ISSN 2278-3091
Volume 9, No.1.4, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse9991.42020.pdf
https://doi.org/10.30534/ijatcse/2020/9991.42020

A Mini Review on Supervised Learning Models for E-Commerce Product Classification

Norsyela Muhammad Noor Mathivanan[1], Nor Azura Md. Ghani[2,3], Roziah Mohd Janor[3]
[1]Postgraduate Student, Center for Statistical and Decision Sciences Studies, Faculty of Computer & Mathematical Sciences Universiti Teknologi MARA, Malaysia. syelamohdnoor@gmail.com
[2]Member, National Design Centre, Universiti Teknologi MARA, Malaysia. azura@tmsk.uitm.edu.my
[3]Professor, Center for Statistical and Decision Sciences Studies, Faculty of Computer & Mathematical Sciences Universiti Teknologi MARA, Malaysia. azura158@uitm.edu.my, roziahmj@uitm.edu.my

ABSTRACT
Nowadays, consumers use e-commerce websites to find their desired products. Some researchers and related services collect data from these websites for research purposes. The misclassification of the product may distract them from finding the right products. Thus, product classification is essential in e-commerce websites. Machine learning models are widely used to classify the products according to their categories. This review presents the supervised learning model and its applications in this field. Various articles were analysed on the use of different supervised learning models for product classification. The review provides different types of supervised learning models and its enhancement in dealing with e-commerce products data. The findings are crucial for other researchers and practitioners that can be used to improve their current supervised learning model for e-commerce product classification.

Key words: E-Commerce**,** Supervised Learning Model, Product Classification, Review

1. INTRODUCTION

Online commerce has rapidly grown since the past decade. Goods can be purchased not only from physical stores but also via online shopping. Consumers are provided with the ease and flexibility of shopping as they only have to search for products using specific keywords and know the product's availability. There are millions of products on e-commerce websites such as Amazon, e-Bay, 11street, and Lazada that are participated by thousands of sellers.

Furthermore, many new products are registered on these websites on a daily basis. The key component of the success of these websites is its quick and accurate retrieval of the desired products [1]. Each product is presented by metadata such as title, description, category, image, and price that are mostly assigned manually by the sellers. Unlike the title and price, it is possible to classify the product categories from the metadata. The automatic product categorisation can reduce time and costs besides improving the accuracy of categorisation of the same product by different sellers [2]. These are the reasons that categorising products become the key issue in e-commerce. Product classification is defined as a text classification with large product taxonomy [3]. It is a classic topic for natural language processing where predefined categories are assigned to text inputs with machine learning techniques. The classification is based on significant words or features extracted from the text, such as the title and descriptions of the products.

In particular, product classification has three main issues as follows: 1) the products are sparsely distributed in a large number of categories and the data distribution is quite skewed; 2) the product titles and descriptions have different length; and 3) there is a possibility that the available pairs of the current product title and assigned category are incorrect [2]. Researchers had conducted studies to overcome these problems and provide a good product classification model using different methodologies. The product classification is based on supervised learning models that are commonly used to classify products using either text or image data [4]. Most of the studies implement several important steps, such as data pre-processing, feature extraction and selection, and model selection. Therefore, it is crucial to provide a comprehensive review of all the steps in accordance with previous studies.

2. FRAMEWORK FOR E-COMMERCE PRODUCT CLASSIFICATION

Figure 1 shows the common overall framework for product classification. The framework is constructed based on the review of previous literature and a general framework regarding product classification. There are several processes before classifying the products. First, researchers have to deal with the data and feature that are used in the study. The data have to be extracted from their source and undergo several crucial tasks in the data pre-processing step. Then, the feature selection technique reduces its dimensionality. There is a need to find an efficient technique to deal with the increasing amount of large text data sets. An optimal number of attributes or features are required to classify any text document as the preliminary condition. There are two ways in

