METHOD OF CONTENT CLASSIFICATION BASED ON SUPERVISED MACHINE LEARNING IN ONLINE COMMENT MINING OF CUSTOMER

Le Trieu Tuan
lttuan@ictu.edu.vn
University of Information and Communication Technology - Thai Nguyen University, Thai Nguyen, Vietnam
Dr. Pham Minh Hoan
hoanpm@neu.edu.vn
National Economics University, Hanoi, Vietnam

Abstract

The study aims to apply the supervised machine learning method to the classification of product review content in online customer comment mining. The entire study was conducted automatic data collection with 2,241 customer reviews on products on Lazada.vn, then trained with Supervised Machine Learning models to find the most suitable model with the training dataset and apply this model to predict the reviews content for the dataset. The results show that the machine learning methods, those are Support Vector Machines (SVM), Decision Tree (DT) and Neural Network (NN) have the best performance with classifying customer comments in Vietnamese. The research results have reference value for applications of comment mining in the field of online business.

Keywords: *Content classification, Comment classification, Using supervised machine learning.*

1. Introduction

With the advancement of information technology has changed the way communication makes it easy for customers to access information and exchange content about products and services on a large scale in real time. The advent of social networks and e-commerce websites allows customers to evaluate products online through comments, such as: Lazada.vn, shopee.vn, tiki.vn, etc (Ochilbek Rakhmanov, 2020). With the explosion of Big Data, the comment of the online community needs to be collected and exploited automatically, allowing merchants to track shopping customer behavior, detect customer preferences and support customers to buy products and services in the best way (MehdiGolzadeh & et al, 2021).

Content classification is an important step in machine learning to research and exploit online customer comments. Currently, there are many authors who have studied the method of content classification at different levels. From results of domestic and foreign studies, the

author finds that there are two approaches to classifying online comment contents by machine learning method: (1) Supervised Machine Learning and (2) Unsupervised Machine Learning (Sun & et al, 2017). Research direction, the method of exploiting customer comment contents is not newest, however, each method has its own advantages and disadvantages, no method is considered to be absolutely accurate. This study applies a supervised machine learning method to classify online customer comment contents with automatically collected data sources, including 2,241 customer comments about products on the Lazada.vn website.

2. Literature Review

2.1. Exploit the customer comments

Exploiting customer comments is a field of research to analyze and evaluate customers' opinions on objects such as products, services, organizations, individuals, events, topics and their properties (Pang & Lee, 2008; Liu, 2012). A customer comment mining process typically consists of three main steps: (1) Comment Retrieval, (2) Content Classification, and (3) Comment Summarization (K.M. Kavitha & et al, 2020; Kumar & Reddy, 2016). In which, content classification is considered as the most important step for the purpose of classifying comment according to the following levels: Positive; Negative; and Neutral. According to Liu (2012), customer comments mining is divided into three levels: (1) Document Level, at this mining level, it is assumed that each document represents the comment content about a single entity. Therefore, the analysis will not be applicable to documents that cover many subjects; (2) Sentence Level, at this mining level, it is assumed that each sentence represents the content of an object, however, the analysis will ignore the sentences with many clauses, each of which represents comments on different subjects; and (3) the entity/aspect level, instead of exploiting the comment according to the linguistic structure (document, sentence, clause...), this level of analysis looks at the content in target, the target of the comment can be an object or an aspect (attribute) of the object. Today, with the explosion of Big Data, the exploitation of customer comments has become a great concern of businesses, especially companies with websites that allow users to respond on the internet. Customer comments mining can also be added to Recommender Systems to recommend products with positive comment and not recommend product categories that receive a lot of negative comment (Özlem & Tutku, 2021; Pang & Lee, 2008).

2.2. Customer comment contents classification using Supervised Machine Learning

Supervised Lachine Learning is a Computer Science technique for constructing a function from training data. The training data consists of pairs of input objects (usually vectors), and desired outputs. The output of a function is to predict a label for an object (Reynaldo & et al, 2019).

