size 2449 x 1094.

3.3.2. Term frequency–inverse document frequency (Tf – idf)

Another statistical measure, TF-IDF is used to evaluate relevance of a word to a document in a text document dataset. TF-IDF is a metric that is composed by the two quantities Term Frequency (TF) and Inverse Document Frequency (IDF). TF estimates the occurrence probability of a word normalized by the total word frequency in the document. On the other hand, idf valuates log of the inverse probability of the total of the documents in dataset divided by the total of documents where the specific word appears. Searching document and extracting information by TF-IDF works based on the increase in the number of occurrences of a word in the document and the number of documents that contain the word.

$$TF * IDF(t_i, d_j, D) = TF(t_i, d_j) * IDF(t_i, D) \quad (1)$$

The formula of weight:

$$\begin{cases} W_{ij} = \left(1 + \log(f_{ij})\right) \log\frac{N}{df_i}, Nếu\ f_{ij} \geq 1 \\ W_{ij} = 0 \qquad\qquad\qquad\qquad\quad , Nếu\ f_{ij} = 0 \end{cases} \quad (2)$$

4. MODELING

We program and test four machine learning models in order to attempt to predict the incorrect price flow in support chain to the customers. Through of progress, we also incorporate the analysis of the product's price history and the company's price in evaluating the performance of model for taking accuracy at the end of training. The mean prediction accuracy achieved using the proposed models is different. The KNN algorithm is showing the dominance in this price prediction.

4.1. K-Nearest Neighbor (KNN)

K-nearest neighbor (KNN) (Fix, 1952) was first introduced in 1951 and 1952 and then further study conducted by Cover and Hart in 1967. The algorithm was inspired by the hypothesis that "things that look alike must be alike" (Cover and Hart). Difference from other classification algorithms, KNN is a classification method based on data that was located closest to the objects without searching for a predictor within some predefined class of functions to determine the label on test point (Cover and Hart, 1967).

For a number k, the binary classification of KNN rule is defined as follows:

Input: sample of training $S = (x1, y1), (x2, y2),..., (x\text{m}, y\text{m})$

Output: for every point $x \in X$, return the majority label within $\{y\pi i(x)\ ;\ i \leq k\}$

If k = 1, then the 1-NN rule:

$$hs(x) = y\pi i(x) \quad (3)$$

Since the output depends on the number of k, therefore the appropriate of k has a significant impact on the result of the KNN algorithm. Figure 4 shows an illustration of the K-Nearest Neighbors (KNN) algorithm.

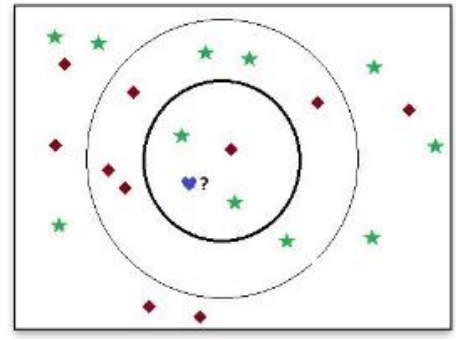


Figure 4. Illustration of KNN Algorithm

(Guo et al., 2003)

4.2. Random Forest (RF)

Random forest machine learning algorithm (Breiman, 2001) trains the classification model (Figure 5) through the following main steps: First, from a dataset with m elements and n attributes, take k random samples to build t decision trees. Next, t decision trees will be built individually for each sample, and each decision tree will produce an

output. Finally, using the output of t decision trees presents majority voting.

