International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-13 Issue-10, September 2024

Classification of Vietnamese Reviews on E-Commerce Platforms

Phan Thi Ha, Trinh Thi Van Anh

Abstract: ***The research team used machine learning models to classify Vietnamese reviews on products on the e-commerce platform as positive or negative. To classify and evaluate the effectiveness of Support Vector Machine (SVM), Random Forest, Logistic Regression machine learning models on different platforms, the authors have built their own training and test data sets as well as a set of stopwords to classify Vietnamese web reviews [9]. This can then be applied to building a webapp that allows entering a link of any online products and then categorizing its user reviews, helping sellers evaluate their products/services, understand consumer behavior and make changes, improvements to the products accordingly.***

Keyword: Text Classification, SVM, Random Forest, Logistic Regression, CNN.

I. INTRODUCTION

In this modern age, the need to update and use information is an essential part of every person's daily activity. The role of information is eminently evident in every aspect of modern life including work, study, business, e-commerce and especially scientific research around the world… In Vietnam, with the emergence of information technology in recent years, the need to read newspapers, search, and shop on the internet has become a daily habit. This is thanks to various superior features that internet information provides: compact storage, long storage time, quick search and convenience in information exchange. E-commerce platforms are a method of exchanging, buying and selling that is very familiar to many people over the internet. A prominent feature of shopping on e-commerce platforms is that buyers can leave reviews of products on the seller's site. These reviews will help sellers evaluate their products/services, understand consumer behavior and make changes, improvements to the products accordingly. In addition, reviews from previous buyers help future customers create an overview of all aspects of the product from quality, appearance to related services, which will greatly influence their purchasing decisions after.

Manuscript received on 01 August 2024 | Revised Manuscript received on 07 August 2024 | Manuscript Accepted on 15 September 2024 | Manuscript published on 30 September 2024.
* Correspondence Author
Dr. Phan Thi Ha*, Lecturer, Faculty of Information Technology at Posts and Telecommunications Institute of Technology (PTIT), Ha Noi, Vietnam, and Computing Fundamental Department, FPT University, Hanoi, Viet Nam. Email ID: hapt@ptit.edu.vn, hapt37@fe.edu.vn and hathiphan@yahhoo.com, ORCID ID: 0009-0003-2521-0717
Trinh Thi Van Anh, Lecturer, Faculty of Information Technology at Posts and Telecommunications Institute of Technology (PTIT) in Ha Noi, Vietnam, and Computing Fundamental Department, FPT University, Hanoi, Viet Nam. Email ID: vanh22@yahoo.com, anhttv@ptit.edu.vn and anhttv20@fe.edu.vn, ORCID ID: 0009-0005-8062-2014

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

Therefore, analyzing product reviews brings many benefits to both buyers and sellers on e-commerce platforms. However, with the development of e-commerce platforms comes huge number of products reviews and products, some of which have up to several thousand reviews. Manually evaluating each review is not feasible, which is where machine learning and deep learning models are applied to replace manual labor. The purpose of the article is to use machine learning and deep learning models to classify Vietnamese reviews on phone products, tablets, e-readers, etc. as positive or negative. pole. To classify and evaluate the effectiveness of Support Vector Machine (SVM), Random Forest, Logistic Regression, CNN models [1,2,3,4,5,6,7] [10] [11] [12] [13] [14], the authors have built their own training and test data sets as well as a set of stopwords to classify Vietnamese web reviews.

II. BUILDING DATA

A. Building a Review Classification Model

The model Figure [1] built by the authors describes the review classification process, including basic steps such as: Data collection, data labeling, data preprocessing, feature vector extraction, model training, model evaluating, and saving the trained model to use in predicting new inputs.

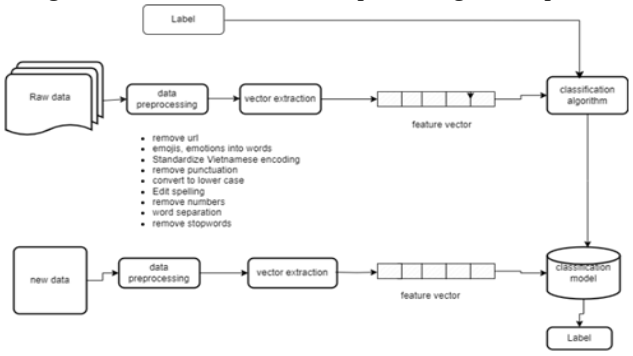

Figure 1. Review Classification Model

Raw data was collected from 18,604 reviews from a electronic platform [9], a mobile device retail chain established in 2024 with the number 1 market share in Vietnam.

