

A Comparative Study of Machine Learning Techniques for Automatic Product Categorisation

Chanawee Chavaltada¹, Kitsuchart Pasupa¹(✉), and David R. Hardoon²

¹ Faculty of Information Technology, King Mongkut's Institute of Technology
Ladkrabang, Bangkok 10520, Thailand

Kitsuchart@it.kmitl.ac.th

² PriceTrolley Pte. Ltd., 573969 Singapore, Singapore

Abstract. The revolution of the digital age has resulted in e-commerce where consumers' shopping is facilitated and flexible such as able to enquire about product availability and get instant response as well as able to search flexibly for products by using specific keywords, hence having an easy and precise search capability along with proper product categorisation through keywords that allow better overall shopping experience. This paper compared the performances of different machine learning techniques on product categorisation in our proposed framework. We measured the performance of each algorithm by an Area Under Receiver Operating Characteristic Curve (AUROC). Furthermore, we also applied Analysis of Variance (ANOVA) to our results to find out whether the differences were significant or not. Naïve Bayes was found to be the most effective algorithm in this investigation.

Keywords: Product classification · Product categorisation · Machine learning

1 Introduction

The revolution of the digital age has resulted in e-commerce and purchasing of goods has shifted from buying at physical stores to buying from virtual outlets via online shopping where consumers are facilitated with shopping ease and flexibility such as having an ability to enquire about product availability and get instant response as well as having a flexibility to search for products using specific keywords while also being able to access a description and perform a call to action. In addition, intelligent search functions can provide consumers some suggested products that are relevant to the search keyword. Therefore, having an easy and precise search capability along with proper product categorisation through keywords allow potential customers to have an overall better shopping experience.

The United Nations Standard Products and Service (UNSPC) is a product and service taxonomy standard that was established according to the United

Nations' Common Coding System (UNCCS) and the Dun & Bradstreet's Standard Product and Service Codes (SPSC) [1]. Furthermore, common products have also been categorised by product domain experts; however, these categorisation approaches have proven effective only if the number of products is small.

The challenge of large-scale, accurate and automated categorisation has motivated exploration of computerised approaches where machine learning is a natural avenue given a framework for a computer algorithm to learn from a set of data while continuously optimising the categorisation operation to reduce error, time and cost [2]. Supervised learning is a widely used type of machine learning that requires learning from a set of training data in order that the trained model will be efficient before it is used for an actual analysis. In this case, products can be classified into appropriate categories. However, a product dataset is usually represented as a corpus of documents that possesses an a priori text processing challenge to be overcome before a classification model can be developed [3]. Examples of text processing techniques are number removal, punctuation removal, stop word removal, conversion to lowercase, and tokenisation. Then, n -gram model is used for feature extraction, i.e., it counts the frequency of words that are subsequently vectorised for use in text classification [4]. The common classification techniques for document analysis include Naïve Bayes [5], Support Vector Machine [6], Artificial Neural Networks [7], Latent Discriminant [8] Regression and Logistic Regression [9].

In this paper, we focused comparing the performances of multiple machine learning methods on product categorisation. In our experiment, we collected product name and category data from three online shopping websites. Prior to the classification process, the data were pre-processed with text processing techniques mentioned above and an n -gram model was used in feature extraction. Subsequently, classification models were built from popular techniques including NB, SVM, ANNs, and LR which were described in Sect. 2. The experiment framework is discussed in Sect. 3. In Sect. 4, we presented our results and finally conclude the paper with a discussion and conclusion in Sect. 5.

2 Methodology

In this paper, we followed the overall product classification methodology illustrated in Fig. 1.

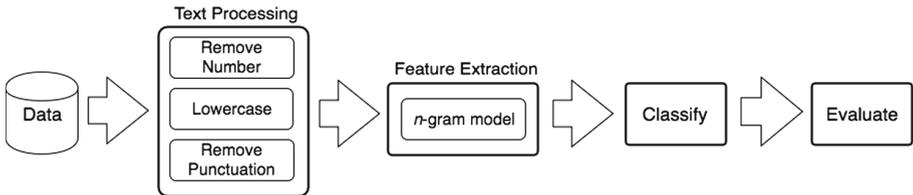


Fig. 1. Product classification processes.

2.1 Text Processing

Document data format is often done by converting the data into a compatible format for each respective process or text processing. This approach manipulates text into utility data. There are a number of text preprocessing techniques such as number removal, punctuation removal, conversion of letters to lowercase, and tokenisation. Usually, they are applied for information retrieval, information extraction and data mining.

2.2 Feature Extraction

A feature is an individual measurable attribute of an occurrence being observed [10]. This step is necessary for building effective algorithms. Effective features are discriminant, independent and informative. For extracting features from documents, count vectorisation is a good method. It counts word frequency. Here, we used an n -gram model as a linguistic probability model for predicting items in the same sequence order as that of a Markov model [11] and extracting features.

