Identifying Gender of Internet Users Based on Access History

Trung-Hieu Le, PhD
Dai Nam University, Vietnam
hieult@dainam.edu.vn
Nguyen Thi Hang, PhD
Thai Nguyen University, University of Information and Communication Technology, Vietnam
nthnag@ictu.edu.vn
Dinh Tran Ngoc Huy, MBA., (corresponding)
Banking University HCMC, Ho Chi Minh city Vietnam - International University of Japan, Japan.
Dtnhuy2010@gmail.com
Nguyen Thi Phuong Thanh, Master
Thai Nguyen University of Information and Communication Technology, Vietnam.
ntpthanh@ictu.edu.vn
Nguyen Thuy Dung, Master
Thai Nguyen University of Information and Communication Technology, Vietnam.
ntdung@ictu.edu.vn

Abstract:

The use of activities and internet access differs between men and women. On average, men spend more time on the Internet a day. Men also have some of the same online activities as women. However, there are specific differences such as men's tendency to access features such as breaking news, football, or games and men's products. On the contrary, women are more interested in shopping, e-commerce, chatting and participating in social networking sites and blogs. The study aims to identify and predict gender of internet users based on their access history.

With SVM method, the correct classification rate is the highest compared to the other two models Accuracy = 87.67%, in addition, the Precision, Recall, and F-Score parameters also give outstanding rates. This result allows us to believe in the ability of the SVM machine learning model to effectively handle the classification and gender identification problem with large-dimensional data.

Key words: internet users, gender, access history, technology

1. Introduction

Today, with the continuous development of science and technology in the world in general and in Vietnam in particular, there have been great strides. The infrastructure and equipment are relatively modern and constantly developing. According to the summary report of the Ministry of Information and Communications in 2016, the percentage of Internet users in Vietnam reached 62.76% of the population, of which the percentage of households with Internet access reached 24.38%, that is for every 5 families. One household uses fixed broadband. In which, according to statistics of the Department of Telecommunications (Ministry of Information and Communications) in November 2016, the total number of fixed broadband Internet subscribers reached more than 9 million subscribers and the number of mobile broadband subscribers reached more than 12.6 million subscribers. .
Besides, according to the statistics of "wearesocial.net", in January 2015, Vietnamese people ranked 4th in the world in terms of time spent using the Internet with 5.2 hours per day, only after the Philippines ranked first with 6 hours per day. hours, followed by Thailand with 5.5 hours, and Brazil with 5.4 hours/day.

