Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Huế, ngày 07-08/6/2019
DOI: 10.15625/vap.2019.00063

TOWARDS MACHINE LEARNING APPROACHES TO IDENTIFY SHRIMP DISEASES BASED ON DESCRIPTION

Luyl-Da Quach[1], Long Quach Hoang[2], Nghia Duong-Trung[3], Chi-Ngon Nguyen[4]

[1] FPT University, Can Tho city, Vietnam
[2]Tay Do University, Can Tho city, Vietnam
[3]Can Tho University of Technology, Can Tho city, Vietnam
[4]Can Tho University, Can Tho city, Vietnam

luyldaquach@gmail.com; qhlong1997@gmail.com; dtnghia@ctuet.edu.vn,duong-trung@ismll.de; ncngon@ctu.edu.vn

ABSTRACT: *Shrimp farming is a key sector in economic development in Mekong Delta provinces. Unfortunately, there are many problems in shrimp farming, especially shrimp diseases which cause a considerable loss. Shrimp diseases are expressed through symptoms and manifestations of shrimp. Recognizing the importance of shrimp symptoms to help raise an early warning, in this research the authors apply several state-of-the-art text classification algorithms such as Logistic Regression, Random Forest, Naïve Bayes, Support Vector Machines, and Multilayer Perceptron on a collection of 1098 observations categorizing into 14 distinct classes. Several thorough evaluation scenarios have been conducted including a process tokenization and models' comparison on the obtained data set with different ratios. The results show that Support Vector Machines achieves the highest classification accuracy (81.27%), followed by Multilayer Perceptron, Random Forest, Logistic Regression, and Naïve Bayes. Through the results of the study, it is feasible to apply machine learning algorithms to diagnose shrimp diseases entirely based on textual symptom descriptions.*

Keywords*: Diagnosis of shrimp diseases, text classification, comparison of machine learning algorithms.*

I. INTRODUCTION

Aquaculture occupies an important position in Vietnam, of which shrimp farming is considered a major component. The Mekong Delta has great potential and benefits in developing shrimp farming. However, the development of shrimp farming is still difficult. The first reason is natural problems such as drought and saline intrusion. Another cause comes from spontaneous shrimp farming. Poor irrigation development, undeveloped infrastructure, and indiscriminate use of fertilizers might lead to water pollution and uncontrolled effects. This is the cause of shrimp diseases and consequently shrimp deaths.

In order to effectively prevent shrimp diseases, it is common for shrimp farmers to perform daily monitoring and understanding of signs of shrimp diseases in order to effectively detect and prevent them. In particular, the detection can be investigated through the states on shrimp (Figure 1) or through disease symptoms. For example shrimp eats a lot of abnormality for a few days, then stops eating. A few days later, shrimps are death [1].

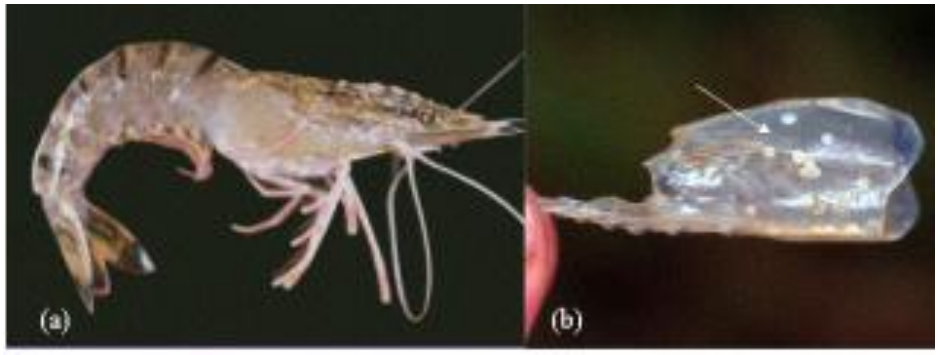

Figure 1. White Spot Syndrome virus [2]

Recognizing the problem of shrimp manifestation is a necessary issue. The study proposes text-based classification solutions to predict shrimp diseases. Text classification is an important technique in data mining and natural language processing. It involves building a text classification system to predict categories based on textual descriptions. Descriptive text classification has been applied in many research fields including emotional analysis [3], understanding user intent [4] and especially health issues. An intelligent heart disease prediction system is based on information about medical records using data mining algorithms [5]. Many effective techniques for predicting heart disease based on the description have been investigated in [6]. However, no research has been conducted to predict shrimp diseases based on text descriptions.

In this study, the authors propose a practical pipeline to identify and label shrimp diseases by applying state of the art machine learning algorithms on textual sources. Some text samples are described in Table 1.