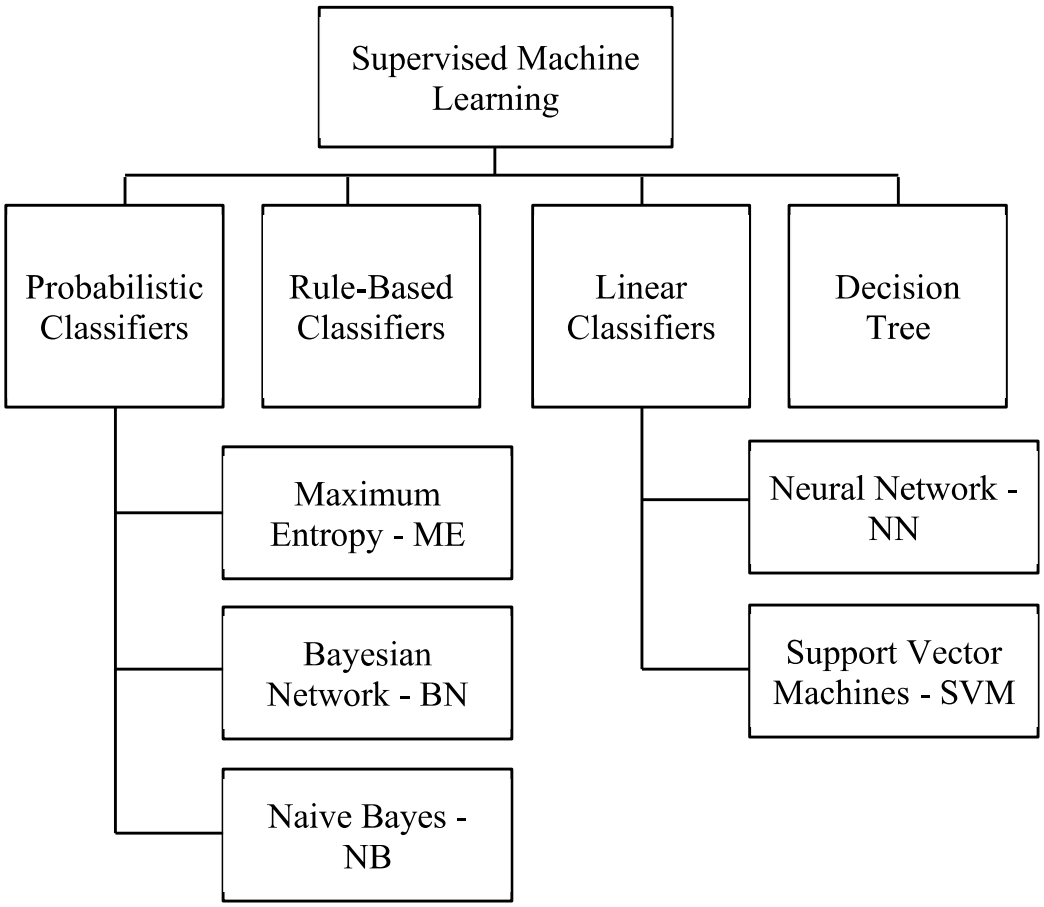

Figure 1: ***Customer comment contents classification using Supervised Machine Learning***

2.3. Some tools to support classification

No.	Tool name	Uses	Reference resource
1	Natural Language Toolkit (NLTK)	For text processing, coding classification, Stemming, tagging, parsing, and easy-to-use interface with more than 50 word and content resources vocabulary.	http://www.nltk.org/
2	GATE	Useful in developing a pipeline. Language analysis modules for different languages are contributed by developers. They are available to be used integrated into the pipeline.	https://gate.ac.uk/

3	CoreNLP	Perform the most common natural language processing tasks, such as: Part-of-Speech Tagging, Named Entity Extraction, Chunking and Co-Reference.	http://nlp.stanford.edu/software/corenlp.html
4	OpenNLP	This is a JAVA library for natural language processing, supporting common tasks, including: encoding, sentence decomposition, word type labeling, object recognition, parsing.	https://opennlp.apache.org
5	WEKA	Algorithms data mining, data preprocessing, classification, clustering, regression, association rules, visualization.	http://www.cs.waikato.ac.nz/ml/weka/
6	VnTokenizer	This is a specialized tool to separate words, assign word categories to Vietnamese, developed by Le et al (2008). VnTokenizer is written in JAVA, can be used as Tools Command Line or Programming.	http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer
7	Underthesea - Vietnamese NLP Toolkit	An open source set of Python modules, datasets and tutorials that support research and development in Vietnamese Natural Language Processing.	https://underthesea.readthedocs.io

3. Method

This study was conducted according to the method of knowledge mining from Knowledge Discovery in Databases - KDD. The steps in the research process are carried out as shown in Figure 2. The experimental environment is installed in Python programming

language with the support of Python Vietnamese Toolkit (for Vietnamese language) and other tools available library.