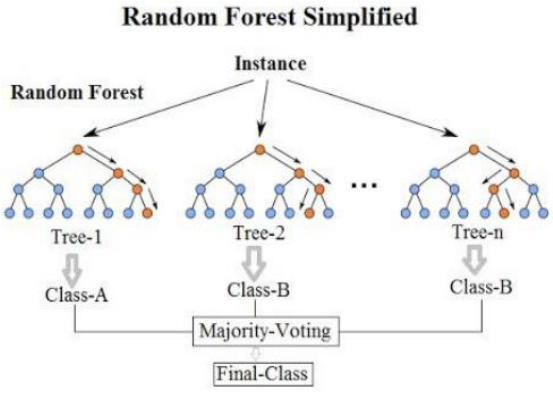


Figure 5. Showing Random Forest in simplified way

(Negandhi et al., 2019)

In this study, we used the built-in Random forest model in the scikit-learn library. We used the tool GridSearchCV employing a 10-fold cross-validation method to perform hyper-parameters tunning for the model and select the optimal set of parameters. The set of hyper-parameters for the RF model included n_estimators values of 300, 400, 500, criterion= 'gini', max_features='sqrt'. The SVM model achieved a mean cross-validated score of 0.7475 with the best estimator using the hyper-parameter settings of n_estimators = 300, criterion= 'gini', max_features='sqrt'

4.3. Support Vector Machine (SVM)

Consider the linear binary classification task by SVM model (Figure 6). Set M are the number of training data points $\overrightarrow{x_i}$, i = 1,..M. Having corresponding labels +1 or -1. Lets N is number of dimensions in the Feature Space. For this problem, data train denoted as follows:

$$\{(\overrightarrow{x_i}, y_i): \overrightarrow{x_i} \in R^N, y_i = \pm 1\} \quad (4)$$

The SVM model proposes a hyperplane concept and provides solutions for finding hyperplanes. The result is to produce the different domains space. Each of them contains the datatypes which belong to. In term, the hyperplane separates the multidimensional space into two layers by using the following formula:

$$\vec{w}.\overrightarrow{x} + b = 0 \quad (5)$$

Where $\vec{w}$ is a weight vector, b is bias. When changed $\vec{w}$ and b are the direction and distance from orgin to hyperplane changed. $\overrightarrow{x_i}$ assigned label +1 if $\vec{w}.\overrightarrow{x} + b \geq 1$, $\overrightarrow{x_i}$ assigned label +1 if $\vec{w}.\overrightarrow{x} + b \leq -1$. The distance between the two groups is $\frac{2}{\|\vec{w}\|}$ which is called the margin.

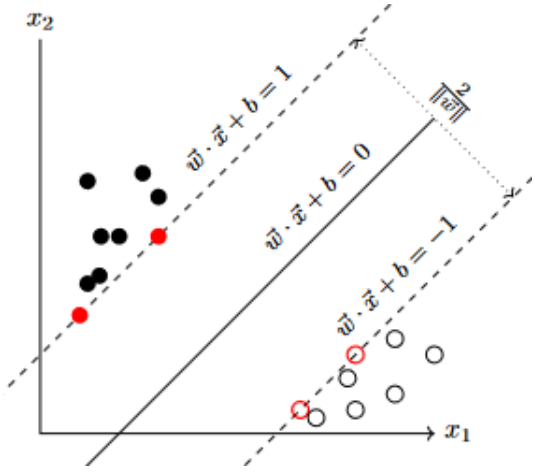


Figure 6. An example classification SVM with N=2

(Mahesh, 2020)

We utilized the built-in SVM model in the scikit-learn library along with the GridSearchCV tool, employing a 10-fold cross-validation method to perform hyper-parameter tuning for the model and select the optimal hyper-parameters. The set of hyper-parameters for the SVM model included Kernel values of 'linear' and 'rbf', C values of 0.1, 10, 100, 1000, and 10000, and Gamma values of 1, 0.1, 0.01, 0.001, and 0.0001. The SVM model achieved a mean cross-validated score of 0.768 with the best estimator using the hyper-parameter settings of Kernel = 'linear', C = 0.1 and gamma=1