Table 1. Exemplification of shrimp's disease description

Vietnamese description	English description	Label
Tôm có màu xỉn, vỏ bị mềm có khi rất mềm, vỏ rời thịt, thường yếu, kém hoạt động, dễ bị con khác ăn thịt.	Shrimp's color is faded. Shells are very soft, shell-meat separation, Shrimps are weak, inactive and easily be eaten by others.	Chronic soft-shell syndrome
Mang tôm bị đổi sang màu nâu hay đen, phụ bộ và vỏ bị mờ đục, nếu vỏ có màu xanh.	The shrimp plaque changes to brown or black. Some parts are opaque and green.	Plaque disease in shrimp
Tôm thường nổi đầu, dạt bờ và chết rải rác, không lột xác được.	Shrimp raises their heads. They have washed ashore and die scattered. They can not moult.	Filamentous bacteria
Bơi lội nhanh nhẹn, không định hướng và thân tôm trắng mờ đục.	Shrimp swims fast without orientation. Shrimp's body turns opaque and white.	White faeces disease

In this study, we perform the following steps:

- Step 1: Collect shrimp manifestations and label corresponding diseases. However, for Vietnamese text, there will exist compound words. So in order to make a complete evaluation, the research team has obtained data with and without compound words. Then 2 sets of data have independently experimented. The procedure of training models and making a prediction is shown in Figure 2 and Figure 3 respectively.
- Step 2: Perform machine learning algorithms, e.g. linear regression, decision trees, random forest, multilayer perceptron and support vector machines, to classify text sources.
- Step 3: Compare prediction accuracy.

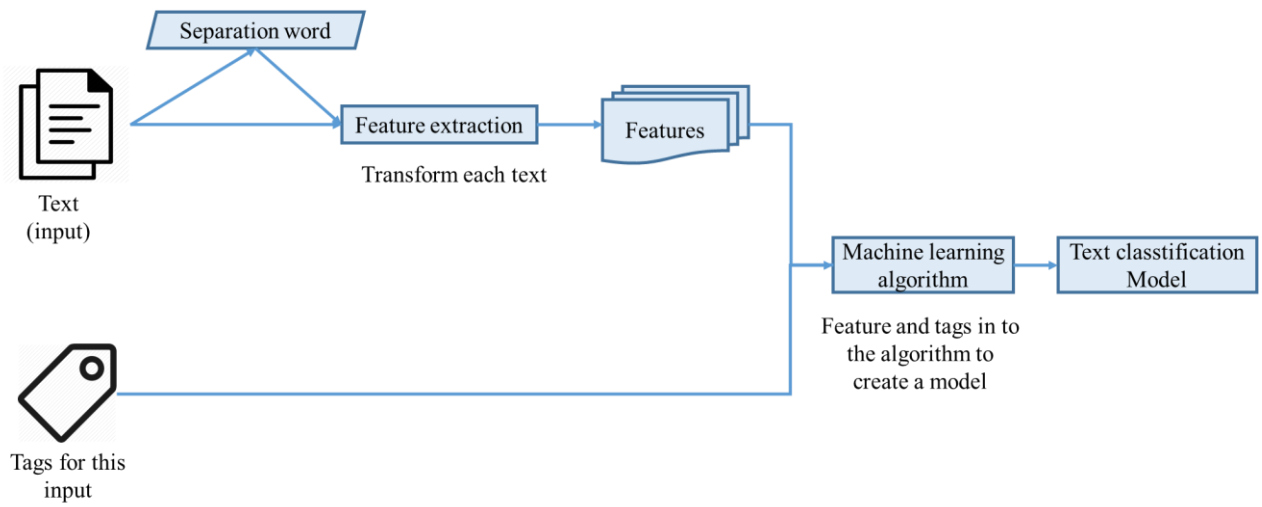

Figure 2. Training model for disease identification system on shrimp

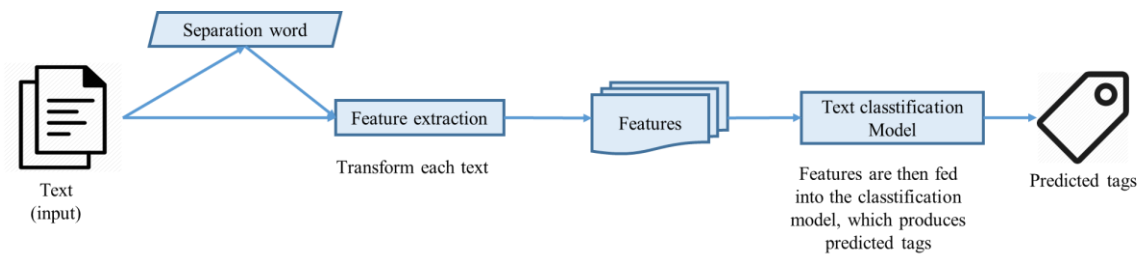

Figure 3. Shrimp disease prediction model

The main contribution of the study is as follows. Firstly, the authors propose a practical procedure in developing an effective shrimp disease prediction system based on textual sources. Secondly, the experiment results help assess effective classification algorithms in similar circumstances. And lastly, the authors compare the accuracy of algorithms based on separated and non-separated compound words in Vietnamese.