B. Labeling Data

Reviews on the website [9] come with the users' scores for the product (lowest is 1 star, highest is 5 stars). However, some highest scorers have negative reviews and some of the lowest scorers highly praise the products. Therefore, we only took the review content and then proceeded to assign labels to each review.

© Copyright: All rights reserved.

The collected data set includes 18,604 unlabeled reviews. After obtaining the data set, we will assign labels: positive or negative to each review. The labels assigned are based on personal opinions. After labeling, the data set now has 14544 positive labels and 4060 negative labels.

C. Data Preprocessing

Raw data from the dataset needs to go through data preprocessing before being analyzed using machine learning algorithms. User reviews in text form, without specific structure, will be preprocessed according to Figure [2]

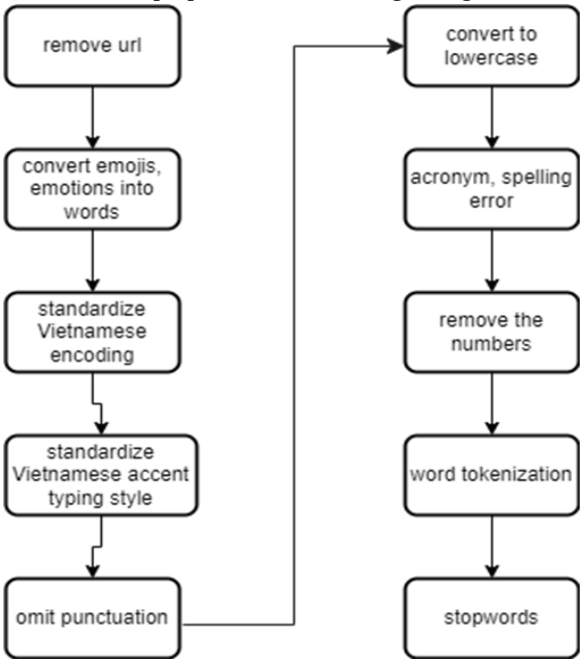

Figure 2. Preprocessing Steps

i.Removing URL

Uniform Resource Locator (url) is a link or address that refers to a resource page on the internet. During product reviews, some customers include url links in their comments. These links do not carry information about the reviewer's feelings about the products. In addition, url paths often contain characters and parameters that are meaningless to humans, which will affect the accuracy of the model. Therefore, the authors removed these url paths to help reduce redundant information and reduce the size of the feature vector.

ii.Converting Emojis, Emoticons Into Words

Emojis, emoticons are images, symbols of animals, objects... expressing emotions. These emojis often appear in text messages, websites, social networks... Emoji help convey the writer's emotional messages succinctly; so to avoid losing this emotional information, the authors replaces emojis and emoticons with corresponding emotion words. For example, ":)" is replaced by "laugh happily", "X-D" is replaced by "laugh loudly", "(;_;)" is replaced by "sad and cry"... This replacement is done by building encoding Unicode dictionaries of emojis and emoticons with their corresponding replacement words.

iii.Standardizing Vietnamese Unicode Code Set

Currently, the Unicode code set is most commonly used in text editing. When editing Vietnamese documents, there are two most popular Unicode code sets: built-in Unicode and combined Unicode. To display the letter "á", the built-in Unicode code set will build the letter "á" in the code table, while the combined Unicode code table will combine the letter "â" and the tone. On screen, the two letters "á" appear the same, but the computer identifies these as two different words. The built-in Unicode code set is more popular and widely used, however there are still some people and devices that use the combined Unicode. It is unavoidable to collect text from both sets of codes during the process of collecting data from the internet. Therefore, the project unifies the code set for the entire data set by replacing the combined Unicode code set with the built-in Unicode code set. Sharing the same Unicode code set helps to avoid cases where the same word is encoded differently, resulting in an increase in the size of the feature vector. The way to unify the Unicode is to build a dictionary that replaces combined Unicode with the corresponding built-in Unicode and includes vowels in Vietnamese and tones (high-rising, low-falling, high-broken, low-rising and heavy tones).