2.3 Classification

Classification is a data mining and supervised learning technique with an objective to predict an outcome by learning a statistical model of historical data attributes (also known as training data). Each classification method has different tuning parameters that affect the efficiency of the model.

Naïve Bayes (NB). This technique is based on Bayes' theorem with strong independence assumptions between features. It is usually used for text classification by calculating the probabilities of occurrences of items or posterior probabilities given that an occurrence and the previous occurrence are independent. Then the occurrence with highest probability is chosen. Moreover, some models such as text classification model has multiple labels, so a multinomial model has to be used. This model is used for prediction of frequency of corpus occurrences with an assumption that the length of the document is related to label according to Bayes' theorem. Therefore, each document represents a bag of words. The words are counted so that the probability for each label can be calculated [12].

Support Vector Machine (SVM). SVM selects and utilises proper representative instances from a training set as a support vector. SVM constructs hyperplanes between support vectors of each class, which can then be used basically for linear classification. In order to make a model as a non-linear classifier, a kernel function is applied. Kernel function maps input data in a lower dimensional feature space to a higher dimensional feature space. Some common kernel functions include Polynomial and Radial basis function (RBF) [13].

Artificial Neural Networks (ANNs). ANNs is a model of biological neural structure that receives an input through an axon into a cell body and send an output to the next neuron via a synapse. This process involves a lot of neurons that are connected in parallel and have an ability to learn from a mistake in order to improve themselves by adjusting the weight of each neuron. A neural networks model has three main type of layers. The first layer is an input layer for receiving data and sending them to next layer. The second layer consists of hidden layers that are responsible for computing and improving nodes. The performance and accuracy of a model depend on this the characteristics of this layer such the number of hidden layers and nodes. After the data were processed by the hidden layers, the output layer determines the answer by using an activated function for a specified problem.

Logistic Regression (LR). LR is a regression model where dependent attributes are categorical. Commonly, a regression model is used for analysing an event probability by using an expect value that affects event. There are two types of LR: Binary Logistic Regression and Multinomial Logistic Regression. The differences between both types are in the types of labels which are binary and multinomial, respectively. In the case that the labels are multiple values, we must use the multinomial logistic regression [14].

3 Experimental Framework

3.1 Data Collection

The data have been collected from three online shopping websites. It consisted of product names and categories. Details of each data are explained in Table 1.

Table 1. Details of dataset A, B and C.

Dataset	Product names	Categories
A	5,863	58
B	11,658	89
C	28,355	468

3.2 Data Preprocessing

The collected data were transformed to a structured format. This was done by applying text processing techniques on the product names. Punctuations and numbers were removed from the product names. Then, all of the letters were converted to lowercase. Then, we were able to extract features by using an n -gram model to transform the data into feature vectors for use in the models as shown in Fig. 1. The data were normalised by z -score. The product names were used as input data, but product categories were encoded into numerical data for use as labels, and the labels were used for predicting the target for classifier.

3.3 Experiment Setting

Data were split into two sets. Eighty and twenty percent of the data were used as a training set and a test set, respectively. They were pre-processed and feature extracted by methods explained in Sect. 2.1. Since all algorithms required parameter tuning, five-fold cross validation was applied to find optimal parameters for the best model on training set. The performance was evaluated by Area Under Receiver Operating Characteristic Curve (AUROC) measure that is more suitable for handling imbalance data than accuracy measure is and also more statistically consistent [15]. We evaluated the performances on different sets of features which were $\{1\}$ -gram, $\{1, 2\}$ -gram, $\{1, 2, 3\}$ -gram and $\{1, 2, 3, 4\}$ -gram. Subsequently, the parameters for each algorithm were varied as follows:

- ANNs: Hidden layers were $\{1, 2, 3\}$, and the number of neuron for each layer were $\{10, 20, 30, \dots, 100\}$
- SVM: C value was in range $\{10^{-4}, 10^{-3}, \dots, 10^5, 10^6\}$. We evaluated three types of kernel which are Linear, Polynomial, and Radial Basis Function (RBF). Degrees of Polynomial were in range 1–6 and Gaussian width range was $\{10^{-6}, 10^{-5}, \dots, 10^5, 10^6\}$.
- LR: Regularisation parameter was in the range of $\{10^{-4}, 10^{-3}, \dots, 10^5, 10^6\}$.