II. DATA DESCRIPTION AND PRE-PROCESSING

According to OIE - World Organization for Animal Health statistics [7], there are 22 types of viruses causing shrimp diseases in the world, including White Spot Syndrome Virus (WSSV), Yellow Head Virus (YHV), Infectious Myonecrosis Virus (IMNV). Therefore, to accurately get the description of disease symptoms of shrimp, the authors

take samples description from experts in the fisheries sector (especially shrimp) at the Department of Fisheries, Can Tho University. We obtain 1098 symptoms of 14 diseases equivalent to 14 classes. After tokenizing by VnTokenizer tool [8], the final text source contains 13,646 words. Several descriptions are listed in Table 1, e.g. the Vietnamese description column.

Table 2. Statistical data describing shrimp disease manifestations

#	Scientific name	Vietnamese description	Pathogen	Number of samples	Reference
1	Infection with Vibrio	Bệnh do Vibriosis	Vibrio bacteria.	75	[9]
2	Filamentous bacteria	Bệnh do vi khuẩn dạng sợi	The disease symptom concentrates in a shell.	65	[10]
3	Plaque disease in shrimp	Bệnh đóng rong hay mảng bám	Clinging shells, gills and parts of shrimps make stress. If serious, shrimps will not be able to peel and be weak resistance to other diseases.	76	[11]
4	Chronic soft-shell syndrome	Bệnh mềm vỏ ở tôm	Shrimp is deficient in vitamins and minerals, especially lack of calcium and phosphor.	121	[12]
5	Vitamin C deficiency in shrimp	Bệnh thiếu vitamin C ở tôm	Inadequate supply of vitamin C in food.	66	[13]
6	Black gill disease	Bệnh đen mang	Infertility factors from Fusarium SPP.	86	[14]
7	Taura syndrome	Hội chứng taura	Taura syndrome virus.	81	[15]
8	White spot disease	Bệnh đốm trắng	White spot syndrome virus.	63	[16]
9	Necrotizing hepatopancreatitis	Bệnh Hội chứng hoại tử gan tụy cấp tính	Proteobacteria.	107	[17]
10	White feces syndrome	Hội chứng phân trắng	Infected with V.vulnificus, V.fluvialis, V.alginolyticus.	67	[18]
11	Nuclear polyhedrosis baculovirosis	Bệnh còi do vi rút có nhân đa diện	Baculoviridae virus: Baculovirus penaei, Monodon baculovirus.	63	[19]
12	Yellow dead disease	Bệnh đầu vàng	Yellowhead complex virus.	79	[20]
13	Infectious myonecrosis	Bệnh hoại tử cơ, bệnh đục cơ, bệnh cong thân	Infectious myonecrosis virus.	73	[21]
14	Luminous bacteria disease	Bệnh phát sáng	Luminescent vibrio group bacteria: Vibrio Harveyi.	76	[22]

III. LEARNING SOLUTIONS APPLICABLE TO THE TEXTUAL DESCRIPTION

A. Logistic Regression

Logistic regression is a simple and useful machine learning algorithm that frequently apply in a wide range of practices [23]. Logistic regression algorithm is used a lot in classifying email, images, especially in the field of natural language processing. It uses linear equations with independent predictors to predict a value in the area between the negative infinity and the infinity.

B. Random Forest

Random forest is a popular supervised machine learning algorithm because of its flexibility, simplicity, and ease of use. It is used in both classification and regression tasks [24]. The general idea of the random forest algorithm is that it combines multiple machine learning models, e.g. decision trees, to increase the prediction accuracy of the algorithm. A random forest consists of multiple random decision trees. Two types of randomness are built into the trees. First, each tree is built on a random sample from the original data. Second, at each tree node, a subset of features are randomly selected to generate the best split. In our implementation, we carried out a random forest with 200 random trees and using entropy information gain (Equation 1).

$$entropy(S) = -\sum_{i=1}^{n} p_i log_2 p_i \quad (1)$$

C. Naïve Bayes

Naïve Bayes algorithm is a simple supervised machine learning algorithm that applies the Bayes theorem [25]. In particular, the Bayes probability theorem performs the list of relationships between variables y and the dependent characteristic vector x (Equation 2). The different versions of Naïve Bayes algorithms are done by applying multiple assumptions regarding probability distribution must include Gaussian Naïve Bayes [26], Complement Naive Bayes [27], and Bernoulli Naive Bayes [28].