Step 1. Data collection and preprocessing

The study was conducted to collect data by an automatic program, the data was taken from Lazada.vn website. This is a method of automatically collecting content from HTML pages of any internet resource by special programs or scripts. With the object and scope of the research aimed at the Vietnamese language, the data only uses customer comment in Vietnamese. Next, the study carried out data preprocessing by removing the missing data, the responses do not contain the necessary information to proceed to the next processing step.

Step 2. Data Labeling

This step is to prepare a labeled (or classified) dataset large enough to be used as a training data set. Usually for studies applying machine learning methods, this dataset will be built manually. However, in this study, after randomly reviewing the content of the collected comment dataset and based on the results of the rating scores (the rating field in the dataset), this study found that the comment with rating score less than 7.0 has negative meaning (Negative), and conversely, comment with rating score greater than 7.0 has positive meaning (Positive). Therefore, the training dataset was determined to have 2,241 comments, of which 81 were negative (labelled 0) and 2,160 comments were positive (labelled 1).

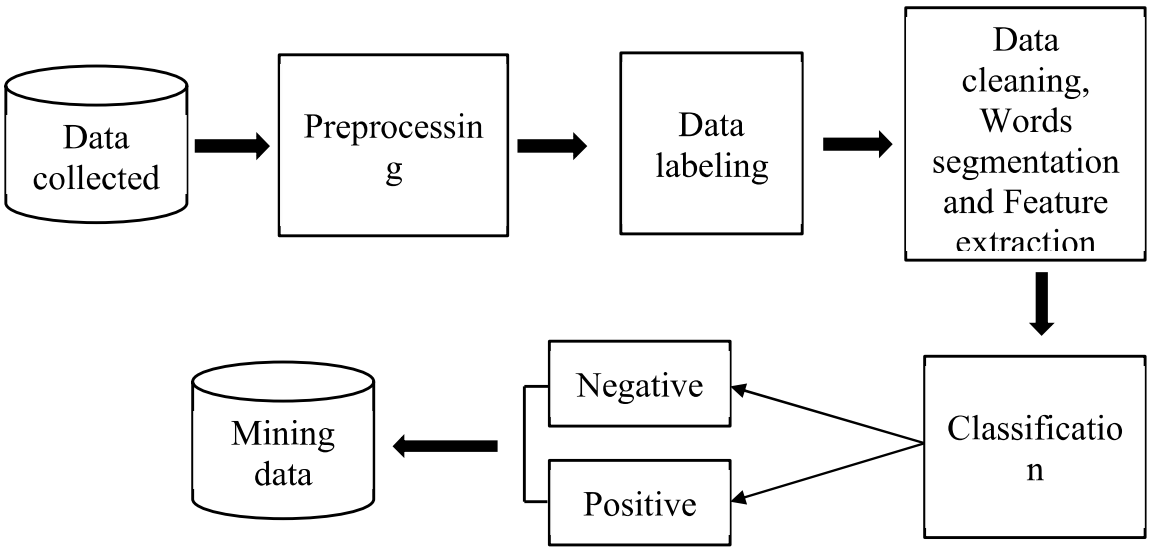

Figure 2: Research process

Step 3. Data cleaning, Words segmentation and Feature extraction

Data Cleaning: This step conducts data cleaning before starting processing on the dataset, including some natural language processing steps such as removing stop words, or checking spelling, etc.

Words Segmentation: This step is very important in natural language processing, and especially for Vietnamese language because there are many compound words, separating words in different ways can cause semantic ambiguity. This study inherits the separated library from Python Vietnamese Toolkit.