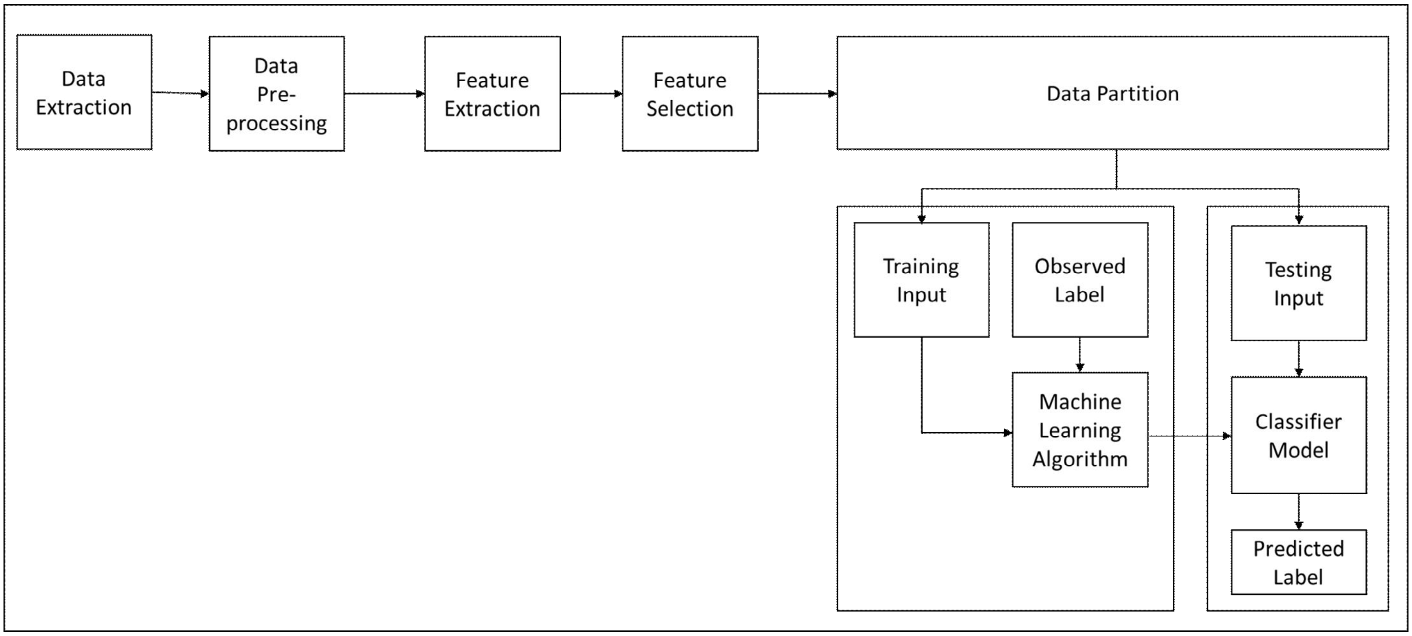

Figure 1: Framework for E-Commerce Product Classification

Fulfilling this requirement, which are feature extraction and selection. In feature extraction, researchers can identify a new set of feature space that is more than the original feature space [5]. On the other hand, researchers can select a subset of the original feature set in feature selection [6]. Both techniques provide significant impact in increasing the accuracy of a classification model and its efficiency of the processing time. The classification model has to be trained properly to classify products into the right category. Hence, researchers proposed and applied various machine learning models to classify the products.

2.1 Data Extraction

Data extraction can be defined as the process of retrieving data from any sources and transform them into a more useful format for further processing. This process does not refer to the processing or analysis that may happen after obtaining the data. Normally, data that were extracted for product classification originated from online store websites, where the data were obtained from the website pages. Web data extraction takes almost unfeasible human time and effort in collecting data [7] for ease of data retrieval. There are two aspects that become the main concerns in dealing with product classification, which are the data source and their format. There are two data sources used in previous studies. First, they used confidential data that were extracted from e-commerce websites or permissions to use data extracted by third party companies [8-18]. Studies that used this kind of data extraction intended to secure their data from being wrongly used by others. However, there were researchers who used publicly accessible e-commerce products data sets to access the performance of supervised learning models on the data sets. This second type of data source makes it possible for researchers to explore suitable models to deal with product classification, and they can compare their results with previous studies that use the same data sets. Meanwhile, researchers need to know the format of the data before classifying the products. There are two common forms of data utilised by previous researchers, which are text and image data sets. Out of these two data formats, text data is widely used for product classification rather than image data [4]. The former data format needs less storage and time to be extracted compared to the latter data format. However, some researchers prefer to use image data for its high dimensionality as features with limited content analysis, heterogeneity, and other nuance factors in image-based classification performance [19-21]. Table 1 shows a part of the open-access data sets used by previous researchers to perform their product classification. For the next steps, this study focused on explaining the use of text data to classify the product.

2.2 Data Pre-Processing

Data pre-processing is a crucial step in dealing with product classification. The main objective of this step is to use suitable techniques in transforming original textual data into an understandable format. Pre-processing is crucial to maintain the retrieval performance and improve the accuracy of the model. The space to store the document and time requirement to process the data can be efficiently decreased with this process. It is a complex process for the representation of the feature extracted from the textual input. The extraction of key features or key terms can enhance the relevancy of word on the category and document used in the study. Therefore, pre-processing is important to prevent additional problems in classifying the products.

Table 1: Accessible Data Sets used in Previous Studies

Data Source	Data Type	Related Work
Rakuten Data Challenge (RDC) This data set is taken from www.rakuten.com that contains 800,000 product titles in English with their respective multi-level category labels.	Text	[2, 22, 23]
Open Product Data (OPD) This data set is taken from product-open-data.com that contains over 900,000 products and their associated brands, which can be entirely downloaded.	Text	[24]
Product databases of Amazon, Flipkart, Snapdeal, and Paytm These data sets are taken from https://github.com/sam-chirag/Data-Classification-Using-Machine-Learning-Dataset that contains 40,000 products with 1,000 leaf classes.	Text	[25]
Amazon Review Data This data set contains 233.1 million reviews and metadata with 29 pre-categories. It is provided based on the per-category files requested by interested researchers.	Text	[26, 27]
PI100 This data set is taken from http://research.microsoft.com/users/xingx/PI100.aspx that contains 12,000 images, and they are evenly distributed in 100 categories from MSN shopping website.	Image	[4, 28 – 31]
Fashion-MNIST This data set is taken from https://github.com/zalandoresearch/fashion-mnist that contains 70, 000 unique products based on the assortment in Zalando's website.	Image	[32, 33]

A. General Processes

There are several processes involved in pre-processing, such as data cleaning, data integration, data transformation, data reduction, and data discretisation. Data cleaning is the initial process that solves problems such as missing values, inconsistent and noisy data, and detecting error and outliers in the raw data. Data integration combines data from different sources into one data store. Data transformation has two steps: data normalisation and aggregation. The data collected for product classification tend to have a skewed distribution, and researchers usually normalise the data to fit within the range [34]. The next process is data reduction in which the volume of the data is reduced without changing the end result of the analysis. Data discretisation replaces the numerical attributes with nominal ones to reduce possible values of continuous that can contribute to a slow and ineffective process of machine learning [35]. It is not necessary to perform all the steps in pre-processing because the necessity depends on the complexity and requirement to transform the data into the data-mining-ready structure.