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(y)P(x_1, x_2, \dots, x_n|y)}{P(x_1, x_2, \dots, x_n)} \quad (2)$$

The Gaussian Naïve Bayes method used in the study with each class is distributed according to the Gaussian distribution. The application of the normal distribution is expected by μ_y and the variance σ_y^2 with i is the dimension of the instance and the corresponding y. See Equation (3) as follows:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\left(x_i - \boldsymbol{\mu_y}\right)^2}{2\sigma_y^2}\right) \quad (3)$$

Maximum likelihood is used to calculate σ_y^2và μ_y as follows:

$$\left(\mu_y, \sigma_y^2\right) = \underset{\mu_y, \sigma_y^2}{argmax} \prod_{n=1}^{N} p(x_i^{(n)}|\mu_y, \sigma_y^2) \quad (4)$$

D. Support vector machines

Support vector machines (SVM) is a supervised machine learning method, used for classification, regression and exception detection [29]. SVM is considered an effective approach to solve classification problems with large dimension data. The basic SVM algorithm solves the problem of linear classification, however, if we combine SVM with kernels, it will allow solving some nonlinear problems by data mapping which result in a space with a higher dimension. Without any necessary technical changes, the only thing to do is to replace the scalar products of two vectors u.v by a kernel function K (u, v) (Table 3). In this study, we use SVM with a linear kernel function [30], with the error parameter c = 2.0.

Table 3. Several common kernels used in SVM

Type of function	Kernel
Linear	$K(u,v) = u.v$
Polynomial with degree d	$K(u,v) = (u.v + c)^d$
Radial basis function	$K(u,v) = exp(-\gamma\|u-v\|^2)$

E. Multi-layer Perceptron

The artificial neural network model is a supervised machine learning method, based on simple mathematical models of the human brain [31]. The artificial neural network consists of a set of units dealing with the activation or output states of the processing unit (Figure 4). In neural networks, there are 3 types of units: input units, output units and hidden units.

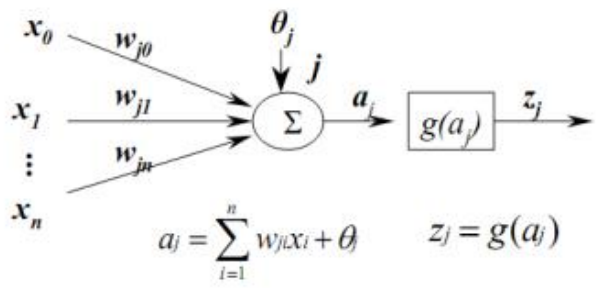

Figure 4. One computing unit in a neural network

Multi-layer Perceptron (MLP) is a class of feedforward artificial neural networks with 3 layers: an input layer, an output layer, and a hidden layer. Each node exists a nonlinear activation function (Figure 5).

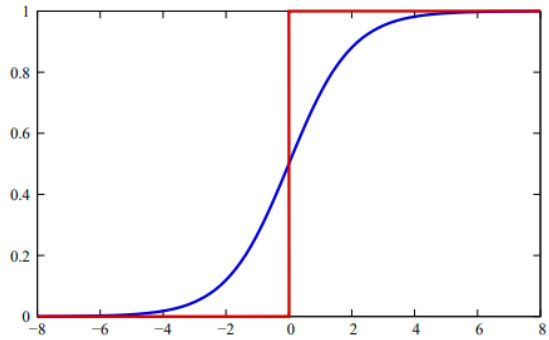

Figure 5. Activation curve

In this study, we use multi-layer perceptron with the number of nodes in the hidden layer is 42, the maximum number of iterations is 1000, the error parameter is 0.001, the learning rate is 0.1.

F. VnTokenizer tool

The VnTokenizer tool [8] was developed in the VLSP project done by a research group lead by Prof. Ho Tu Bao. This tool is based on the method of maximum matching with the data set used is the Vietnamese syllabary and Vietnamese vocabulary dictionary. It is also built on the Java language that can be easily integrated into other Vietnamese analysis systems. The process of performing word separation according to the maximum matching method is shown in Figure 6.