iv.Standardizing Vietnamese Diacritic Typing Style

In Vietnamese there are 5 tone marks: high-rising, low-falling, high-broken, low-rising and heavy tones. The placement of these tone marks also has clear rules. Currently, there are two ways to place these diacritics: "old style" and "new style". With the "old style" diacritics, there are some rules as follows: tone mark is placed on the vowel for syllables with one vowel, on first vowel for syllables with two vowels, on second vowel for syllables with three vowels or two vowels and one consonant. Lastly, for syllables with "ê" or "ơ" vowels, tone mark is placed on those two vowels. According to the "new style" diacritics, the rules are as follows: tone mark is placed on the vowel for syllables with one vowel, on vowels with diacritic marks (ă, â, ơ, ư…) for syllables with those vowels, on second vowel for syllables with two vowels ending in a consonant and lastly, on last vowel for syllables ending in "oa", "eo", "uy".

Table [1] examples of two different ways to place diacritics mentioned above

Table 1. New and Old Diacritic Typing Styles

Old Style	New Style
òa, óa, ỏa, õa, ọa	oà, oá, oả, oã, oạ
òe, óe, ỏe, õe, ọe	oè, oé, oẻ, oẽ, oẹ
ùy, úy, ủy, ũy, ụy	uỳ, uý, uỷ, uỹ, uỵ

In the data set, there are two ways of placing diacritics; for example, "hủy" and "huỷ", "hỏa" and "hoả" are the same word but the computer identifies them as different words because of different diacritic placements. Therefore, authors unified way to place diacritics across the data set to avoid word repetition.

v.Removing Punctuation and Special Characters

Punctuation is a grammatical device in Vietnamese, an indispensable part when creating documents. Common punctuation marks include period (.), comma (,), question mark (?), exclamation mark (!), etc. However, punctuation marks do not carry expressive information and also increase the size of the data set.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

For example, “like” and “like!” have the same meaning, but the computer identifies them as two different words, which results in two different values when extracted into a feature vector. The authors removed punctuation marks and special symbols (for special characters that replace words, replace the characters with the corresponding correct words) to reduce the vocabulary size of the data set and reduce the dimensional feature vector.

vi.Converting to all Lowercase

Converting all letters in a data set to lowercase is a technique frequently applied to preprocessing text data. This is because computer recognizes “tốt”, “Tốt” and “TỐT” as different words while they are the same. To avoid this, authors converted all letters in the data set to lowercase (can also convert to all uppercase). Converting all letters to lowercase helps avoid word repetition in the feature vector, reduces the dimensionality of the vector, and increases accuracy when calculating the TF-IDF value of each word.

vii.Spelling out Abbreviations and Correcting Some Simple Spelling Errors

In the data set, there are lots of reviews like “sp ok, chất lượng, phù hợp giá tiền, hàng như mong đợi, nv giao hàng nhiệt tình”, “Sp rất tốt”, “Đt e đang xài tự nhiên nó xanh mh”, “Máy bập một lúc ms lên (từ lúc mua về đã thế)”, “Sản phẩm chất lượng, giao hàng nhanh,phù hợp với giá tiền,nên mua nhaaa”,... These reviews are very common online, because:

- When typing, many people want to write quickly or do not want to write down all the words they need to write, so they tend to abbreviate some words. For example, some common abbreviations in the data set are: no - ko (không), phone - đt (điện thoại), product - sp (sản phẩm), customer service - cskh (chăm sóc khách hàng) …
- During the product review process, customers might misspell words, type the wrong key... leading to misspellings when writing. Some common misspelled words in the data set are: sản phẩn (sản phẩm), xạc pin (sạc pin), sài (xài) …
- Some customers, when giving a review, want to emphasize an issue and will write phonetics such as: nhaaaaa, nónggggg, likeeeee, ... to express their feelings.