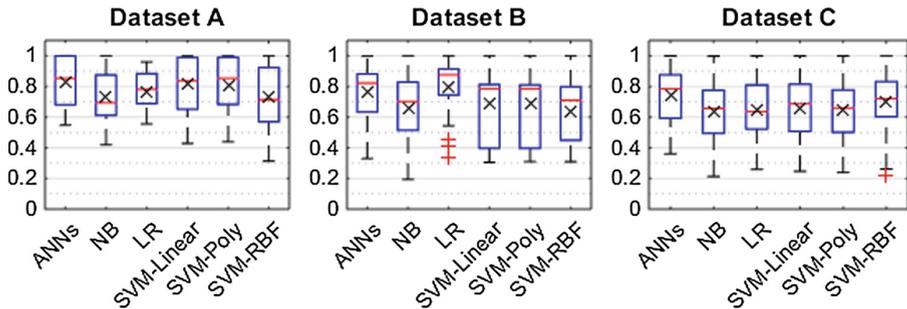
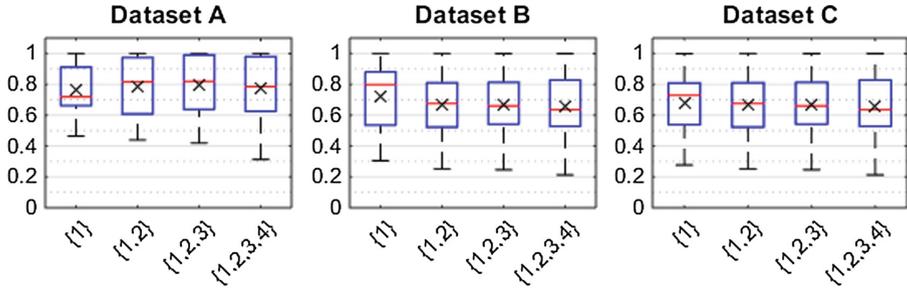


Fig. 2. The box plots show average of AUROC across feature and algorithms for each datasets with 10 runs

Once the optimal parameters had been set, they were used to train a model which was later tested and evaluated on the test set. We ran the experiment 10 times, each with a different random split.

4 Results and Discussions

Figure 2(a) shows the average AUROC for each set of features across all considered classification techniques. Using the set of features 1, 2, 3-gram gave the best performance for dataset A, while for the dataset B and C, using only the unigram ($n = 1$) set was needed for a good performance. However, from the results for dataset A, B and C, the determination of the best set of features was inconclusive at $p = 0.79, 0.35, 0.96$ respectively with one-way analysis of variance (ANOVA). Respectively, as determined by one-way analysis of variance (ANOVA, a statistical comparison technique that provides a capability to compare differences between means [16]).

Furthermore, we compared the performances of six different techniques on three datasets (averaged across four sets of features) as shown in Fig. 2(b). NB showed the highest average AUROC for sdataset A and C but for dataset B, the highest was LR. It was found that there were interactions between the sets of features and the algorithm used in this framework for every dataset–dataset A and B at $p < 0.01$ and C at $p < 0.05$ by two-way ANOVA. This means that

Table 2. Multiple comparison–it shows mean difference (MD) and its p -value. Bold face indicates statistically significant.

Algorithm 1	Algorithm 2	DataSet A		DataSet B		DataSet C	
		MD	p -value	MD	p -value	MD	p -value
NB	LR	0.063	<0.05	-0.025	0.889	0.097	<0.01
NB	ANNs	0.091	<0.01	0.114	<0.01	0.104	<0.01
NB	SVM Linear	0.013	0.990	0.080	<0.01	0.083	<0.01
NB	SVM Poly	0.025	0.843	0.089	<0.01	0.095	<0.01
NB	SVM RBF	0.092	<0.01	0.134	<0.01	0.045	0.086
LR	ANNs	0.029	0.746	0.138	<0.01	0.007	0.998
LR	SVM Linear	-0.050	0.166	0.105	<0.01	-0.013	0.967
LR	SVM Poly	-0.038	0.459	0.106	<0.01	-0.002	0.999
LR	SVM RBF	0.029	0.742	0.159	<0.01	-0.052	<0.05
ANNs	SVM Linear	-0.079	<0.01	-0.034	0.671	-0.021	0.817
ANNs	SVM Poly	-0.067	<0.05	-0.034	0.722	-0.009	0.995
ANNs	SVM RBF	<0.001	0.999	0.02	0.949	-0.059	<0.01
SVM Linear	SVM Poly	0.012	0.992	0.020	0.999	0.012	0.981
SVM Linear	SVM RBF	0.079	<0.01	0.054	0.172	-0.039	0.187
SVM Poly	SVM RBF	0.067	<0.05	0.052	0.204	-0.051	<0.05