4.4. Multinomial Naive Bayes

One of two classic variants of Naive Bayes algorithm (Loesche, 1994) is Multinomial Naive Bayes (MNB) for multinomial distributed data. Let C be the classes set, N be the dictionary size. Basing on Bayes' rule, each of test documents D_i is assigned to the highest probability class by MNB algorithm. The probability $\Pr(C_i|\, D_i)$ is given below:

$$\Pr(C_i|D_i) = \frac{\Pr(C_i)\,.\,\Pr\,(D_i|C_i)}{\Pr\,(D_i)} \quad C_i \in C \quad (6)$$

Estimating the class prior Pr(Ci) is implemented by a division between the total of documents belonging to class C_i and the total number of documents. Let n be a specific word, f_{ni} be the word count n in test document D_i, and $\Pr(w_n \mid C_i)$ be the probability of word n given class C_i, appearance probability of Di

document in class C, Pr(Di | Ci) aka, is calculated as:

$$\Pr(D_i|C_i) = \left(\sum_n fni\right)! \prod_n \frac{\Pr(W_n|C_i) f_{ni}}{f_{ni}!} \quad (7)$$

5. ANALYZING THE PRODUCE PRICE

Our experiment conducted in 2 main phases in Figure 1. (1) The data were collected from the Shopee e-commerce site through a series of cleaning steps, also known as pre-processing of stop words and special characters. The information features extracted from products are the result of the next step after the data cleaning. The data features are used as the input of the four classification algorithms mentioned above to retrieve data classifications. (2) We build an evaluation application and run on the testing set of the same source above for evaluation activities and make predictions about the selling mis-price of the company's products announced previously.

After the program has identified the product, the program will analyze the price of the product and suggest a product list that violates the price of the company policy. The algorithm defined formally as follows:

Algorithm1: The algorithm analyzes the price of the product and suggest a product list that violates the price of the company policy

Input:

pi: The public price of a shop's products

pj: The public price of a company's products

t: threshold discount allows selling on Shopee analysis:

1: cki = 1 - (pi /pj)

2: if cki > t:

Add to products list that violates the price of the company policy

6. RESULTS AND DISCUSSION

The purpose in this research is aim to identify (the same) products that sell at different prices. For this aim, we modeled and evaluated the proposed methods basing on accuracy of the classifier models (KNN, SVM, RF, MNB) on the two datasets

Collected from the Shopee site.

We implemented these classified models in Python using the library Scikit-learn. We spitted the dataset randomly into train-set (75%) and test-set (25%). Each machine learning algorithm was experienced ten epochs on each dataset, then took their averages to evaluate each model. All experiments were conducted in the same setting, on a machine configuration of Intel (R) Core (TM) i7-7500U, 8GB RAM, and Windows 10 Pro 64bit.