In general, unstructured data such as text and image are more complex than structured data. It is hard to classify the contents due to their complexity. Text data is one of the simplest forms of data that can be generated from various sources [36]. It is created in free-form styles, and it is a difficult and tedious task to find the attributes to describe text data. It is easy for a human to process and perceive text data; however, machines have difficulty to understand it. The analysis and reports on such data cannot be generated due to the lack of structure and dominant characteristics to represent each item. Hence, the process requires higher processing time to mine the data. This is the reason that pre-processing can improve the effectiveness of the mining algorithm, especially for text mining.

B. Crucial Tasks

For text data, the pre-processing step in text mining consists of three tasks: tokenisation, stop word removal, and stemming [37]. Most studies applied these steps to ensure that the data are well-prepared before proceeding with text classification [38]. Tokenisation divides the whole statement into words by removing spaces, commas, and the period from the metadata. The purpose of tokenisation is to identify meaningful keywords from the textual data [39]. Stop words are used to make the text looks heavier and less crucial for analysis, and they are removed to reduce the dimensionality of term space. Several common words are treated as stop words such as articles, pronouns, and prepositions that are not measured as keywords in text mining applications [40].

For stemming, it obtains the root or stem of derived words by removing the common suffixes and reducing the number of words to match the stems accurately. For example, words like attraction, attracted, attracting, and attracts all are from the word "attract". Over the years, many stemming algorithms have been developed to provide services from different domains. Each algorithm has a common way of finding the stems of the word variants. The most popular stemming algorithm for text document in English is Porter's stemmer [36]. This stemmer is a popular algorithm by Porter (1980), which is a suffix stripping sequence of systematic steps for stemming an English word. It can reduce the vocabulary of the training text by approximately one-third of its original size [38].

Previous studies on product classification emphasised the importance of pre-processing in managing the data before extracting the features [8]. The process begins with tokenisation of product titles and elimination of punctuations. In addition, a study found that numbers are used to differentiate between single and wholesale categories, and prepositions are used to judge product and accessory categories [41]. On the other hand, the pre-processing step excludes numbers and prepositions. The data sets used for product classification are usually skewed category distribution [34]. Hence, the normalisation step is needed to deal with the data. Most of the product classifications are under text mining application; however, the steps that are required in the pre-processing depend on the structure and behaviour of the data.

2.3 Feature Extraction

A. Common Approaches

In feature extraction, the original features are replaced with a smaller representative set without deleting them to reduce the space. This technique is often used when dealing with a large number of features in input data that affects the processing time [42]. It can reduce a huge amount of memory and computation as a result. Several techniques are used in extracting the feature such as Bag-of-Word, n-grams, Term Frequency-Inverse Document Frequency (TF-IDF), cosine similarity, Jaccard similarity, Levenshtein distance, and feature hashing. N-grams and TF-IDF are commonly used by researchers to extract text features [10]. The basic tool for feature generation is Bag-of-Word model that transforms text data into a "bag of words" where the most similar type of characteristics or features are calculated in the form of term frequency. However, this technique only considers the counts of words and ignores any spatial information from the data. This technique causes the loss of word, which is particularly problematic in text classification [43]. Therefore, the n-gram model is used as an alternative to capture the information within the text.

Bag-of-Word model can be treated as a special case of the n-gram model where the n value equals to one. The purpose of n-gram model is to observe the difference in information from the model using one or more words. If there is only one word such as pen, cup, and umbrella in the document corpus, it is known as 1-gram. Meanwhile, 2-gram or bigram model describes the text according to the following units and stores the term frequency of each unit as previous. Some of the examples for the features of the bigram model are red pen, big cup, and colourful umbrella. Normally, the maximum is 3-grams, and this model depends on the structure of the text document used in the study. The standard weighting measurement for this technique is term frequency (TF); however, researchers can use the term, frequency-inverse document frequency (TF-IDF), to judge the class of the text documents based on the count of words and content. According to the basic concept of TF-IDF, a learning algorithm gets more information from the rarely occurring terms than frequently occurring terms from a data set. It helps to score the importance of terms using a weighted scheme. This goal can be achieved because TF-IDF gives weight to each word that is represented as a feature. The word counts are replaced with the scores obtained from the document corpus. Hence, TF-IDF is a useful measurement to extract representative features for product classification.

B. Approaches used for Product Classification

Feature extraction plays an important role in product classification using text data. Table 2 summarises the various feature extraction techniques used for product classifying purpose. Among the various research attempts, n-grams model is often used to extract data. The preferable models are unigram and bigram models because the data are in the form of product title or description that does not contain lengthy sentences. Hence, both models are useful for product classification based on text data.

Table 2: Feature Extraction Techniques for Product Classification

Feature Extraction	Related Work
Unigram	[1, 2, 4, 8 – 10, 14, 20, 25, 41, 42, 44, 45, 48 - 50]
Bigram	[20, 25, 42, 44, 51]
n-gram	[15, 22 – 24]
Skip-gram	[10, 17]
TF-IDF	[12, 17, 25, 44, 51, 52]
Others	[1, 12, 42]

2.4 Feature Selection

In feature selection, the dimensionality of original features extracted from the data is reduced by selecting a subset of original features. Then, the remaining features are mapped into a new feature space.

A. Common Approaches

There are three approaches in performing feature selection: filter, wrapper, and embedded. The commonly used approach regardless of fields is filter approach [53]. This approach finds the relevance feature index by estimating the relevance of a feature to target categories. Then, the features are ranked according to their importance, and an action is done on irrelevant features. The action includes measures like correlation, mutual information, and entropy to analyse general data characteristics in choosing the optimal features for the data set. Researchers and practitioners prefer this approach because the related measurements are simpler and faster compared to the other two feature selection approaches [54]. The limitation of this approach is that the calculation index is based on a single feature where the interaction between features is often neglected. This limitation can lead to the discard of the best pair. Hence, the multivariate filter approach is introduced to overcome the problem of dependencies; however, the time complexity is increased compared to the univariate approach [55]. Several techniques can be categorised under filter approach, such as the Pearson correlation coefficient, chi-square, odd ratio, information gain, and mutual information.