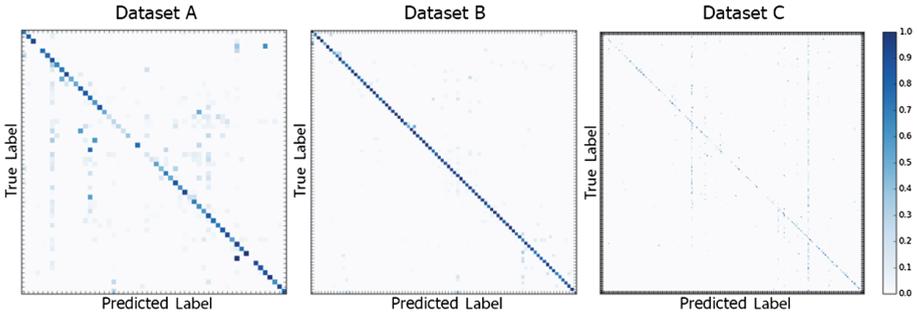


Fig. 3. Confusion matrix of NB on all three datasets.

there was a significant difference between at least one pair of means for each dataset; therefore, we subsequently conducted multiple comparison by two-way ANOVA on each dataset as illustrated in Table 2. Clearly, NB yielded a better performance than those of the others in 3/5 cases for dataset A, 4/5 cases for both dataset B, and C ($p < 0.01$). It can be seen, for dataset B, LR performances were significantly different better in 4/5 cases ($p < 0.01$), but inconclusive when comparing to NB ($p = 0.889$)—NB and LR are comparable in this case. ANNs were found to be worse than NB, SVM-Linear and SVM-Poly ($p < 0.05$) for dataset A, while they were worse than NB and LR for dataset B ($p < 0.01$), and they were worse than NB and SVM-RBF for dataset C ($p < 0.01$). It was inconclusive which algorithm was the worst after all. Moreover, the confusion matrices of NB on all three datasets in Fig. 3 show that it was the best algorithm in this framework.

5 Conclusion

This paper proposes a framework for automatic product categorisation. We evaluated and compared well-known machine learning techniques on three datasets obtained from the online websites and based on AUROC. We have found that the performance of NB was the best—statistically significant. Furthermore, it is inconclusive whether a set of proposed features was the best.

References

1. Ding, Y., Korotkiy, M., Omelayenko, B., Kartseva, V., Zykov, V., Klein, M., Schulten, E., Fensel, D.: GoldenBullet: automated classification of product data in e-commerce. In: Proceedings of the 5th International Conference on Business Information Systems (BIS 2002) (2002)
2. Simon, P.: Too Big to Ignore: The Business Case for Big Data. Wiley, Hoboken (2013)
3. Shankar, S., Lin, I.: Applying machine learning to product categorization. Technical report, Stanford University (2011)

4. Kozareva, Z.: Everyone likes shopping! multi-class product categorization for e-commerce. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1329–1333 (2015)
5. Zhang, H., Li, D.: Naïve bayes text classifier. In: Proceedings of the 2007 IEEE International Conference on Granular Computing (GRC 2007), p. 708 (2007)
6. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.* **2**, 45–66 (2001)
7. Wermter, S.: Neural network agents for learning semantic text classification. *Inf. Retr.* **3**(2), 87–103 (2000)
8. Wang, Z., Qian, X.: Text categorization based on LDA and SVM. In: 2008 International Conference on Computer Science and Software Engineering, vol. 1, pp. 674–677 (2008)
9. Cheng, W., Hüllermeier, E.: Combining instance-based learning and logistic regression for multilabel classification. *Mach. Learn.* **76**(2), 211–225 (2009)
10. Bishop, C.: *Pattern Recognition and Machine Learning*, vol. 128, 1st edn. Springer, New York (2006). pp. 1–58, ISSN 1613-9011
11. Jurafsky, D., Martin, J.H.: *Speech and language processing*. Int. Ed. **710**, 117–119 (2000)
12. Lewis, D.D.: Naive (Bayes) at forty: the independence assumption in information retrieval. In: Nédellec, C., Rouveirol, C. (eds.) *ECML 1998*. LNCS, vol. 1398, pp. 4–15. Springer, Heidelberg (1998). doi:[10.1007/BFb0026666](https://doi.org/10.1007/BFb0026666)
13. Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. *Neural Process. Lett.* **9**(3), 293–300 (1999)
14. Yuth, K.: Principle and using logistic regression analysis for research. *RMUTSV Res. J.* **4**(1), 1–12 (2012)
15. Ling, X.C., Huang, J., Zhang, H.: AUC: a statistically consistent and more discriminating measure than accuracy. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003), vol. 3, pp. 519–524 (2003)
16. Viaene, S., Derrig, R.A., Baesens, B., Dedene, G.: A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *J. Risk Insur.* **69**(3), 373–421 (2002)



<http://www.springer.com/978-3-319-59071-4>

Advances in Neural Networks - ISNN 2017
14th International Symposium, ISNN 2017, Sapporo,
Hakodate, and Muroran, Hokkaido, Japan, June 21-26,
2017, Proceedings, Part I
Cong, F.; Leung, A.C.-S.; Wei, Q. (Eds.)
2017, XXII, 583 p. 238 illus., Softcover
ISBN: 978-3-319-59071-4