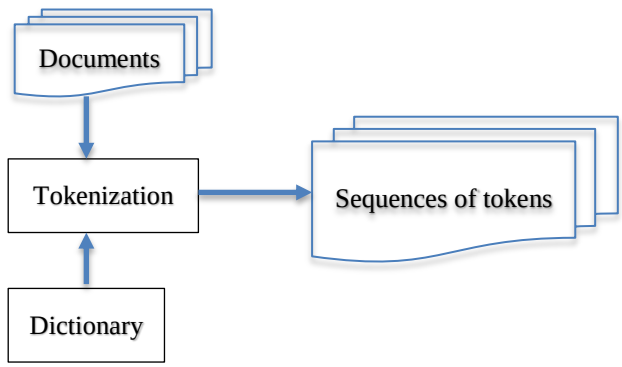

Figure 6. Word tokenization by maximum matching

The word units generated from this tool include words in the dictionary, numeric strings, foreign character strings, punctuation, other mixed characters in the text, new words, words that are generated, or a string of symbols not listed in the dictionary. This tool separates words for the accuracy of 96% -98% [8], used in document [31,32].

IV. EXPERIMENTAL RESULTS

The study of shrimp disease classification based on symptom description has been carried out on 1098 observations associated with 14 classes. The authors tokenize words by spaces resulting in Dataset 1 and Dataset 2. The difference between these two datasets is that the training-test split ratio is 70-30 and 80-20 in case of Dataset 1 and 2 respectively. We also apply VnTokenizer tool [8] to form Dataset 3 and 4. The difference between them is similar to the previous explanation. Within each dataset's investigation, we randomly shuffle data into training and test sets, execute the machine learning models, and take average in the end. Results are shown in Tables 4, 5 and comparison graphs (Figure 7). In this study, the highest accuracy of 81.27% is achieved by SVM.

- For Dataset 1: The highest accuracy is 78.55% achieving by SVM, and the lowest accuracy is 72.30% achieving by logistic regression.
- For Dataset 2: When the training and test ratio is 80-20, the highest accuracy is 81.27%, and the lowest accuracy is 73.64% obtaining by Naïve Bayes model. The overall results are higher than those of Dataset 1.
- For Dataset 3 and 4: Tokenization done by VnTokenizer tool does not provide any improvements in all cases.

The accuracy of the SVM algorithm is the highest (81.27%); however, it has an approximate accuracy of MLP. The reason is that the MLP algorithm uses only 3-layer model (including 1 hidden layer) and the prediction is basically the multiplication of two weight matrices. Whereas, the SVM algorithm performs a boundary determination via support vectors that classifies a given data point. SVM clearly defined the boundary directly from the training data. This minimizes the distance between points and support vectors. However, this accuracy compared with that of MLP is not significantly different. The next high precision is done by random forest algorithm (79.73%) which is the building of

decision trees and performing random removal of properties. Finally, the accuracy of logistic regression (73.73%) and naïve Bayes (73.64%) are the next. Both models use mathematical basics. Logistic regression is based on logic functions, while naïve Bayes is based on probability. It overcomes the disadvantages of the qualitative models, expressing objectiveness and consistency.

Regarding the processing time of all experimental algorithms on 4 datasets, SVM gains the highest accuracy, despite the fact that the processing time of SVM is rather slower than that of the remaining algorithms. This is explained by the natural behaviour of SVM that the determination of support vectors takes a lot of time. Meanwhile, Logistic Regression and Naïve Bayes algorithms are purely based on linear combination and probability respectively. As a result, their execution time is faster and Logistic Regression's execution time is the fastest. The excecution time difference on datasets splitting with VnTokenizer and on datasets that have not been tokenized by VnTokenizer is not significantly different. In some cases, tokenizing datasets via VnTokenizer does not provide much help.

Table 4. Comparison results on data sets tokenized by black spaces

Models	Dataset 1 (Accuracy (%) / execution time (s))						Dataset 2 (Accuracy (%) / execution time (s))					
	1	2	3	4	5	Average	1	2	3	4	5	Average
Logistic Regression	70.91	73.64	72.12	73.33	71.52	72.30	75.91	73.18	74.55	71.36	73.64	73.73
	0.10	0.10	0.11	0.11	0.11	**0.11**	0.67	0.11	0.11	0.17	0.10	**0.23**
Random Forest	74.55	73.64	75.76	76.67	72.12	74.55	82.27	81.36	79.09	79.55	76.36	79.73
	1.12	0.88	0.89	1.01	0.80	0.94	1.25	0.78	0.77	0.83	0.89	0.90
Naïve Bayes	68.79	72.12	69.39	74.24	70.91	71.09	74.55	74.09	73.64	73.64	72.27	73.64
	0.31	0.28	0.29	0.29	0.29	0.29	0.31	0.29	0.26	0.28	0.29	0.29
SVM	77.88	78.79	78.18	80.91	76.97	**78.55**	82.73	82.73	80.00	80.45	80.45	**81.27**
	0.69	0.70	0.70	0.69	0.93	0.74	0.83	1.01	0.84	0.83	0.92	0.89
MLP	75.76	76.67	77.88	80.91	76.97	77.64	80.91	83.18	78.64	79.09	81.36	80.64
	8.55	8.24	8.93	8.56	9.36	8.73	10.5	10.7	10.2	10.6	12.6	10.9