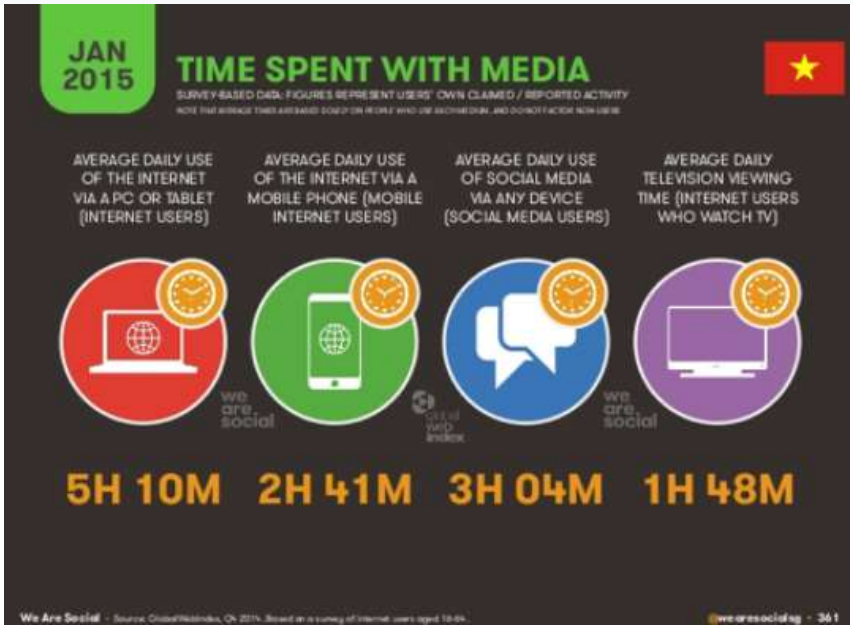


Figure 1 - Average time using the internet per day of Vietnamese people

Because of the continuous development of information technology and the popularity of the internet today, information to users is extremely rich and continuous. Internet users today often have the habit of accessing and searching for the issues they are interested in. Most of the information is saved as a session on the network. Such information can be articles, business documents, products, economic information, e-commerce, other personal information, etc. From that fact emerged the needs. require information analysis to classify that information for different purposes such as learning, research, business, commercial marketing...

With that in mind, we must identify and categorize useful information from rich data sources from users' sessions, internet usage, to suit specific audiences, and to support specific technologies. automation tools that aid in knowledge discovery and information mining. Gender Prediction (or Determination Gender or Gender Prediction) is a method of classifying and identifying activities accessed by either Male gender or Female gender from other activities with known labels. For example, an article in a web page can be accessed by either male or female gender (such as sports, education, law, information technology, cosmetics, clothing, etc.). The classification can be done manually: read the content of each activity and assign it to a certain label. However, for a system with a lot of records, this method will take a lot of time and effort. Therefore, it is necessary to have an automatic method for gender classification. This method helps to determine gender with high accuracy and is used for purposes such as learning, research, business, and commercial marketing.

2. Literature review

Available methods of sex determination

In the world, a number of previous works have studied methods based on text analysis such as De Vel et al. (2001) used 221 features to identify the author of the email. Argamon and Koppel et al. (2003) studied differences in male and female writing styles in 604 British National Corpus documents. Schler et al. (2008) explores the use of features and content-based to predict the gender and age of bloggers on a dataset of more than 71,000 blog posts from blogger.com. This model achieved 80% results for gender predictions and 76% for age predictions. Nguyen et al. conducted a study to predict gender and age of twitter messages and forum posts using regression method with about 80% accuracy. (source: Dong Nguyen, Rilana Gravel, Theo Meder, Dolf Trieschnigg – TweetGenie: Automatic Age Prediction From Tweets. Available from: http://dolf.trieschnigg.nl/papers/SIGWEB.2013.nguyen.pdf)

Method of determining gender using blog posts

In the past years, Blog is a kind of diary, popular personal website sharing life experiences or something in people's daily life. This is a very large type of data containing articles and texts created by hundreds of thousands of user authors. This information contains many features that can be exploited for the classification problem, specifically here is the gender determination of bloggers. A research paper specifically on demographics and gender was developed in 2007 by Schler et al (2008) with a dataset of all blogs visited in a single day in August 2004.

The study focused on differences in blogging and differences between men and women among bloggers of different ages. The stylistic and content features are given as the core to solve the problem.

Research using MCRW (Multi-Class Real Winnow) model. For each class, ci, i = 1, ..., m, wi is a weight vector <wi 1, ..., wi n>, where n is the size of the feature set. Each wi j, is initialized starting with 1. The training sets are randomly ordered and processed once. The

algorithm runs a continuous training loop, randomly resetting the examples after each cycle. After every ten cycles, the Algorithm checks the number of correctly classified training examples. If this number has decreased, the algorithm will roll back. If no improvement is found after five rounds of 10 cycles, the algorithm is terminated. Research shows that MCRW model is more efficient than SVM in classifying a large number of documents.

The test results show that bloggers can be classified by gender according to age groups, writing style and content. In the cases given, the combination of stylistic and content features provides the best classification accuracy.

The method of sex determination through data from daily mobile communications was studied according to the article Demographic Prediction Based on User's Mobile Behaviors (2012) in the MDC Data Set competition. In this paper, the research team proposes a new model namely Multi-Level Classification Model to solve the problem of existing unbalanced classes in the data. Based on this model, the user's gender prediction results will be obtained by combining multiple classification models into a multi-level structure.