Feature Extraction: This step will select the typical features (keywords) that are representative of the data set as input for the classification algorithm. This study selects keywords according to TF-IDF method (Term Frequency/Inverse Document Frequency), the TF-IDF value of a keyword is a number obtained through statistics showing the importance of this keyword in a comment. The TF-IDF of the keyword w_i in the response d is calculated using the following formula:

$$if\_idf_{id} = f_{id} \times log\frac{N}{n_i}$$

With f_{id}: The frequency of occurrence of the keyword w_i in the comment d

N: Total number of the comment

n_i: Number of the comment where the keyword w_i appears

Comment classifier model training: This phase aims to determine whether a customer comment is Positive or Negative. This study applies the classification algorithms belonging to the group of supervised machine learning which is considered to be the best, they are Naive Bayes, Support Vector Machines, Neural Network and Decision Tree algorithms. Based on the combined results from previous studies related to the topic, to find the most suitable model for the dataset, which is the classified comments, and then to make predictions for the data unclassified comments or new comments data generated without retraining. The training process is carried out by the K-fold method, randomly dividing the data into K non-intersecting subsets. For each experiment (out of K times), one subset is used as the test data, and (K-1) the remaining subset is used as the training data. This study was conducted with K=5.

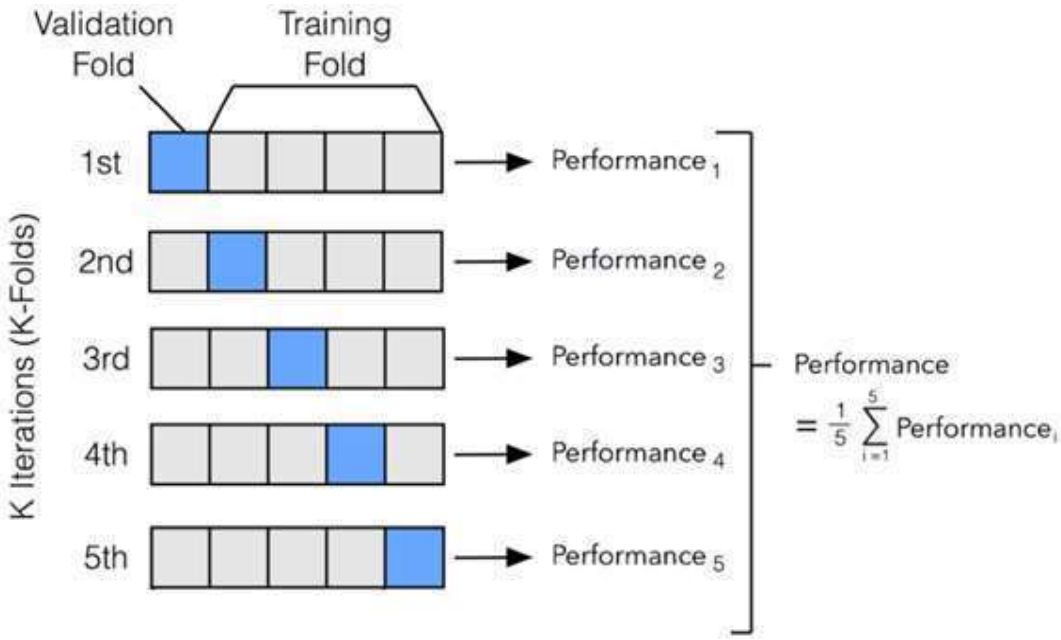

Figure 3: K-Fold method

With Performance: An average performance of 5 experiments

K Iterations: To repeat K times

Validation Fold: The dataset is used for testing

Training Fold: The dataset is used for training

Step 4. Evaluation of classification efficiency

The study used the method of evaluating the classification model which is based on the calculated indexes in the Confusion Matrix as Table 1.

Table 1: Confusion Matrix

	Reality: Positive	**Reality: Negative**
Predict: Positive	True Positive (TP)	False Negative (FN)
Predict: Negative	False Positive (FP)	True Negative (TN)

The effectiveness of the comment classification model is evaluated based on 4 indicators: Accuracy, Precision, Recall, and Average value of harmonics (F1). In addition, this study also considers the training time factor (Time) of each model.

With in:

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision \times Recall}$$

4. Results

4.1. Results of data collection and preprocessing

The results of data collection were 2,241 customer comments on products in Vietnamese of 15 items across 5 stores. The data is distributed as shown in Table 2.