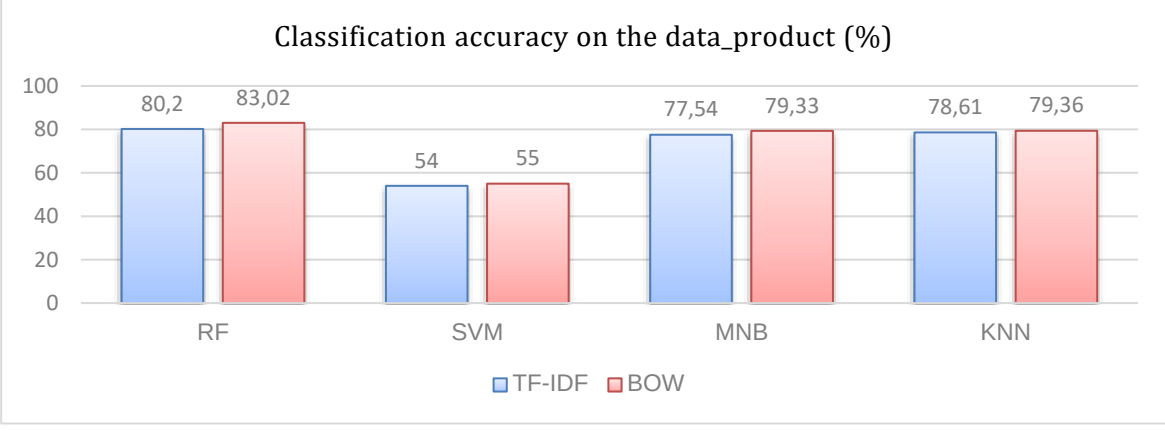


Figure 7. Classification accuracy on the data_product

We compare the performance classification of a model using two text vectorization techniques: BoW and TF-IDF. According to the results (Figure 7), the average classification accuracy for the data_product dataset using BOW is: RF - 83.02%, SVM - 55%, MNB - 79.33%, and KNN - 79.36%. Similarly, for the TF-IDF technique, the correctness of classification for RF is 80.02%, SVM is 54%, MNB is 77.54%, and KNN is 78.61%. The results show the RF model achieves the highest accuracy for both techniques, with 83.02% accuracy using BOW and 80.02% accuracy using TF-IDF.

The classification accuracy on the data_filter dataset (Figure 8). The results show the RF model achieves

the highest accuracy for both techniques, with 98.06% accuracy using BOW and 96.58% accuracy using TF-IDF.

The results have also emphasized that RF model and BoW method are effective choices for building classifier models on the two datasets. The results suggest that the BoW method performs slightly better than TF-IDF for these particular datasets. However, the difference in accuracy between the two techniques is not significant. In order words, RF and BoW methods are quite suitable to problems of identifying and predicting products that sell incorrect prices of the company by machine learning model combined price analysis.

Basing on results of the model, companies, distributors can correctly supervise price flows in their support chain into the market. This activity helps to monitor and ensure that the product price reaches the consumer right or not compared with the companies' price announced by the company. It is a basis for them to propose a timely and appropriate supply price management plan. Our experiments process in three cosmetic brands of the 200 distributors which are selling in Shopee e-commerce. The correct prices of these products are announced by the three brands on the three websites (http://cosmeheal.vn,http://mibitiprudente.vn, https://drceutics.vn).

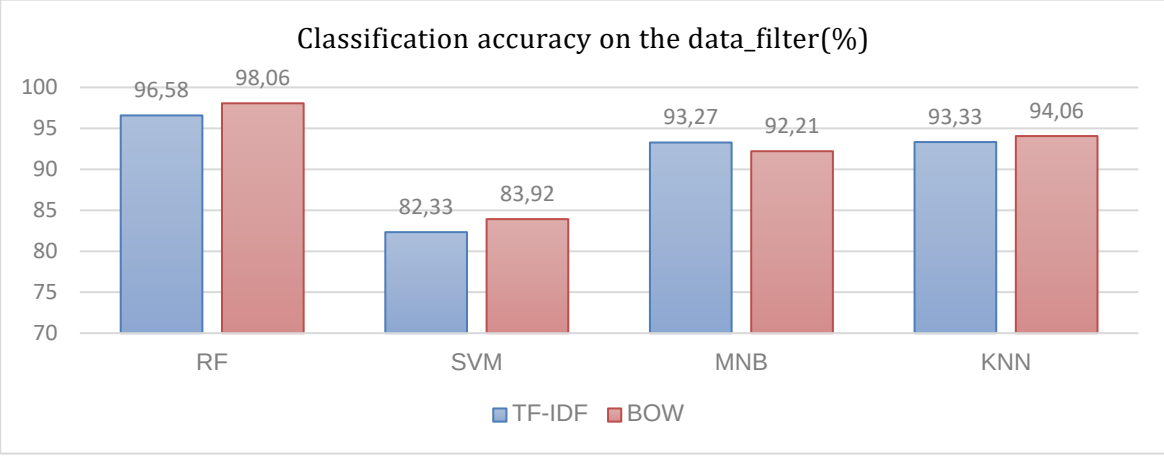


Figure 8. Classification accuracy on the data_filter

7. CONCLUSION AND FUTURE WORKS

We started our paper with a practical problem for companies that have multiple online distributors. While companies want to set consistent price strategies, online distributors and retailers may set different prices and discounts, which confuse customers and may harm the companies in the long run. We align with the work of (Park et al., 2015) in contributing a housing price prediction method based on machine learning algorithm included C4.5, RIPPER, Naïve Bayesian, and AdaBoost and comparing their classification accuracy performance. Performances of the machine learning models Support Vector Machines, Random Forest, Gradient Boosting, AdaBoost are also experimented by (Kohli et al., 2019) to predict the behavior of stock price and influence the stock trend. Since, we use these studies as the basis for choosing our problem solving method through machine learning model.