The selection of features using the wrapper approach is conducted by using the prediction from classifier for a given subset. This approach detects similar potential information between features [54]. The choice of classifier is important to obtain a good feature subset because it is a classifier dependent approach. When dealing with text classification, the chosen classifier can handle high dimensionality features. In addition, researchers need to ensure that the classifier can manage noise and multi-label data set. The efficiency of this approach depends on the strategy of selecting the classifier that achieves high accuracy performance with minimal computation time [56]. The crucial problem in overcoming the use of the wrapper approach is the over-fitting issue, especially for large scale data set. There are two types of wrapper feature selection, which are deterministic and randomised. Normally, the learning process in the deterministic way uses the heuristics way. This action is prone to local optimum because it is difficult to determine the upper limit on the number of feature in a subset [57]. There are various techniques under this approach, such as beam search, bidirectional search, sequential forward selection, and backward sequential elimination. On the other hand, the wrapper approach uses the randomised technique to find the subsets with some kind of randomness using the Las Vegas wrapper, genetic algorithm, or Monte Carlo models [58].

The embedded approach is similar to the wrapper approach. However, this approach has two elements in a single algorithm, which are either filter or wrapper approach and a classifier. Specifically, it applies the wrapper approach on features that are chosen by the filter approach [59]. It becomes a part of the objective function from the classification algorithm itself. The embedded approach can reduce the computational time for feature selection compared to the wrapper approach. Several combinations that can be used under this approach are Odd Ratios + SVM-RFE, Support Vector Machine with Recursive Feature Elimination, and Weighted Naïve Bayes [55]. All feature selection approaches have their strengths and weaknesses. Nonetheless, it is concluded that embedded or hybrid feature selection approach is the most beneficial approach among the other approaches because it can overcome the limitations in filter and wrapper approaches.

B. Approaches used for Product Classification

Feature selection plays an important role in product classification using text data. Table 3 summarises the various feature selection techniques used for product classification. Most studies used filter approach compared to the others. It is easier to implement this approach compared to other approaches. The preferable techniques include correlation-based and information gain.

Table 3: Feature Selection Techniques for Product Classification

Feature Selection	Related Work
Correlation-based	[9, 46, 48, 49]
Information Gain	[14, 41, 50, 51]
Mutual Information	[10, 50]
Chi-Square	[25, 50]
PCA	[52, 60, 61]
K-Means	[9, 60]

2.5 Data Partition

The supervised machine learning model is used to build a model to assign data to their predefined classes. The model is normally tested on independent data set, and the prediction accuracy provides the information about the performance of the classifier [62]. Hence, researchers have to split the overall data into multiple chunks such as training, validating, and testing parts in developing the model. However, many studies for product classification classified their data set into two chunks, which are training and testing parts [1, 2, 4, 9, 10, 12, 14 – 16, 20, 23, 24, 32, 33, 41, 42, 44 – 46, 48 – 51, 60, 63 – 67]. These researchers used various partition ratios that randomly split the data into approximately 70% and 30% for training and testing parts, respectively [68]. Some researchers mention that the validation part is important to avoid model overfitting [2, 3, 8, 11, 17, 18, 22]. It is necessary to have a validation data set as well as the training and test data sets when adjusting any classification parameter [69].

Training and validation are usually conducted with cross-validation to find the best parameters for a classifier [70]. It is a resampling procedure to evaluate supervised learning models on a limited data sample. The procedure consists of a single parameter known as k, where it defines the number of groups to be split into a given data sample. K-fold cross-validation uses various k values, but there is a bias-variance trade-off related to the chosen value [71]. Thus, researchers often use k = 5 or k =10 in performing k-fold cross validation because these values are not affected by the excessively high bias and variance [71]. Previous researchers used this procedure to build their supervised learning model for classifying products [4, 15, 23, 44, 50, 66]. Besides k-fold cross-validation, a bootstrapping technique is also useful in validating supervised learning models for product classification [69].

2.6 Classification

Data classification is conducted using the machine learning model. Machine learning is the study of computational methods to improve the performance by utilising the knowledge from experience [72]. It aims to teach machines to handle data efficiently. Various studies revealed the methods to construct machines that can learn by themselves [73 – 75]. Machine learning has been applied not only for educational purposes but also for aiding industries to make good decisions based on the availability of their project-related data. Machine learning algorithms can either be supervised or unsupervised. There are several machine learning algorithms that use different approaches, such as semi-supervised and reinforcement learning [76]. However, supervised and unsupervised learning models are the main approaches in applying machine learning model. The difference between these two main models is the label in the training data subset.

Technically, the unsupervised learning model is used to group or cluster unlabelled data. On the hand, the supervised learning model is used to classify data with predefined categories. When performing classification, researchers use algorithms under supervised learning models; some of them use unsupervised models to support and enhance their supervised learning model performance. When a supervised learning algorithm achieves an acceptable performance level, the learning process will stop. It performs an analytical task using the training data before constructing the contingent functions to map new attribute instances [29]. The algorithms need sufficient pre-specifications in describing data behaviour. Then, they can be used to produce the desired outcome and high-level performance model.