Table 5. Compare results on the dataset tokenized by VnTokenizer tool

Models	Dataset 3 (Accuracy (%) / execution time (s))						Dataset 4 (Accuracy (%) / execution time (s))					
	1	2	3	4	5	Average	1	2	3	4	5	Average
Logistic Regression	71.52	75.15	71.52	72.12	70.61	72.18	75.91	76.36	72.27	74.09	72.27	74.18
	0.12	0.13	0.11	0.12	0.12	**0.12**	0.12	0.13	0.11	0.11	0.12	**0.12**
Random Forest	76.97	74.24	75.45	78.48	73.64	75.76	80.91	80.45	79.09	81.82	74.55	79.36
	1.03	0.82	1.04	0.91	0.92	0.94	0.90	0.85	0.79	0.77	0.89	0.84
Naïve Bayes	68.79	70.61	69.09	74.24	70.00	70.55	75.45	73.18	70.45	75.00	70.00	72.82
	0.29	0.34	0.34	0.44	0.32	0.35	0.33	0.31	0.33	0.32	0.51	0.36
SVM	76.67	78.79	76.97	80.30	77.27	**78.00**	82.27	83.64	78.64	78.64	77.73	**80.18**
	0.75	0.79	0.81	0.88	0.77	0.80	0.94	0.76	0.92	0.92	1.14	0.94
MLP	76.36	76.67	76.36	79.70	76.36	77.09	80.91	82.27	78.64	77.73	75.91	79.09
	8.71	10.5	9.43	10.5	9.53	9.73	12.3	11.5	11.1	12.2	12.1	11.84

V. CONCLUSION

The shrimp diagnosis system based on the description gains the highest classification accuracy (81.27%) using the SVM model. One important thing to note is that the accuracy of MLP is slightly lower than SVM although it only uses 1 hidden player. This opens a research direction in the application of deep learning algorithms to improve the classification accuracy. The intensive comparison of 5 machine learning algorithms on 4 different datasets of tokenization and ratios leads to the following conclusions:

- Tokenizing Vietnamese without separation of compound words does not improve the prediction accuracy. However, it helps reduce the number of words (22,761 words compared with 25,815 words in case of separation of compound words). The reason is that the dataset is not large enough. Hence, the amount of separated words is not much which leads to incomparable processing time.
- Among 5 models used, SVM still gains a certain advantage when obtaining the highest accuracy. The accuracy of SVM model, e.g. 81.27%, shows that the application of the system is practically feasible.

The results of the study open up a research direction in applying deep learning algorithms in shrimp disease diagnosis study based on descriptions. The future of research can be built on the images and the combination of images with symptom description for higher accuracy.

Based on the results of an intensive comparison of the time and accuracy of the five machine learning algorithms, the study can provide guidance on the development of a diagnostic system based on textual data sources. Depending on the characteristics of datasets, it is possible to select a reasonable algorithm for high accuracy and low waiting time to fulfil expected results.

VI. ACKNOWLEDGEMENT

Thank you to the experts at the Department of Fisheries, Can Tho University for supporting the process of sampling and editing.

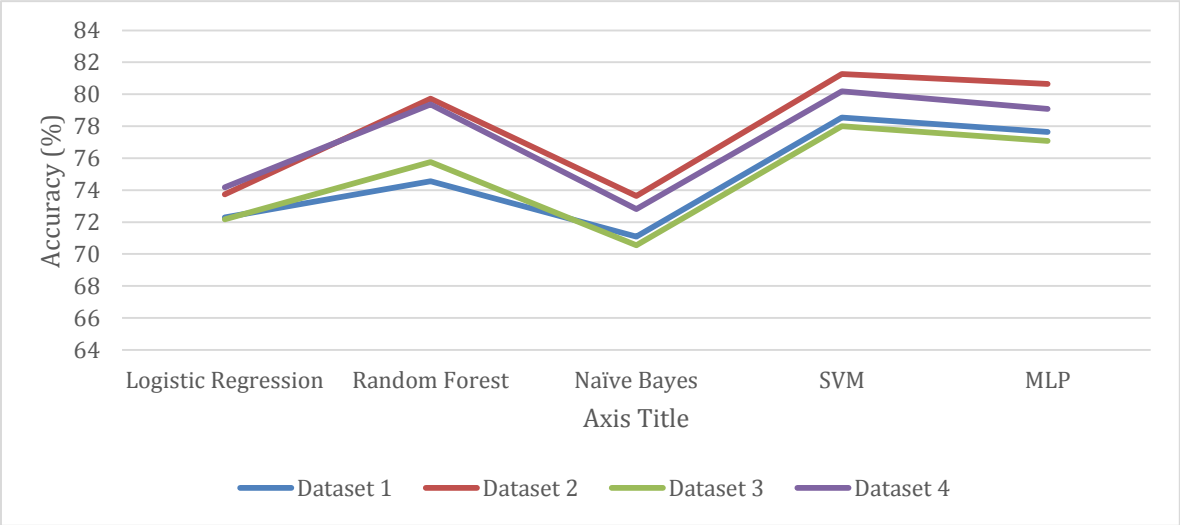

Figure 7. Comparison among experimented learning algorithms across different datasets