3. Research Methodology

To conduct a general gender determination classification, we will perform the following steps:

Step 1: Build a training dataset based on pre-classified user data set. Conducting learning for the data set, processing and collecting the data of the learning process are distinct features for each content.

Step 2: The data to be classified is processed, and the features combined with the previously learned feature are drawn to classify and produce results.

The outstanding feature of this problem is the diversity of activities and characteristics of men and women. The features make the classification relative and somewhat subjective, which, if done by humans, can be prone to ambiguity. For example, there is access to clothing shopping information at an e-commerce website, which can still be accessed by men or women.

In recent years, the classification method using Support Vector Machine (SVM) has been interested and used a lot in the fields of identification and classification. The SVM method was born from the statistical learning theory built by Vapnik and Chervonenkis and has great potential for development in theory as well as in practical applications. Practical tests show that the SVM method has good classification ability for 2-class and multi-class classification problems as well as in many other applications (such as text classification by topic, face detection). in images, regression estimation, software error prediction, etc.). Compared with other classification methods, the classification ability of SVM is equivalent or significantly better. For these reasons, I have chosen this method for predicting the gender of internet users, the specific algorithm and application will be presented in the following chapters.

4. Main findings

4.1.The main steps of the SVM . method

Data preprocessing: The SVM method requires data to be expressed as vectors of real numbers. So if the input is not real, then we need to find a way to convert it to SVM's digital form. Avoiding too large numbers, it is usually advisable to normalize the data to convert to the range [-1,1] or [0,1].

-Select the multiplication function: It is necessary to choose the corresponding multiplication function for each specific math problem to achieve high accuracy in the classification process.

-Perform cross-checking to determine application parameters.

-Use parameters for training the sample set.

-Testing the Test dataset.

Advantages of SVM method in data classification

As is known, data classification is a process of putting unlabeled data into corresponding labeled data classes. Each label is identified by some sample data set of that label. To perform the classification process, training methods are used to build a classifier set from sample records, then use this classifier to predict the class of new records with unknown labels.

We can see that two-class classification algorithms such as SVM all have the common feature of requiring data to be represented in the form of feature vectors, but other algorithms must use parameter estimation and the optimal threshold while the SVM algorithm can find these optimal parameters on its own. Among the methods, SVM is the method that uses the largest feature vector space (more than 10,000 dimensions) while other methods have much smaller dimensions (such as Naïve Bayes 2000, k-Nearest Neighbors is 2415…).

In his work in 1999, Joachims compared SVM with Naïve Bayesian, k-Nearest Neighbour, Rocchio, and C4.5 and in 2003 Joachims demonstrated that SVM works very well with the characteristics previously mentioned properties of the data set. The results show that SVM gives the best classification accuracy when compared with other methods. (source: Making Large-Scale SVM Learning Practical - Thorsten Joachims. Available from: https://www.cs.cornell.edu/people/tj/publications/joachims_99a.pdf)

According to Xiaojin Zhu, in the works of many authors (such as Kiritchenko and Matwin in 2001, Hwanjo Yu and Han in 2003, Lewis in 2004) have shown that the SVM algorithm gives good results of text classifier.

4.2 Experimental results

Experimental results using 5 models and 4 evaluation criteria to provide effective SVM machine learning model for gender classification. The evaluated models are recorded after using the method of Cross validation or k-fold Cross validation for testing and training. The obtained results show that the classification ability has high accuracy but decreases gradually with discrete models from category A model to product category model D.

The reason is because the data noise is quite large for discrete models, the more features the model has, the larger the noise. Specific results with discrete models are collected in the tables below:

Table 1 - Discrete model results

LABELS	SVM with model A			
	Precision	Recall	F-Score	Accuracy
Male	77.4 %	55.3 %	64.5 %	**86.51 %**
Female	88.2 %	95.4 %	91.7 %	
Weighted Avg	85.8 %	86.5 %	85.6 %	