Algorithms related to supervised learning model can predict and classify the predetermined attribute. The model efficiency can be evaluated using a range of performance metrics such as accuracy, precision, recall, F1-score, and Area under the ROC Curve (AUC). All of these evaluation measures have their own technical interpretation and well-defined relationship; they can be linked to each other according to certain use of cases. For product classification, most studies used accuracy as the main evaluator for their supervised learning models [2 – 4, 9 – 11, 14, 32, 33, 41, 46, 49 – 51, 60, 63 – 65, 67]. Most of them supported their results with good precision and recall values besides accuracy to evaluate the performance. Several studies argued on the limitation of predictive accuracy as the main standard in classification [45]. There were researchers that used other measurements such as F1-score [1, 10, 12, 16 – 18, 22 – 25, 47, 48], Area Under Receiver Operating Characteristic Curve (AUROC) [15], and error analysis [3, 4, 44]. There was a study that proposed a new approach for performance measurement or also known as average revenue loss to evaluate product classification model [45].

3. SUPERVISED LEARNING MODELS APPLICATION ON PRODUCT CLASSIFICATION

Product classification can be seen as a supervised classification problem where the target classes are the product categories, and the features are extracted from product information such as title, description, or image. Figure 2 shows the supervised learning models used in common models by previous researchers without any improvisation. These models were the benchmark models to compare the proposed models.

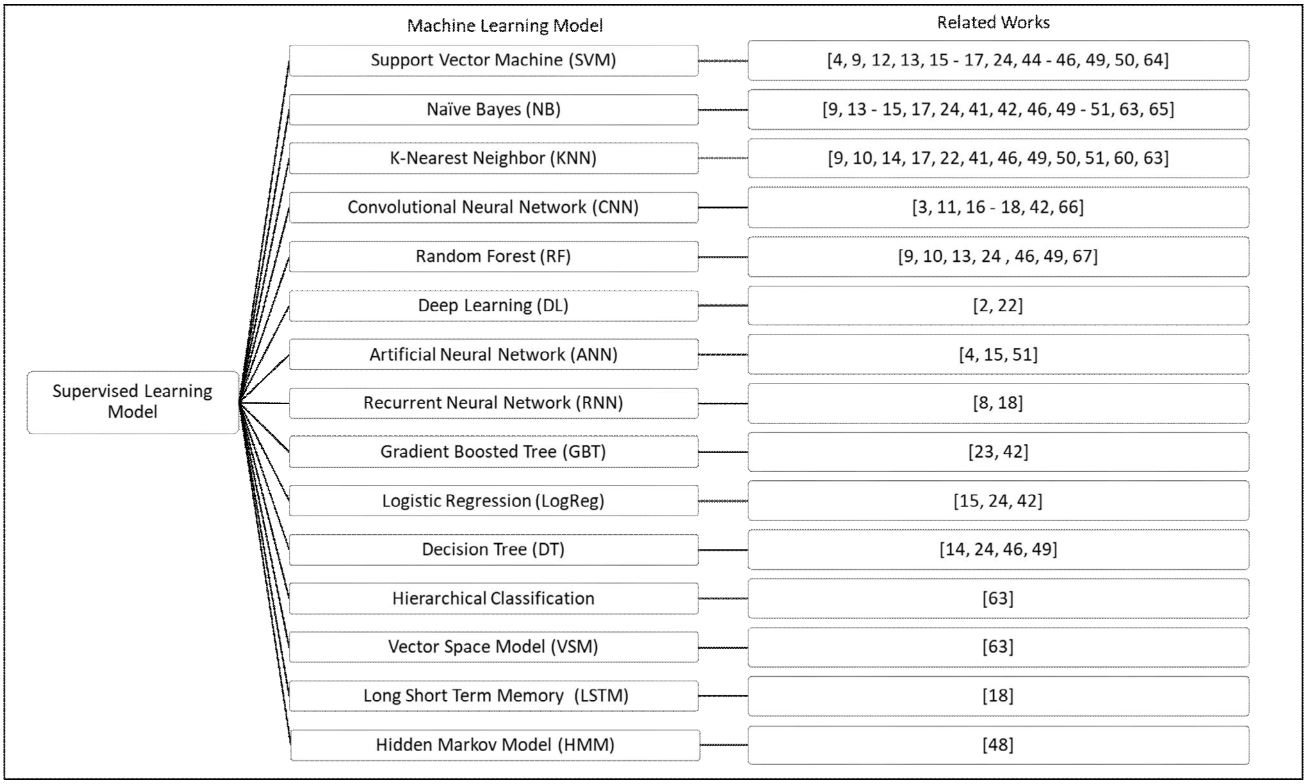

Figure 2: Product Classification based on Machine Learning Models

There is a considerable body of literature on product classification. Most studies used text data in classifying products. Besides, there were initiatives that used image data [32, 33]. Some researchers argued that one source of data is insufficient to provide a good classification model [3, 20, 24]. Hence, they used both text and image data sets to build their classification models. They had to collect the data by themselves because most of the data were available in either one of the forms. Figure 2 shows the supervised learning models used to classify various e-commerce products. Most studies used the support vector machine (SVM) and naïve Bayes (NB) models. Meanwhile, the convolutional neural network (CNN) was commonly used when researchers dealt with product images as the features.