Thus, in terms of meaning, “sp ok, chất lượng, phù hợp giá tiền, hàng như mong đợi, nv giao hàng nhiệt tình” and “sản phẩm ok, chất lượng, phù hợp giá tiền, hàng như mong đợi, nhân viên giao hàng nhiệt tình”, “sp” and “sản phẩm” (product), “vv” and “vui vẻ” (happy), “dễ dàng” and “rõ ràng” (clear), “nhaaaaa” and “nha”, “nónggggg” and “nóng” (hot) are the same. Humans would understand that, but the computer will consider them as different and extract two different feature vectors. To avoid the above situations, the authors built a dictionary of abbreviations and corresponding clearly spelled-out words, some basic misspelled words and their corresponding correctly spelled words. In addition, the dictionary also builds synonym lists, for example "10 points", "100 points" mean "excellent"; “thank you”, “thanks” mean “thanks”; “iphone13”, “12prm” mean “iphone”.... After that, they wrote a function to replace abbreviations and misspelled words in the data set with their corresponding correct words. This replacement helps reduce the number of words in the dictionary of the data set, thereby reducing the dimensionality of the feature vector and speeding up training for the model.

viii.Eliminating Numbers

In the project's data set, numbers often indicate dates and product prices. For reviews with product prices, in "can be bought for 2 million VND", the word expressing emotions is the word "can be"; "phone up to 30 million VND", the word expressing emotions is “up to". For reviews with dates, they usually indicate the date of purchase of the product. These numbers do not carry much emotional information about the consumer. Therefore, the authors proceed to delete numbers from the data set.

ix.Word Separation

English is an associative language, meaning between words are either spaces or punctuation. Unlike English, Vietnamese is an isolated language, meaning that spaces do not separate words but separate syllables, meaning a word can have one or more syllables. There are three main solutions to the word separation problem: statistics-based, dictionary-based and a combination of multiple methods (hybrid-based). In Vietnam, there are lots of research groups that have developed libraries for Vietnamese lexical tokenization: Pyvi (Python), CocCocTokenizer (C++), Underthesea (Python language), VnTokenizer (Java). The authors used the Underthesea [8] library that supports the Python language to perform lexical tokenization for the entire data set with an F1 score accuracy of 97.65%.

D. Eliminating Stopwords

Stopwords are words that are common and frequently appear in a language but usually are not important in language analysis. Removing stopwords aims to extract meaningful information from raw data and also helps reduce the dictionary size of the data set, reduce the dimensionality of the feature vector, and focus on important keywords that carry a lot of meaningful information. Stopword example includes "and", "is", "has", "be",... Table [2] below shows the top 10 stopwords that appear most often in the data set.

Table 2. Top 10 Most Frequently Occurring Stopwords in the Data Set

Word	Frequency
máy	7203
không	7117
rất	7007
tốt	6213
hài lòng	5419
được	4678
mua	4624
cực kỳ	3707
ok	3592
dùng	3544

Within the scope of research, the authors built a separate list of stopwords. This list is created based on the frequency of occurrence of words in the data set and the stopword list includes words that appear many times but do not carry important information. For example, according to the table above, the words “máy”, “mua”, “dùng” will be considered stopwords, and the words “không”, “tốt”, “hài lòng”, “được”, “ok”, “rất”, “cực kỳ” will remain. The choice of stopwords is developed by the authors. The project builds 4 different stopword sets for training. Stopwords are selected from 50, 70, 100, 120, 150, 170, 200 words that appear the most to create 4 sets of stopwords.

© Copyright: All rights reserved.

III. TRAINING AND ASSESSMENT

A. Feature Vector Extraction

Extracting feature vectors for a data set is the process of converting text data into numeric vectors for training classification models. The authors used the TF-IDF technique to extract feature vectors. After data preprocessing, there are a total of 311,440 words in the data set, and the dictionary of the data set has 11,753 words. The project uses the Tfidf Vectorizer() function of the scikit-learn library to extract feature vectors for the data set. After extracting the feature vector, the data set includes 18604 vectors, each of which has 11735 dimensions. Each vector represents a review in the data set and each dimension in the vector is the TF-IDF value of a word in the dictionary of the review.