REFERENCES

[1] Lu, Y., Tapay, L. M., Brock, J. A., & Loh, P. C. (1994). Infection of the yellow head baculo- like virus (YBV) in two species of penaeid shrimp, Penaeus stylirostris (Stimpson) and Penaeus vannamei (Boone). Journal of Fish Diseases, 17(6), 649-656.

[2] Durand, S., Lightner, D. V., Redman, R. M., & Bonami, J. R. (1997). Ultrastructure and morphogenesis of white spot syndrome baculovirus (WSSV). Diseases of Aquatic Organisms, 29(3), 205-211.

[3] Wang, J., Wang, Z., Zhang, D., & Yan, J. (2017, August). Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification. In IJCAI (pp. 2915-2921).

[4] Hu, J., Wang, G., Lochovsky, F., Sun, J. T., & Chen, Z. (2009, April). Understanding user's query intent with wikipedia. In Proceedings of the 18th international conference on World wide web (pp. 471-480). ACM.

[5] Sa, S. (2013). Intelligent heart disease prediction system using data mining techniques. International Journal of healthcare & biomedical Research, 1, 94-101.

[6] Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. International Journal of Engineering, 1(8), 1-4.

[7] OIE - World Organisation for Animal Health (2018). Manual of diagnostic tests for aquatic animals. Paris: Office International des Epizooties.

[8] Huyen, N. T. M., Roussanaly, A., & Vinh, H. T. (2008, March). A hybrid approach to word segmentation of Vietnamese texts. In International Conference on Language and Automata Theory and Applications (pp. 240-249). Springer, Berlin, Heidelberg.

[9] Jayasree, L., Janakiram, P., & Madhavi, R. (2006). Characterization of Vibrio spp. associated with diseased shrimp from culture ponds of Andhra Pradesh (India). Journal of the world aquaculture society, 37(4), 523-532.

[10] Meng, F., Zhang, H., Yang, F., Li, Y., Xiao, J., & Zhang, X. (2006). Effect of filamentous bacteria on membrane fouling in submerged membrane bioreactor. Journal of Membrane Science, 272(1-2), 161-168.

[11] Lakshmi, B., Viswanath, B., & Sai Gopal, D. V. R. (2013). Probiotics as antiviral agents in shrimp aquaculture. Journal of pathogens, 2013.

[12] Baticados, M. C. L., Coloso, R. M., & Duremdez, R. C. (1987). Histopathology of the chronic soft-shell syndrome in the tiger prawn Penaeus monodon. Diseases of aquatic organisms, 3(1), 13-28.

[13] He, H., & Lawrence, A. L. (1993). Vitamin C requirements of the shrimp Penaeus vannamei. Aquaculture, 114(3-4), 305-316.

[14] Dewangan, N. K., Gopalakrishnan, A., Kannan, D., Shettu, N., & Singh, R. R. (2015). Black gill disease of Pacific white leg shrimp (Litopenaeus vannamei) by Aspergillus flavus. Journal of Coastal Life Medicine, 3(10), 761-765.

[15] Cao, Z., Wang, S. Y., Breeland, V., Moore, A. M., & Lotz, J. M. (2010). Taura syndrome virus loads in Litopenaeus vannamei hemolymph following infection are related to differential mortality. Diseases of aquatic organisms, 91(2), 97-103.

[16] Selvin, J., & Lipton, A. P. (2003). Vibrio alginolyticus associated with white spot disease of Penaeus monodon. Diseases of aquatic organisms, 57(1-2), 147-150.

[17] Nunan, L. M., Pantoja, C., & Lightner, D. V. (2008). Improvement of a PCR method for the detection of necrotizing hepatopancreatitis in shrimp. Diseases of aquatic organisms, 80(1), 69-73.

[18] Sriurairatana, S., Boonyawiwat, V., Gangnonngiw, W., Laosutthipong, C., Hiranchan, J., & Flegel, T. W. (2014). White feces syndrome of shrimp arises from transformation, sloughing and aggregation of hepatopancreatic microvilli into vermiform bodies superficially resembling gregarines. PloS one, 9(6), e99170.