A recent study revealed that a neural machine translation was a better model to classify e-commerce data from Rakuten Data Challenge (RDC) and Rakuten Ichiba [69]. This classification model works by translating a natural language for product into token sequences that constitute a root-to-leaf path in product taxonomy. The model was compared to the Deep Belief Network (DBN), K-Nearest Neighbor (KNN), and the fusion of these two models. Besides, another study proposed a model known as Attention CNN [23]. This model was proven superior compared to (Gradient Boosted Tree) GBT model. Additionally, a fusion of KNN and SVM was proposed to classify products from eBay.com [41]. The benchmark models for the study were NB and KNN where the performance of both models was lower than the proposed model. Previous researchers revealed that the current models might not be sufficient to classify products, especially large scale data sets. Besides, researchers often faced difficulties when applying the conventional procedures on product title classification [44]. They used many combinations to increase the accuracy of the classification model. Recent studies showed that the combination of classification and clustering models could provide better classification result [77, 78]. The clustering algorithm is often used to group unlabelled data into a homogeneous group based on the selected features. It is a crucial tool to solve unsupervised learning problems. The purpose of the clustering algorithm is to classify the data into groups that share almost similar features. There are some properties that should be fulfilled to ensure that the performance of a clustering algorithm can deal with noise and uncertainty besides data with high dimension and different types of attributes.

4. CONCLUSION

This paper provides a review of various supervised learning models that are a part of machine learning to classify e-commerce products. First, all steps for product classification were studied. Pre-processing is a crucial step to manage and clean the raw data set before extracting the features for the input of the classification models. Several feature extraction and selection techniques were used for product classification to provide good features in developing the models.

Then, various supervised learning models that were currently used in product classification were studied. Several studies described the issues. Based on the issues, this study listed the common models used to solve product classification problems. In addition, this study found some improvisation on supervised learning models, specifically for product classification. The summarisation of improvised models can provide insights on an efficient and effective classification model for product classification.

ACKNOWLEDGEMENT

The research is financially supported by the University Teknologi MARA and Ministry of Education Malaysia under the Grant Scheme (600-IRMI/FRGS 5/3 (120/ 2019)).

REFERENCES

1. Z. Kozareva. **Everyone likes shopping! multi-class product categorization for e commerce**. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 168, pp. 1329–1333, 2015.
2. A. Cevahir, and K. Murakami. **Large-scale multi-class and hierarchical product categorization for an e-commerce giant.** In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 525–535, 2016.
3. T. Zahavy, A. Krishnan, A. Magnani, S. Mannor. **Is a picture worth a thousand words? A Deep Multi-Modal Fusion Architecture for Product Classification in e-commerce**, The Thirtieth AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI-18), pp. 7873–7870, 2018.
4. S. A. Oyewole and O. O. Olugbara. **Product image classification using eigen colour feature with ensemble machine learning**. Egyptian informatics Journal, vol. 19, no. 2, pp. 83-100, July 2018.
5. M. Abdullah, A. Agal, M. Alharthi, M. Alrashidi. **Arabic Handwriting Recognition Model based on Neural Network Approach**. International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), vol.8, no.1.1, pp. 253-258, 2019.
6. D. Saxena, S. K. Saritha and K. N. Prasad. **Survey paper on feature extraction methods in text categorization**. International Journal of Computer Applications, 166(11), pp. 11–17, 2017.
7. M. Syamala, N. J. Nalini. **A Deep Analysis on Aspect based Sentiment Text Classification Approaches**. International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), vol.8, no.5, pp. 1795-1801, 2019.
8. J. W. Ha, H. Pyo, and J. Kim. **Large-scale item categorization in e-commerce using multiple recurrent neural networks**. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 107–115, 2016.
9. N. M. N. Mathivanan, N. A. M. Ghani, and R. M. Janor, **Improving Classification Accuracy Using Clustering Technique**. Bulletin of Electrical Engineering and Informatics. vol. 7, no. 3, pp. 465–470, 2018.
10. V. Gupta, H. Karnick, A. Bansal, P. Jhala. **Product classification in ecommerce using distributional semantics**. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016.
11. R. K. Khanuja. Optimizing e-commerce product classification using transfer learning. Master Thesis, San Jose State University 679.
12. G. Moiseev. **Classification of e-commerce websites by product categories**. In AIST (Supplement). 237–247, 2016.
13. I. Partalas, and G. Balikas. **E-commerce product classification: our participation at cdiscount 2015 challenge**. ArXiv, abs/1606.02854, 2016.
14. S. Shankar, I. Lin. **Applying machine learning to product categorization**. Technical report, Stanford University, 2011.
15. C. Chavaltada, K. Pasupa and D. R. Hardoon. **A comparative study of machine learning techniques for automatic product categorisation**. Lecture Notes in Computer Science, 10–17, 2017.
16. V. Umaashankar, S. GirishShanmugam, and A. Prakash **Atlas: A dataset and benchmark for e-commerce clothing product categorization**. ArXiv, abs/1908. 08984. 2019.
17. L. Yang, Y. Yang, H. Yu, and G. Zhu. 2019. **Anonymous market product classification based on deep learning**. In Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing (AIIPCC '19). Association for Computing Machinery, New York, NY, USA, Article 17, 1–5, 2019.
18. W. Yu, Z. Sun, H. Liu, Z. Li, and Z. Zheng. **Multi-level deep learning based e-commerce product categorization**. In SIGIR 2018 Workshop on e-commerce. 2018.
19. Y. Han, S. M. Choi. **A content recommendation system based on category correlations**. Fifth international multi-conference on computing in the global information technology; p. 66–70, 2010.
20. A. Kannan, P. P. Talukdar, N. Rasiwasia, and Q. Ke. **Improving product classification using images**. In ICDM, pp. 310–319, 2011.
21. S. A. Oyewole, O. O. Olugbara, E. Adetiba, T. Nepal. **Classification of product images in different color models with customised kernel for support vector machine**. In: Third international conference on artificial intelligence, modelling and simulation. pp. 153–7, 2015.
22. M. Y. Li, S. Kok, L. Tan. **Don't classify, translate: Multi-level ecommerce product categorization via machine translation**. arXiv preprint arXiv:1812.05774, 2018.