B. Model Training and Evaluation

The data set is divided into two subsets: the training set and the test set. The training set is used to train the model, the test set is used to evaluate the model. The two sets, training and test, are divided in a ratio of 8:2, meaning the training set accounts for 80% and the test set accounts for 20% of the project data set. The training set includes 14,883 samples, including: 11625 samples with positive labels, 3258 samples with negative labels. The test set includes 3721 samples including: 2919 samples with positive labels, 802 samples with negative labels After dividing data into two subsets, put the training set into the models for training. The models are trained with the stopword set, which is divided into different times to find the standard set. After the models are trained, use the models to predict the test set. The trained models are evaluated in Table[3] with the results of the model evaluation indexes trained without using stopwords and stopword sets from 50, 70, 100, 120 , 150, 170, 200 most frequently occurring words.

Table 3. Results of Model Evaluation Indicators

Set of Stopwords	Algorithms	Results			
		Accuracy	Precision	Recall	F1-Score
Do not use stopwords	SVM	0.936	0.952	0.967	0.959
	Random forest	0.926	0.933	0.957	0.954
	Logistic Regression	0.93	0.954	0.967	0.956
	CNN	0.881	0.94	0.906	0.923
Set of stopwords from the 50 most frequently occurring words	SVM	0.936	0.952	0.967	0.959
	Random forest	0.927	0.936	0.974	0.954
	Logistic Regression	0.927	0.942	0.966	0.954
	CNN	0.858	0.874	0.95	0.913
Set of stopwords from the 70 most frequently occurring words	SVM	0.934	0.95	0.967	0.959
	Random forest	0.927	0.929	0.97	0.954
	Logistic Regression	0.929	0.944	0.966	0.955
	CNN	0.872	0.894	0.95	0.921
Set of stopwords from the 100 most frequently occurring words	SVM	0.934	0.949	0.967	0.958
	Random forest	0.926	0.935	0.973	0.954
	Logistic Regression	0.927	0.942	0.966	0.954
	CNN	0.904	0.949	0.928	0.938
Set of stopwords from the 120 most frequently occurring words	SVM	0.934	0.948	0.968	0.958
	Random forest	0.927	0.937	0.971	0.953
	Logistic Regression	0.927	0.942	0.966	0.954
	CNN	0.848	0.875	0.942	0.907
Set of stopwords from the 150 most frequently occurring words	SVM	0.934	0.948	0.968	0.958
	Random forest	0.927	0.937	0.973	0.954
	Logistic Regression	0.923	0.942	0.966	0.954
	CNN	0.888	0.907	0.954	0.93
Set of stopwords from the 170 most frequently occurring words	SVM	0.934	0.948	0.969	0.958
	Random forest	0.927	0.937	0.973	0.955
	Logistic Regression	0.927	0.942	0.966	0.954
	CNN	0.863	0.943	0.933	0.908
Set of stopwords from the 200 most frequently occurring words	SVM	0.933	0.947	0.968	0.958
	Random forest	0.926	0.935	0.973	0.954
	Logistic Regression	0.927	0.942	0.966	0.954
	CNN	0.864	0.879	0.959	0.917

From the table above, it is evident that the results of the measurements of the SVM, random forest, and logistic regression algorithms do not change significantly on different stopword sets. However, with the CNN convolutional neural network, the best results are achieved with stopword set from the 100 most frequently occurring words. Therefore, the authors decided to use a set of stopwords built from the 100 most frequently occurring words.

IV. CONCLUSION

Based on the understanding of the application of machine learning in text classification, the authors have researched machine learning models for the classification of customer reviews on e-commerce platforms, from data collection from websites to data preprocessing, training Support Vector Machine, Random Forest, Logistic Regression models and Convolutional Neural Network deep learning networks. Specifically, the authors built a set of stopwords, experimented and evaluated the results on the set of stopwords and collected data. From there, a website can be built integrating the trained models. Users only need to enter a product review or enter a product link on the e-commerce platform to choose any classification algorithm and the system will automatically classify and give results.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures that the research is conducted with objectivity and without any external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.

© Copyright: All rights reserved.

- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Authors Contributions:** The authorship of this article is attributed equally to all participating authors.