We extend the work by (Kohli et al., 2019) by conducting with SVM, KNN, RF, and MBN methods to identify and predict products that sell at incorrect prices to the three cosmetics companies. We conduct a comparative study of machine learning algorithms included SVM, KNN, RF, MBN and 2 text-based information extraction methods BoW and TF-IDF to detect the best method for our problem. Algorithms are tested on data_filter and data_product datasets. Basing on the results of the assessment of accuracy from the models, we make the assertion that RF model and BoW method are an effective choice for building an application to identify and predict products that sell at incorrect prices of the company by machine learning model combined price analysis. Experimental results also show that the text datasets in Shopee site is input effectively on BoW classification technique and RF machine learning model. This leads to the average accuracy being more superior than to the rest of the solutions.

We contribute to the theory of (Mahesh, 2020) by proposing an experience in raw data obtained real e-market business. Specially, macro factors such as commodity price, market history, and announced price of companies are vital input factors to predict whether retailers, distributors sell at the wrong or not in Shopee ecommerce through machine learning model combined price analysis. The results provide the following contributions significantly. First, we first use a real dataset in circulation. Specifically, this dataset is collected from e-commerce platform Shopee, which divided into two random subsets. We train in popular machine learning models to solve the problem being studied by the product suppliers themselves. Second, the accuracy of experiments serves as the basis for selecting and evaluating the type of machine learning algorithm suitable for the data set that we are studying. Last but not least, this result supports a basis for companies to suggest an optimal strategy for price fluctuations in the market and to plan to more effectively control the product life cycle through circulating prices.

In the future, we plan to perform the combination of image and text features to improve the accuracy of machine learning models. Further, we plan to experiment by incorporating more images and other data characteristics, such as production date, lifetime or product performance. These features will help to improve data classification to increase the accuracy of the machine learning algorithm in the incorrect product price prediction in the supply chain that manufacturers and companies have announced.

ACKNOWLEDGMENT

This work has received support from the Duoc Si Tien Limited Liability Company.

REFERENCES

Breiman, L. (1998). Arcing classifier (with discussion and a rejoinder by the author). *The annals of statistics*, *26*(3), 801-849.

Agrawal, M., Khan, A. U., & Shukla, P. K. (2019). Stock price prediction using technical indicators: a predictive model using optimal deep learning. *Learning*, *6*(2), 7.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Agustina, D. A., Subanti, S., & Zukhronah, E. (2021). Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter Terhadap Marketplace di Indonesia Menggunakan Algoritma Support Vector Machine. *Indonesian Journal of Applied Statistics*, *3*(2), 109-122.

Fix, E., & Hodges Jr, J. L. (1952). *Discriminatory analysis-nonparametric discrimination: Small sample performance*. California Univ Berkeley.

Gielens, K., & Steenkamp, J. B. E. (2019). Branding in the era of digital (dis) intermediation. *International Journal of Research in Marketing*, *36*(3), 367-384.

Hamad, H., Elbeltagi, I., & El-Gohary, H. (2018). An empirical investigation of business-to-business e-commerce adoption and its impact on SMEs competitive advantage: The case of Egyptian manufacturing SMEs. *Strategic Change*, *27*(3), 209-229.

Holsapple, C. W., & Singh, M. (2000). Electronic commerce: from a definitional taxonomy toward a knowledge-management view. *journal of Organizational computing and Electronic Commerce*, *10*(3), 149-170.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Wadsworth Int. *Group*, *37*(15), 237-251.