23. Y. Xia, A. Levine, P. Das, G. Di Fabbrizio, K. Shinzato, and A. Datta. **Large-scale categorization of japanese product titles using neural attention models**. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Association for Computational Linguistics. 2017.
24. M. Karlsson and A. Karlstedt. **Product classification-a hierarchical approach**. LU-CSEX 2016-31, 2016.
25. V. Nair, R. Malhotra, N. Maknoor, S. K. Mohaptra. **A machine learning algorithm for product classification based on unstructured text description**. International Journal of Engineering Research & Technology (IJERT) Volume 07, Issue 06, June 2018.
26. J. Ni, J. Li and J. McAuley. **Empirical Methods in Natural Language Processing**. (EMNLP), 2019.
27. R. He and J. McAuley. **Ups and downs: Modeling the visual evolution of fashion trends with oneclass collaborative filtering**. In Proceedings of the WWW. 507–517, 2016.
28. X. Xie, L. Lu, M. Jia, F. Seide, and M. Y. Ma. **Mobile search with multimodal queries**. Proc IEEE 2008; 96(4):589–601, 2008.
29. H. Zhang and Z. Sha. **Product classification based on SVM and PHOG descriptor**, IJCSNS. International Journal Computer Science Network Security,13(9):1–4, 2013.
30. S. Jia, X. Kong, H. Fu and G. Jin. **Automatic fast classification of product images with class specific descriptor.** J Electron, 6, pp. 7-10, 2010.
31. B. Tomasik, P. Thiha, D. Turnbull. **Tagging products using image classification**. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval. Boston, MA, USA; pp. 792–3, 2009.
32. H. Xiao, K. Rasul, and R. Vollgraf. **Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms**. arXiv preprint arXiv:1708.07747, 2017.
33. S. Bhatnagar, D. Ghosal and M. H. Kolekar, **Classification of fashion article images using convolutional neural networks**. 2017 Fourth International Conference on Image Information Processing (ICIIP), Shimla, pp. 1-6, 2017.
34. P. Papadimitriou, P. Tsaparas, A. Fuxman and L. Getoor. **TACI: Taxonomy-aware catalog integration**. IEEE Transactions on Knowledge and Data Engineering, 25(7), 1643–1655, 2012.
35. S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas. **Data preprocessing for supervised learning**. International Journal of Computer Science, 1(2), 111–117, 2006.
36. M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez and K. Kochut. **A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques**, 2017.
37. S. Vijayarani, J. Ilamathi, and S. Nithya. **Preprocessing Techniques for Text Mining - An Overview** International Journal of Computer Science & Communication Networks, Vol 5(1),7-16, 2015.
38. K. Dalal and M. A. Zaveri. Automatic Text Classification: A Technical Review. International Journal of Computer Applications, 28(2), 37–40, 2011.
39. S. Vijayarani, and R. Janani. **Text Mining: open Source Tokenization Tools - An Analysis**. Advanced Computa tional Intelligence: An International Journal (ACII), 3(1), 37–47, 2016.
40. M. F. Porter. **An algorithm for suffix stripping**, 14(3), 313–316, 1980.
41. D. Shen, J. D. Ruvini, and B. Sarwar. **Large-scale item categorization for e-commerce**. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM ’12, pages 595–604, New York, NY, USA. ACM, 2012.
42. A. Uthaman, T. Gawade, D. Inamadar and S. Khan. **Feature Extraction using Text mining**. International Journal of Emerging Technology and Computer Science, 1(2), 1–4, 2016.
43. R. Johnson and T. Zhang. **Effective Use of Word Order for Text Categorization with Convolutional Neural Networks**. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 103 – 112, Denver, Colorado. Association for Computational Linguistics.2015.
44. H. F. Yu, C. H. Ho, P. Arunachalam, M. Somaiya, and C. J. Lin. **Product title classification versus text classification**. Technical report, Department of Computer Science, National Taiwan University, Taipei, Taiwan. 2013.
45. J. Chen and D. Warren. **Cost-sensitive learning for large-scale hierarchical classification**. In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 1351–1360, 2013.
46. N. M. N. Mathivanan, N. A. M. Ghani, and R. M. Janor,. **E-Commerce Product Classification Using Supervised Learning Models**. International Journal of Engineering & Technology, vol. 8(1.7), pp. 214-218, 2019.
47. P. Das, Y. Xia, A. Levine, G. Di Fabbrizio and A. Datta. **Large-scale taxonomy categorization for noisy product listings**. In Big Data (Big Data), 2016 IEEE International Conference on, pages 3885–3894. IEEE, 2016.
48. N. M. N. Mathivanan, N. A. M. Ghani, and R. M. Janor, **Text Classification of E-Commerce Product via Hidden Markov Model**. Advancing Technology Industrialization Through Intelligent Software Methodologies, Tools and Techniques, IOS Press, 2019.
49. N. M. N. Mathivanan, N. A. M. Ghani, and R. M. Janor, Performance analysis of supervised learning models for product title classification, IAES International Journal of

Artificial Intelligence (IJ-AI), Vol. 8, No. 3, pp. 299-306, September 2019.