REFERENCES

1. Fradkin, Dmitriy; Muchnik, Ilya (2006). "Support Vector Machines for Classification" Discrete Methods in Epidemiology. DIMACS Series in Discrete Mathematics and Theoretical Computer Science. Vol. 70. pp. 13–20
2. Ho, Tin Kam (1995). Random Decision Forests Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995.
3. Joachims T., "Text categorization with Support Vector Machines: Learning with many relevant features", in Proc. of the European Conference on Machine Learning (ECML), 1998, pages 137–142. https://doi.org/10.1007/BFb0026683
4. K. J. Piczak, "Environmental sound classification with convolutional neural networks," in Proceedings of the IEEE 25th International Workshop on Machine Learning for Signal Processing, 2015, pp. 1-6. https://doi.org/10.1109/MLSP.2015.7324337
5. Prinzie, A.; Van den Poel, D. (2008). "Random Forests for multiclass classification: Random MultiNomial Logit". Expert Systems with Applications. 34 (3): 1721–1732. https://doi.org/10.1016/j.eswa.2007.01.029
6. Tolles, Juliana; Meurer, William J (2016). "Logistic Regression Relating Patient Characteristics to Outcomes". JAMA. 316 (5): 533–4. https://doi.org/10.1001/jama.2016.7653
7. Venkatesan, Ragav; Li, Baoxin (2017-10-23). Convolutional Neural Networks in Visual Computing: A Concise Guide. CRC Press. ISBN 978-1-351-65032-8. Archived from the original on 2023-10-16. Retrieved 2020-12-13.
8. https://underthesea.readthedocs.io/en/latest/readme.html
9. Thegioididong.com
10. Sistla, S. (2022). Predicting Diabetes u sing SVM Implemented by Machine Learning. In International Journal of Soft Computing and Engineering (Vol. 12, Issue 2, pp. 16–18). https://doi.org/10.35940/ijsce.b3557.0512222
11. Muthukrishnan, Dr. R., & Prakash, N. U. (2023). Validate Model Endorsed for Support Vector Machine Alignment with Kernel Function and Depth Concept to Get Superlative Accurateness. In International Journal of Basic Sciences and Applied Computing (Vol. 9, Issue 7, pp. 1–5). https://doi.org/10.35940/ijbsac.g0486.039723
12. Nagar, K., & Chawla, M. P. S. (2023). A Survey on Various Approaches for Support Vector Machine Based Engineering Applications. In International Journal of Emerging Science and Engineering (Vol. 11, Issue 11, pp. 6–11). https://doi.org/10.35940/ijese.k2555.10111123
13. Sharma, D. S. K., & Sharma, M. N. K. (2019). Text Classification Using Ensemble Of Non-Linear Support Vector Machines. In International Journal of Innovative Technology and Exploring Engineering (Vol. 8, Issue 10, pp. 3169–3174). https://doi.org/10.35940/ijitee.j9520.0881019
14. Kumar, C. S., & Thangaraju, P. (2019). Improving Classifier Accuracy for diagnosing Chronic Kidney Disease Using Support Vector Machines. In International Journal of Engineering and Advanced Technology (Vol. 8, Issue 6, pp. 3697–3706). https://doi.org/10.35940/ijeat.f9377.088619

AUTHORS PROFILE

Dr. Phan Thi Ha is currently a lecturer at the Faculty of Information Technology at Posts and Telecommunications Institute of Technology (PTIT) in Vietnam, and Computing Fundamental Department, FPT University, Hanoi 10000, Vietnam as well. She received a B.Sc.in Math & Informatics, a M.Sc. in Mathematic Guarantee for Computer Systems and a PhD. in Information Systems in 1994, 2000 and 2013, respectively. Her research interests include machine learning, natural language processing and mathematics applications.

Trinh Thi Van Anh is currently a lecturer and a researcher at the Faculty of Information Technology at Posts and Telecommunications Institute of Technology (PTIT) in Vietnam and Computing Fundamental Department, FPT University, Hanoi 10000, Vietnam as well. She received a B.Sc. in Electronics-Telecommunications in 1993 (HUST), M.Sc. in Computer Science in 1998 (HUST). Her research interests include machine learning, NLP and opinion mining.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© Copyright: All rights reserved.