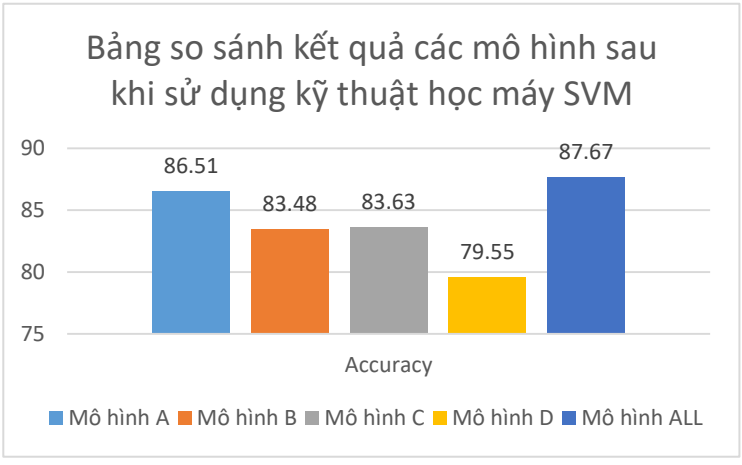
(source: authors calculation)

Table 2 is the table of main results obtained from the SVM machine learning model when using and combining all the features and discrete models of the normalized data set and giving the evaluation criteria. Compared with the results of the four discrete models above, the prediction rate when combining features together gives an accuracy rate of 87.67%. From the above experiments, it is shown that SVM with very high accuracy classifier can meet the requirements of the gender prediction problem.

LABELS	SVM with all features			
	Precision	Recall	F-Score	Accuracy
Male	79.4 %	59.3 %	67.9 %	**87.67 %**
Female	89.3 %	95.7 %	92.4 %	
Weighted Avg	87.1 %	87.7 %	87 %	

***Table 2* Results from model SVM**

Chart 1 - Model accuracy:



Comparison with some other methods

To further evaluate the performance of the predictive model, the thesis has trained the dataset on other popular machine learning models, NaiveBayes and RandomTree, the specific results are given in Tables 3 and 4.

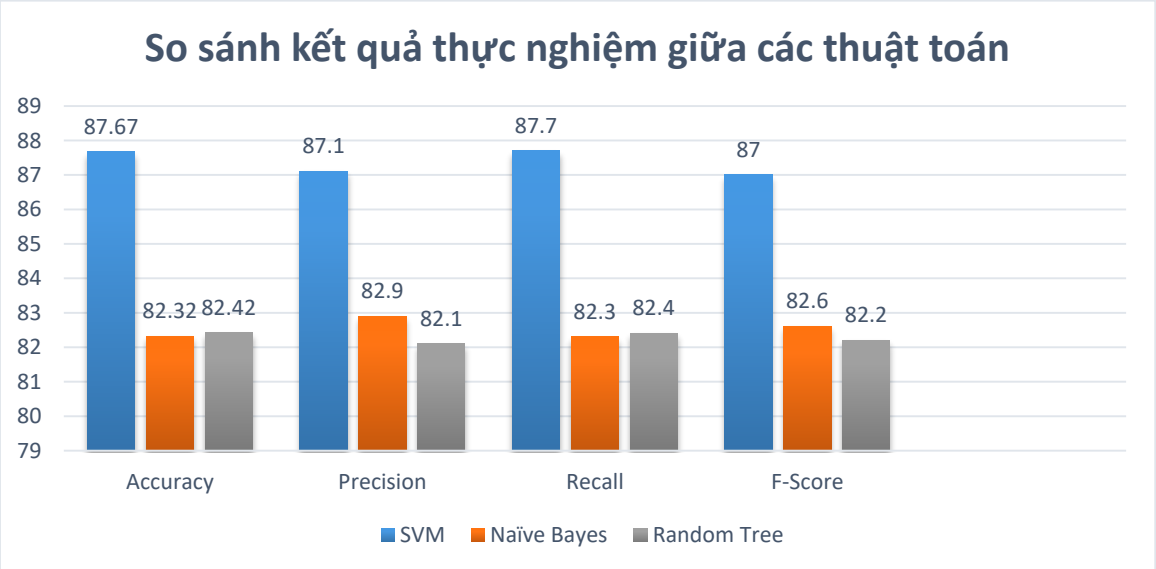
LABELS	NaiveBayes			
	Precision	Recall	F-Score	Accuracy
Male	59 %	64.3 %	61.5 %	82.32 %
Female	89.7 %	87.4 %	88.5 %	
Weighted Avg	**82.9 %**	**82.3 %**	**82.6 %**	