50. D. Vandic, F. Frasincar, and U. Kaymak. **A framework for product description classification in e-commerce**. Journal of Web Engineering 17, 1&2, 001–027, 2018.
51. C. Sun, N. Rampalli, F. Yang, and A. Doan. **Chimera: Large-scale classification using machine learning, rules, and crowdsourcing**. Proceedings of the VLDB Endowment, 7:1529–1540, 2014.
52. N. M. N. Mathivanan, N. A. M. Ghani, and R. M. Janor, **Analysis of K-Means Clustering Algorithm: A Case Study Using Large Scale E-Commerce Products**, In 2019 IEEE Conference on Big Data and Analytics (ICBDA), IEEE, pp. 41-44, 2019.
53. S. Sarkar, S. Goswami, A. Agrwal, J. Aktar. **A Novel Feature Selection Technique for Text Classification Using Naïve Bayes**. International Scholarly Research Notices, vol. 2014, no. 717092, p. 1-10, Apr. 2014
54. G. Kou, P. Yang, F. Xiao, Y. Chen and F. E. Alsaadi. **Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision making methods**. Applied Soft Computing vol. 86, Jan. 2020.
55. Y. Saeys, I. Inza, and P. Larranaga. **A review of feature selection techniques in bioinformatics**. Bioinformatics, vol. 23, no. 19, pp. 2507–2517, 2007.
56. Y. Kuang. **A Comparative Study on Feature Selection Methods and Their Applications in Causal Inference**. Computer, 2009.
57. Y Mao and Y. Yan. **A Wrapper Feature Subset Selection Method Based on Randomized Search and Multilayer Structure**. International Scholarly Research Notices, vol. 2019, no. 9864213, p. 1-9, Nov. 2019.
58. R. Kohavi and G. John. **Wrappers for feature selection**. Artificial Intelligence, 97(1-2):273–324, December 1997.
59. M. B. Imani, M. R. Keyvanpour, and R. Azmi. 2013. **A novel embedded feature selection method: a comparative study in the application of text categorization**. Appl. Artif. Intell. 27, 5, pp. 408–427, May 2013.
60. N. M. N. Mathivanan, N. A. M. Ghani, and R. M. Janor. **A comparative study on dimensionality reduction between principal component analysis and k-means clustering**. Indonesian Journal of Electrical Engineering and Computer Science Vol. 16, No. 2, pp. 752-758, November 2019.
61. V.Saravanan, Sathya Charanya.C. **E-commerce product classification using lexical based hybrid feature extraction and svm**. International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol-9 Issue-1, November 2019.
62. E. Alpaydin. **Introduction to machine learning**. MIT press; 2014 Aug 22.
63. Y. Ding, M. Korotkiy, B. Omelayenko, V. Kartseva, V. Zykov, M. Klein, E. Schulten, and D. Fensel: **GoldenBullet: Automated Classification of Product Data in Ecommerce**. In Proceedings of Business Information Systems Conference (BIS 2002), Poznan, Poland, April 2002.
64. L. Donati, E. Iotti; G. Mordonini; A. Prati. **Fashion Product Classification through Deep Learning and Computer Vision**. Appl. Sci., 9, 1385, 2019.
65. S. Pandey, M. Supriya, and A. Shrivastava. **Data classification using machine learning approach**. In: Thampi SM, Mitra S, Mukhopadhyay J, Li, Kuan-Ching, James AP, Berretti S (editors). Intelligent Systems Technologies and Applications. Cham, Switzerland: Springer, pp. 112-122, 2017.
66. P. Ristoski, P. Petrovski, P. Mika, H. Paulheim. **A machine learning approach for product matching and categorization**. Semantic Web 9, 5, 707–728, 2018.
67. A. Shrivastava, J. Sondhi, and B. Kumar. **Machine learning technique for product classification in e-commerce data using microsoft azure cloud**. International Research Journal of Engineering & Applied Sciences, IRJEAS, Volume 5 Issue 2, Page 11-13, Apr 2017- Jun 2017.
68. Liu, H. and Cocea, M. **Semi-random partitioning of data into training and test sets in granular computing context**. Granul. Comput. 2, 357–386, 2017.
69. C. M. Bishop. **Neural Networks for Pattern Recognition**. Oxford: Oxford University Press, pp. 372, 1995.
70. K. Korjus, M. N. Hebart, R. Vicente R**. An efficient data partitioning to improve classification performance while keeping parameters interpretable**. PLoS ONE 11(8), 2016.
71. G. James, D. Witten, T. Hastie, and R. Tibshirani. **An introduction to statistical learning**, volume 112. Springer, 2013.
72. Y. Singh, P. K. Bhatia, and O.P. Sangwan. **A review of studies on machine learning techniques**. International Journal of Computer Science and Security, Volume (1), Issue (1), pp. 70-84, 2007.
73. S. Chokkadi, M. S. Sannidhan., K. B Sudeepa, A. Bhandary. **A Study on various state of the art of the Art Face Recognition System using Deep Learning** Techniques. International Journal of Advanced Trends in Computer Science and Engineering, 8(4), pp. 1590-1600, July- August 2019.
74. R Jayadi, H. M. Firmantyo, M. T. J. Dzaka, M. F. Suaidy and A. M. Putra. **Employee Performance Prediction using Naïve Bayes** International Journal of Advanced Trends in Computer Science and Engineering, 8(6), pp. 3031-3035, November - December 2019.
75. S. I. Manzoor and J. Singla. **A Comparative Analysis of Machine Learning Techniques for Spam Detection.** International Journal of Advanced Trends in Computer Science and Engineering, 8(3), 810-814, 2019.
76. A. Dey. **Machine Learning Algorithms: A Review**. International Journal of Computer Science and Information Technologies, Vol. 7 (3), 1174-1179, 2016.

77. Y. K. Alapati and K. Sindhu**. Combining clustering with classification: a technique to improve classification accuracy**. Lung Cancer, 32(57), 3, 2016.
78. A. Bansal, M. Sharma and S. Goel. **Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining**. International Journal of Computer Applications, 157(6), 975–8887, 2017.