[19] Lu, M., Farrell, P. J., Johnson, R., & Iatrou, K. (1997). A baculovirus (Bombyx mori nuclear polyhedrosis virus) repeat element functions as a powerful constitutive enhancer in transfected insect cells. Journal of Biological Chemistry, 272(49), 30724-30728.

[20] Sánchez-Barajas, M., Liñán-Cabello, M., & Mena-Herrera, A. (2008). Detection of yellow head disease in intensive freshwater production systems of Litopenaeus vannamei. Comparative Biochemistry and Physiology, Part A, 1(151), S16.

[21] Prasad, K. P., Shyam, K. U., Banu, H., Jeena, K., & Krishnan, R. (2017). Infectious Myonecrosis Virus (IMNV)–An alarming viral pathogen to Penaeid shrimps. Aquaculture, 477, 99-105.

[22] Baticados, M. C. L., Lavilla-Pitogo, C. R., Cruz-Lacierda, E. R., De La Pena, L. D., & Sunaz, N. A. (1990). Studies on the chemical control of luminous bacteria Vibrio harveyi and V. splendidus isolated from diseased Penaeus monodon larvae and rearing water. Dis. Aquat. Org, 9(2), 133-139.

[23] Hsiang-Fu Yu, Fang-Lan Huang, Chih-Jen Lin (2010), Dual coordinate descent methods for logistic regression and maximum entropy models, Machine learning Volume 85, Issue 1–2, 41–75.

[24] L. Breiman (2001), “Random Forests”, Machine Learning, 45(1), 5-32.

[25] H. Zhang (2004). The optimality of Naive Bayes. Proc. FLAIRS.d

[26] Mitchell(1997), Machine Learning. McGraw-Hill Science/Engineering/Math, p.432.

[27] Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In ICML (Vol. 3, pp. 616-623).

[28] A. McCallum and K. Nigam (1998). A comparison of event models for Naive Bayes text classification. Proc. AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41-48.

[29] Cortes, Corinna; and Vapnik, Vladimir N.(1995). Support-Vector Networks. Machine Learning, 20(3). 273–297.

[30] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification Journal of Machine Learning Research 9(2008), 1871-1874.

[31] Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996).

[32] L. Da. Quach and C. Nguyen, "Conversion of the Vietnammese Grammar into Sign Language Structure using the Example-Based Machine Translation Algorithm," 2018 International Conference on Advanced Technologies for Communications (ATC), Ho Chi Minh City, 2018, pp. 27-31.

[33] Da Q. L., Khang N. H. D., Ngon N. C. Converting the Vietnamese Television News into 3D Sign Language Animations for the Deaf. In: Duong T., Vo NS. (eds) Industrial Networks and Intelligent Systems. INISCOM 2018. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 257.2019, Springer.

BƯỚC ĐẦU ỨNG DỤNG MÁY HỌC NHẬN DIỆN BỆNH DỰA TRÊN MÔ TẢ TRIỆU CHỨNG CỦA TÔM

Quách Luyl Đa, Quách Hoàng Long, Dương Trung Nghĩa, Nguyễn Chí Ngôn

TÓM TẮT: *Nuôi tôm là một ngành mũi nhọn trong phát triển kinh tế tại các tỉnh Đồng bằng sông Cửu Long. Tuy nhiên trong nuôi tôm còn gặp nhiều vấn đề nhất là bệnh trên tôm gây thiệt hại đáng kể. Bệnh trên tôm được thể hiện thông qua các triệu chứng và thể hiện trên cơ thể của tôm. Nhận thấy được tầm quan trọng của triệu chứng tôm giúp cảnh báo tốt sớm. Trong nghiên cứu này, nhóm tác giả đã áp dụng thuật toán phân lớp văn bản bằng các mô hình máy học là Logistic Regression, Random Forest, Naïve Bayes, máy học véc tơ hỗ trợ (SVM), Multi-layer Perceptron (MLP) trên tập dữ liệu gồm 1098 mẫu với 14 lớp. Một số kịch bản đánh giá đã được tiến hành gồm xử lý tách từ và mô hình so sánh trên bộ dữ liệu thu được với tỉ lệ khác nhau. Kết quả cho thấy thuật toán máy học véc tơ hỗ trợ có độ chính xác cao nhất (81.27%), tiếp theo là MLP, Random Forest, Logistic Regression và Naïve Bayes. Thông qua kết quả nghiên cứu, có thể áp dụng các thuật toán máy học để chẩn đoán bệnh tôm hoàn toàn dựa trên mô tả triệu chứng văn bản.*

Từ khóa: *Chuẩn đoán bệnh tôm, phân lớp văn bản, so sánh các thuật toán máy học, áp dụng phân lớp ảnh trong chẩn đoán bệnh tôm.*