***Table 3*- Results from model Naïve Bayes**

NHÃN	Random Tree			
	Precision	Recall	F-Score	Accuracy
Nam	60.7 %	57 %	58.8 %	82.42 %
Nữ	88.1 %	89.6 %	88.8 %	
Weighted Avg	**82.1 %**	**82.4 %**	**82.2 %**	

Table 4- Results from the Random Tree . model

To make it easier to visualize, let's look at the following chart:



Comments: Based on tables 2-3-4 summarizing results of sex classification on SVM, NaiveBayes, RandomTree models, we find that NaiveBayes gives the lowest results when classifying despite its ability to give accuracy. quite high with Accuracy = 82.32% but the reality is still not optimal. Random Tree is better, but the classification rate is only 0.1% more

than NaiveBayes. With SVM, the correct classification rate is the highest compared to the other two models Accuracy = 87.67%, in addition, the Precision, Recall, and F-Score parameters also give outstanding rates. This result allows us to believe in the ability of the SVM machine learning model to effectively handle the classification and gender identification problem with large-dimensional data.

5. Discussion

Introduction to SVM

Support Vector Machines (SVM) is a classification method derived from statistical learning theory, based on the principle of structural risk minimization. SVM will try to find a way to classify the data so that the error that occurs on the test set is minimized (Test Error Minimisation). In the early days when SVM appeared, the computing power of computers was very limited, so the SVM method was not considered. However, from 1995 onwards, the algorithms used for SVM developed very quickly, along with the powerful computing power of the computer, had great applications.

a.Ideas

Given a training set, represented in vector space, where each document is a point, this method finds a decision hyperplane f that can best divide the points on this space into two distinct classes. class "+" and class "-" respectively. The quality of this hyperplane is determined by the distance (called the boundary) of the nearest data point of each layer to this plane. Then, the larger the boundary distance, the better the decision plane, and the more accurate the classification. Its idea is to map (linear or nonlinear) data into a space of feature vectors where an optimal hyperplane is found to separate data belonging to two different classes.

The purpose of the SVM method is to find the maximum boundary distance:

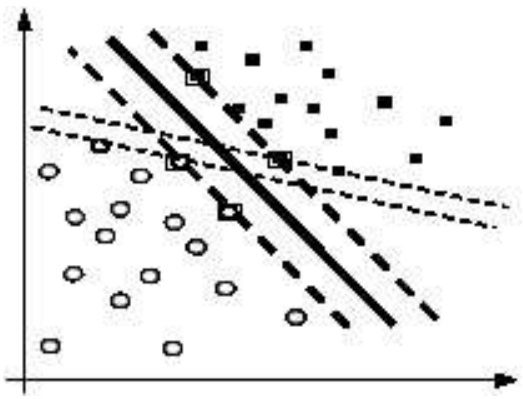


Figure 2 - Description of the SVM . method

The shaded line is the best hyperplane, and the points bounded by the rectangle are the points closest to the hyperplane, they are called support vectors. The dashed lines on which the support vectors lie are called margins.

Theoretical basis

SVM is essentially an optimization problem, the goal of this algorithm is to find a space F and the decision hyperplane f over F so that the classification error is minimal.

Given sample with $D = \{(x_1, y_1), (x_1, y_1), \dots, (x_l, y_l)\}$ with $x_i \in \mathbb{R}^n$, belong to 2 classes $y_i \in \{-1, 1\}$ ar corresponding class labels of the x_i (-1 denote class I, 1 denote class II).