Loesche, W. J. (1994). Periodontal disease as a risk factor for heart disease. *Compendium (Newtown, Pa.)*, *15*(8), 976-978.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR). [Internet]*, *9*, 381-386.

Muljono, M., Artanti, D. P., Syukur, A., & Prihandono, A. (2018). Analisa Sentimen Untuk Penilaian Pelayanan Situs Belanja Online Menggunakan Algoritma Naïve Bayes. Konferensi Nasional Sistem Informasi (KNSI) 2018.

Van, T. P., & Thanh, T. M. (2017, November). Vietnamese news classification based on BoW with keywords extraction and neural network. In 2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES) (pp. 43-48). IEEE.

Quinlan, J. R. (1993). C4. 5 Programs for Machine Learning. Morgan Kaufmann, San Mateo, California.

Shalehanny, S., Triayudi, A., & Handayani, E. T. E. (2021). Public's sentiment analysis on shopee-food service using lexicon-based and support vector machine. *Jurnal Riset Informatika*, *4*(1), 1-8.

Joachims, T. (1998, April). Text categorization with support vector machines: Learning with many relevant features. *In European conference on machine learning* (pp. 137-142). Springer, Berlin, Heidelberg.

Tan, F. T. (2019). Realising platform operational agility through information technology–enabled

capabilities: A resource-interdependence perspective. *Information Systems Journal, 3*(29), 582–608.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

Luo, Y., Zhao, S., Li, X., Han, Y., & Ding, Y. (2016). Text keyword extraction method based on word frequency statistics. *Journal of computer applications, 36*(3), 718.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences, 55*(1), 119-139.

Negandhi, P., Trivedi, Y., & Mangrulkar, R. (2019). Intrusion detection system using random forest on the NSL-KDD dataset. *In Emerging Research in Computing, Information, Communication and Applications* (pp. 519-531). Springer, Singapore.

Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2017, July). Deep IV: A flexible approach for counterfactual prediction. *In International Conference on Machine Learning* (pp. 1414-1423). PMLR.

Nassar, L., Okwuchi, I. E., Saad, M., Karray, F., & Ponnambalam, K. (2020, July). Deep learning based approach for fresh produce market price prediction. *In 2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE.

Imran, I., Zaman, U., Waqar, M., & Zaman, A. (2021). Using machine learning algorithms for housing price prediction: the case of Islamabad housing data. *Soft Computing and Machine Intelligence, 1*(1), 11-23.

Lamon, C., Nielsen, E., & Redondo, E. (2017). Cryptocurrency price prediction using news and social media sentiment. *SMU Data Sci. Rev, 1*(3), 1-22

An, J. (2019). Oil price predictors: Machine learning approach. 670216917.

Good, I. J. (1965). *The estimation of probabilities: An essay on modern bayesian methods*, pp. xi-xii.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 986-996). Springer, Berlin, Heidelberg.

Nguyen, L. (2019). Text classification based on support vector machine. *Dalat university journal of science, 9*(2), 3-19.

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications, 42*, 2928–2934. Advance online publication. https://doi.org/10.1016/j.eswa.2015.03.005.

Kohli, C., Suri, R., & Kapoor, A. (2015). Will social media kill branding? *Business Horizons, 58*(1), 35–44. https://doi.org/10.1016/j.bushor.2014.08.004

Kohli, P. P. S., Zargar, S., Arora, S., & Gupta, P. (2019). *Stock prediction using machine learning algorithms*. In Applications of Artificial Intelligence Techniques in Engineering (pp. 405-414). Springer, Singapore. https://doi.org/10.1007/978-981-13-1819-1_38

An, J. (2019). *Oil price predictors: Machine learning approach. 670216917*.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

Van-Duyet Le. 2017. stopwords: Vietnamese. https://github.com/stopwords/vietnamese-stopwords.