Table 2: Results of data collection and preprocessing

No.	Products	Amount	Number of comments	Medium
1.	Television	14	173	12.4
2.	Fridge	6	93	15.5
3.	Air conditioning	9	196	21.8
4.	Jeans	3	12	4.0
5.	T-shirt	4	54	13.5
6.	Man swimwear	3	45	15.0
7.	Iphone 12	2	246	123.0
8.	Iphone 11 pro max	2	198	99.0
9.	Iphone 10	4	74	18.5
10.	Samsung Galaxy A32	3	297	99.0
11.	Samsung A52	3	157	52.3
12.	Samsung A72	2	65	32.5
13.	OPPO Watch	5	289	57.8
14.	OPPO Reno5	4	32	8.0
15.	OPPO A53	4	310	77.5
Total		68	2.241	

Table 3: Training results by K-Fold (K=5)

No.	Method name	Average accuracy	Standard deviation	Training time (seconds)
1	Naive Bayes (NB)	0.48	0.05	16.02
2	Support Vector Machines (SVM)	0.80	0.02	4.33
3	Neural Network (NN)	0.79	0.03	312.29
4	Decision Tree (DT)	0.70	0.03	315.56

The training results showed that the SVM, NN and DT models have the best accuracy (0.80, 0.79 and 0.70). Meaning these models were relatively suitable for the training dataset. However, if considering the factor of training time, only the NB and SVM models were the best. Therefore, subsequent applications could use two these models as a tool to classify comments for unclassified comment data or newly generated comment data without retraining. The results of this study helped determine the appropriate method and tool for classifying comments.

5. Conclusion

This study has conducted a review of the theoretical basis of the comment classification method and proposed the application of a supervised machine learning method for automatic comment content mining. Experimental results show that Neural Network, Decision Tree and Support Vector Machines methods are the best in training methods. This study has a reference value for applications of comment mining in the field of online sales. Online businesses can use the results to automatically assess which items are the best rated by many customers, thereby increasing the number of items in stock, increasing the frequency of suggestions and marketing that product to prospective customers for increasing sales. And the items that have many negative customer reviews are an opportunity for businesses to consider what factors inside the company or outside lead to that. If the product is not good, the enterprise considers reducing the number of such items. However, this study still has many limitations, which can be continued in the future or in future studies: Firstly, in terms of data collection, this study only collects data that is the customer comment on items on Lazada.vn website. The study can extend to collecting customer comment on any products or services on e-commerce websites or social networking sites; Secondly, about the scale, this study only classifies customer comment on a scale of 2 levels: Positive, and Negative. The next research direction can use a scale of more levels (for examples, a 5-level Likert scale); Third, about the comment content classification technique, this study only uses the supervised machine learning method, if combined with the semantic-based lexical method, it may give better results.

6. References

15. K.M. Kavitha & et al, (2020), *Analysis and Classification of User Comments on YouTube Videos,* International Workshop on Artificial Intelligence for Natural Language Processing (IA&NLP 2020), Vol 177, pp. 593-598.

16. Kumar, S., & Reddy, B, (2016), *An analysis on opinion mining: Techniques and tools*, Indian Journal of Research, 5(8), pp. 489–492.

17. Le, N. M., Do, B. N., Nguyen, V. D., & Nguyen, T. D, (2013), VNLP: *An open source framework for Vietnamese natural language processing*, In Proceedings of the Fourth Symposium on Information and Communication Technology, 88–93.

18. Liu, B, (2012), *Sentiment analysis and opinion mining*, Synthesis Lectures on Human Language Technologies, 5(1), pp. 1–167.

19. MehdiGolzadeh & et al, (2021), *A ground-truth dataset and classification model for detecting bots in GitHub issue and PR comments*, Journal of Systems and Software, Vol 175 pp. 110-125.

20. Ochilbek Rakhmanov, (2020), *A Comparative Study on Vectorization and Classification Techniques in Sentiment Analysis to Classify Student-Lecturer Comments*, Procedia Computer Science, Vol 178, pp. 194-204.

21. Özlem & Tutku, (2021), *Classification of rare diseases: A comment on 'atlas of esophageal atresia'*, Journal of Pediatric Surgery.

22. Pang, B., & Lee, L, (2008), *Opinion mining and sentiment analysis*, Foundations and Trends in Information Retrieval, 2(1-2), pp. 1–135.

23. Reynaldo & et al, (2019), *Gender Demography Classification on Instagram based on User's Comments Section*, 4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), 157, 64-71.

24. Sun & et al, (2017), *A review of natural language processing techniques for opinion mining systems*, Information Fusion, 36, 10–25.