We have, the hyperplane equation contains the vector $\overrightarrow{x_i}$ in space:

$$\overrightarrow{x_i}.\vec{w} + b = 0$$

Put:

$$f(\overrightarrow{x_i}) = sign(\overrightarrow{x_i}.\vec{w} + b) = \begin{cases} +1, \overrightarrow{x_i}.\vec{w} + b > 0 \\ -1, \overrightarrow{x_i}.\vec{w} + b < 0 \end{cases}$$

Hence , $f(\overrightarrow{x_i})$ show class classification $\overrightarrow{x_i}$ int 2 groups as mentioned above.

We said $y_i = +1$ If $\overrightarrow{x_i}$ belong to class I and $y_i = -1$ If $\overrightarrow{x_i}$ belong to class II.

6. Conclusion

In general, the most frequently performed activity on the Internet by users is to gather information, such as reading news or using search sites. More than 90% of Internet users have used search sites, about half of them even use it every day. The Internet is also used for research for school or work by half of all people using the Internet once a week or more often. With new interactive websites and online applications, users not only have the opportunity to find information, but also contribute their own piece of content.

Currently, the number of visits to e-commerce has grown significantly. Most popular sites are auction and sale sites, where 40% of users have visited. Online banking is still in its infancy. The use of online shopping and online banking websites has grown tremendously over the past few years.

In this study, authors has outlined how to describe data and normalize data PAKDD'15 data is used in the thesis. Representation of product category and access time characteristics of internet users to generate training data for inclusion in specific classifier toolkits, LibSVM and Weka. Experimental results are shown in 4 subclassification models and 1 overall classification model combined with 4 evaluation criteria to show the appropriateness of SVM machine learning technique when applied to the problem.

Testing and evaluation results are conducted after training the dataset according to 3 models. Particularly for the SVM model, there is a grid.py tool to help select the optimal parameters. The obtained results show that SVM gives better classification results than NaiveBayes and Random Tree with an accuracy of over 87%.

Research limitation

Due to time constraints, the comparison between other machine learning techniques models only gives the SVM model with all features and two training models, Naïve Bayes and Random Tree.

And The study is based on available data, the data set has a gender imbalance when the number of females is more than the number of males.

The experimental results achieved are not really good compared to expectations.
Data processing speed is still slow when the data set is large

Acknowledgement
Thank you editors, friends and brothers to assist this publishing.

References

[1] Argamon, S., Koppel, M., Fine, J. and Shimoni, A. (2003). Gender, Genre, and Writing Style in Formal Written Texts, Text 23(3), August.
[2] Argamon, S., Koppel, M., Pennebaker, J. and Schler, J. (2008). Automatically Profiling the Author of an Anonymous Text, Communications of the ACM.
[3] Chang, C.C., Lin, C.J, 2001. LIBSVM – a library for support vector machines http://www.csie.ntu.edu.tw/~cjlin/libsvm
[4] De Vel, O., Anderson, A., Corney, M., Mohay, G. M. (2001). Mining e-mail content for author identification forensics. SIGMOD Record 30(4), pp. 55-64.
[5] Do Viet Phuong and Tu Minh Phuong. "*Gender Prediction Using Browsing History".* KSE (1) 2013: 271-283.
[6] Duong Thi Tinh, Nguyen Thu Thuy, Dinh Tran Ngoc Huy. (2021). Doing Business Research and Teaching Methodology for Undergraduate, Postgraduate and Doctoral Students-Case in Various Markets Including Vietnam , Elementary education online, 20(1).
[7] Dinh Tran Ngoc Huy, Nguyen Thi Hang. (2021). Factors that affect stock price and Beta CAPM of Vietnam Banks and Enhancing Management infomation system -Case of Asia Commercial Bank, Revista geintec Inovacao E Tecnologias, 11(2).
[8] Dinh Tran Ngoc Huy, Pham Ngoc Van, Nguyen Thi Thu Ha. (2021). Education and computer skill enhancing for Vietnam laborers under industry 4.0 and evfta agreement, Elementary education online, 20(4).
[9] Dinh Thi Hien, Dinh Tran Ngoc Huy, Nguyen Thi Hoa. (2021). Ho Chi Minh Viewpoints about Marxism Moral Human Resource for State Management Level in Vietnam , Psychology and education, 58(5).
[10] Dinh Tran Ngoc Huy. (2021). Banking sustainability for economic growth and socio-economic development–case in Vietnam , Turkish Journal of computer and mathematics education, 12(2).
[11] Dong Nguyen, Rilana Gravel, Theo Meder, Dolf Trieschnigg – TweetGenie: Automatic Age Prediction From Tweets. Available from: http://dolf.trieschnigg.nl/papers/SIGWEB.2013.nguyen.pdf
[12] Hac, L.D., Huy, D.T.N., Thach, N.N., Chuyen, B.M., Nhung, P.T.H., Thang, T.D., Anh, T.T. (2021). Enhancing risk management culture for sustainable growth of Asia commercial bank - ACB in Vietnam under mixed effects of macro factors , Entrepreneurship and Sustainability Issues, 8(3).
[13] Hang, T.T.B., Nhung, D.T.H., Hung, N.M., Huy, D.T.N., Dat, P.M. (2020). Where Beta is going–case of Viet Nam hotel, airlines and tourism company groups after the low inflation period , Entrepreneurship and Sustainability Issues, 7(3).
[14] Huy, D.T.N. (2015). The Critical Analysis of Limited South Asian Corporate Governance Standards After Financial Crisis, International Journal for Quality Research, 9(4): 741-764.

[15] Huy, D.T.N. (2012). Estimating Beta of Viet Nam listed construction companies groups during the crisis , Journal of Integration and Development,15 (1), 57-71

[16] Huy, D. T.N., Loan, B. T., and Anh, P. T. (2020). Impact of selected factors on stock price: a case study of Vietcombank in Vietnam, Entrepreneurship and Sustainability Issues, vol.7, no.4, pp. 2715-2730. https://doi.org/10.9770/jesi.2020.7.4(10)

[17] Huy, D. T.N., Dat, P. M., và Anh, P. T. (2020). Building and econometric model of selected factors' impact on stock price: a case study, Journal of Security and Sustainability Issues, vol.9(M), pp. 77-93. https://doi.org/10.9770/jssi.2020.9.M(7)

[18] Huy D.T.N., Nhan V.K., Bich N.T.N., Hong N.T.P., Chung N.T., Huy P.Q. (2021). Impacts of Internal and External Macroeconomic Factors on Firm Stock Price in an Expansion Econometric model—A Case in Vietnam Real Estate Industry, Data Science for Financial Econometrics-Studies in Computational Intelligence,vol.898, Springer. http://doi-org-43.webvpn.fjmu.edu.cn/10.1007/978-3-030-48853-6_14

[19] Huy, D.T.N. , An, T.T.B. , Anh, T.T.K. , Nhung, P.T.H. (2021). Banking sustainability for economic growth and socio-economic development – case in Vietnam, Turkish Journal of Computer and Mathematics Education, 12(2), pp.2544–2553

[20] Huy, D.T.N. , An, T.T.B. , Anh, T.T.K. , Nhung, P.T.H. (2021). Banking sustainability for economic growth and socio-economic development –case in Vietnam, Turkish Journal of Computer and Mathematics Education, 12(2), pp.2544–2553

[21] Hu, J., Zeng, H.-J., Li, H., Niu, C., Chen, Z. (2007) "*Demographic prediction based onuser's browsing behavior*", Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada. [viewed 05.09.2016] Available from: http://wwwconference.org/www2007/papers/paper686.pdf

[22] Josh Jia-Ching **Ying**, Yao-Jen Chang, Chi-Min Huang and Vincent S. Tseng (2012) – Demographic Prediction Based on User's Mobile Behaviors. Available from: http://www.idiap.ch/project/mdc/publications/files/mdc-final241-ying.pdf

[23] Kabbur, S., Han, E.-H., Karypis, G. (2010) "*Content-based methods for predicting website demographic attributes*", University of Minnesota Supercomputing Institute Research Report UMSI 2010/98 [viewed 06.09.2016] Available from: http://www.dtc.umn.edu/publications/reports/2010_01.pdf

[24] Le, Trung-Hieu, Thanh-Hai Tran, and Cuong Pham. The Internet-of-Things based hand gestures using wearable sensors for human machine interaction. International Conference on Multimedia Analysis and Pattern Recognition (MAPR). IEEE, 2019.

[25] Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. Available from: https://academic.oup.com/biomet/article-abstract/62/1/207/220350/Mendenhall-s-studies-of-word-length-distribution

[26] Nguyen Thi Hang, Dinh Tran Ngoc Huy. (2021). Better Risk Management of Banks and Sustainability-A Case Study in Vietnam , Revista geintec Inovacao E Tecnologias, 11(2).

[27] Nguyen Thi Hoa, Nguyen Thi Hang, Nguyen Thanh Giang, Dinh Tran Ngoc Huy. (2021). Human resource for schools of politics and for international relation during globalization and EVFTA , Elementary education online, 20(4).

[28] Nguyen, D., Gravel, R., Trieschnigg, D., and Meder, T. (2013). "How old do you think i am?"; a study of language and age in twitter. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. Available from: http://www.dongnguyen.nl/publications/nguyen-icwsm2013.pdf

[11] PENG Qiu-fang, LIU Yang – Research of gender prediciton based on SVM with E-commerce data. Available from: http://lxbwk.njournal.sdu.edu.cn/EN/abstract/abstract3503.shtml

[12] Pham Minh Dat, Nguyen Duy Mau, Bui Thi Thu Loan, Dinh Tran Ngoc Huy. (2020). Comparative China corporate governance standards after financial crisis, corporate scandals and manipulation, Journal of security & sustainability issues, 9(3).

[13] Pham Van Hong, Huynh Xuan Nguyen, Dinh Tran Ngoc Huy, Le Thi Viet Nga, Nguyen Thi Ngoc Lan, Nguyen Ngoc Thach, Hoang Thanh Hanh.(2021). Sustainable bank management via evaluating impacts of internal and external macro factors on lending interest rates in Vietnam, Linguistica Antverpiensia, Issue 1, pp.76-87.

[14] Phung Tran My Hanh, Nguyen Thi Hang, Dinh Tran Ngoc Huy, Le Ngoc Nuong. (2021). Enhancing Roles of Banks and the Comparison of Market Risk and Risk Policy Implications in Group of Listed Vietnam Banks During 2 Stages: Pre and Post-Low Inflation Period , Revista geintec-gestao Inovacao e Tecnologias, Vol.11(2).

[29] Speltdoorn, S. (2010) "*Predicting demographic characteristics of web users using semisupervised classification techniques*" Master's dissertation, Ghent University, Faculty of Economucs and Business Administration. [viewed 14.09.2016] Available from: http://lib.ugent.be/fulltxt/RUG01/001/459/756/RUG01001459756_2011_0001_AC.pdf

[30] Quanzeng You, Sumit Bhatia, Tong Sun, Jiebo Luo (2014) "*The eyes of the beholder: Gender prediction using images posted in Online Social Networks*". Available from: http://www.cs.rochester.edu/u/qyou/papers/gender_classification.pdf

[31] Weka - Available from: http://www.cs.waikato.ac.nz/ml/weka/

[32] Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, Nitesh V. Chawla (2014) "*Inferring User Demographics and Social Strategies in Mobile Social Networks*". Available from: http://www3.nd.edu/~ydong1/papers/KDD14-Dong-et-al-WhoAmI-demographic-prediction.pdf

[33] Yan, X., Yan, L.: Gender classification of weblogs authors. In: Proceedings of the AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, Stanford, CA, March 27-29, pp. 228–230 (2006). Available from: http://aaaipress.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-046.pdf

[34] Ying, J.J.C., Chang, Y.J., Huang, C.M., Tseng, V.S. (2012). Demographic prediction based on users mobile behaviors. Mobile Data Challenge. Available from:http://www.idiap.ch/project/mdc/publications/files/mdc-final241ying.pdf

[35] Zhang, C., Zhang, P. (2010). Predicting gender from blog posts. Technical report, Technical Report. University of Massachusetts Amherst, USA.