A MACHINE LEARNING APPROACH FOR OPINION MINING ONLINE CUSTOMER REVIEWS

Dr. N. Raja Kumar[1]*, Mrs. Pooja Rajesh Chavan[2], Sumayya Begum[3]

Abstract

This study was conducted to apply supervised machine learning methods in opinion mining online customer reviews. First, the study automatically collected 39,976 traveller reviews on hotels in Vietnam on Agoda.com website, then conducted the training with machine learning models to find out which model is most compatible with the training dataset and apply this model to forecast opinions for the collected dataset. The results showed that Logistic Regression (LR), Support Vector Machines (SVM) and Neural Network (NN) methods have the best performance in opinion mining in Vietnamese language. This study is valuable as a reference for applications of opinion mining in the field of business.

[1]*Associate Professor, Department of IT, Lords Institute of Engineering and Technology, Hyderabad, Email: dr.rajkumar@lords.ac.in
[2,3]Assistant Professor, Department of IT, Lords Institute of Engineering and Technology, Hyderabad

***Corresponding Author:** Dr. N. Raja Kumar
*Email: dr.rajkumar@lords.ac.in

DOI: 10.53555/ecb/2022.11.11.214

I. INTRODUCTION

Today, the improvement of information technology has changed the ways of communication, making it easier for customers to access the information and exchange idea about products or services on a large scale in real time. Social networks and online review websites (such as Agoda, TripAdvisor, Yelp, Amazon, etc.) allow customers to give their opinions on products or services through reviews [11]. With the explosion of big data, it is necessary to collect and exploit automatically their online reviews so that business enterprises can easily understand customer purchase behavior, as well as their interests and satisfaction

Level on product or service quality. Opinion mining has become the subject of studies in different areas: market research, e-business, political polls [10]. Currently, the community of scientists have lots of studies on opinion mining methods as well as the application of opinion mining at different levels. From the results of different studies, the author recognized two popular approaches in opinion mining: (I) machine learning and (ii) lexical based method, as see in [1], [10], [12], [13]. Besides, in order to increase the efficiency of opinion mining, the studies used a hybrid method of machine learning and lexical based [12]. Research methodology on opinion mining is not new, however, each method has its advantages and disadvantages and none of them are considered to be absolutely accurate.

In order to enhance customer satisfaction and their shopping experiences, it has become a common practice for online traders to enable their customers to review or to express opinions on the products that they buy. A common user becoming comfortable with the Internet, an increasing number of people are writing reviews. The number of reviews

that a product receives grows rapidly. Some popular products can get hundreds of reviews at some large trading sites.

A feature-based opinion summarization of customer reviews of products sold online. The task is performed in two steps:

1. Identify the features of the product that customers have expressed opinions and rank the features. 2. For each feature, we identify how many customer reviews have positive or negative opinions. The specific reviews that express these opinions are attached to the feature. This makes it very hard for a potential customer to read them to help to make a decision on whether to buy the product.

II. OBJECTIVES

1.Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3.When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow.

In particular, the application of lexical based method for Vietnamese is a big challenge for researchers because the language structure is complex and there are few emotion vocabulary sets and well processing tools for Vietnamese. Therefore, the application of machine learning and the evaluation of the methods' accuracy is necessary to select the most appropriate one for the research's field through the collected dataset. The objective of this study is to review studies on opinion mining and propose the application of machine learning method in opinion mining customer reviews in Vietnamese. The method of knowledge discovery in databases is applied to this study in which 39,976 tourists' reviews on hotels in Vietnam are collected through Agoda.com. Then, the study conducts data pre-processing and training using machine learning methods to find the most suitable model with the training data sets and apply this model to forecast opinions for the entire dataset.

III. INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay,

avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

IV. OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2.Select methods for presenting information.

3.Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

V.SYSTEM ANALYSIS

In the existing system opinion mining focuses or rather we can say that it aimed at analysing and assessing people's perceptions of objects such as products, services, organizations, individuals, events, topics and their attributes. It uses SentiWordNet algorithm. The process of finding user opinion about the topic or product or problem is called as opinion mining. It can also be defined as the process of automatic extraction of knowledge by means of opinions expressed by the user who is currently using the product about some product is called as opinion mining. Analysing the emotions from the extracted opinions is defined as Sentiment Analysis. The goal of opinion mining and Sentiment Analysis is to make computer able to recognize and express emotion.

There are several drawbacks in following the methods specified which is existing like, it is not suitable for online reviews. The performance of this algorithm is very low. The model could be very difficult to train if use the softmax function, since the number of categories is too large (the size of vocabulary).

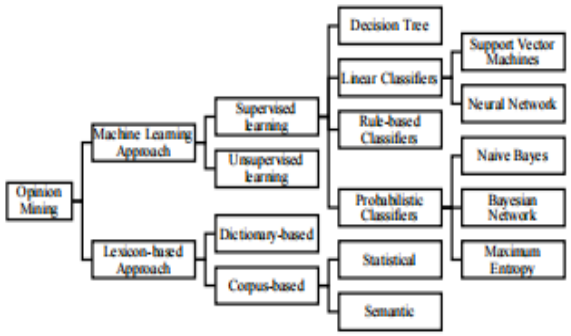
In the mechanism and the proposed technique or the algorithm used is about the SentiWordNet algorithm used with its classic implementation for opinion classification is a text mining technique natural language processing (NLP). Machine learning method plays an important role in opinion mining. Opinion mining at the document and sentence level is used to determine whether a statement is positive or negative. The objective of this study is to review studies on opinion mining and propose the application of machine learning method in opinion mining customer reviews in Vietnamese. The method of knowledge discovery in databases is applied to this study in which 39,976 tourists' reviews on hotels in Vietnam are collected through Agoda.com. Then, the study conducts data pre-processing and training using machine learning methods to find the most suitable model with the training data sets and apply this model to forecast opinions for the entire dataset.

With the implementation of the algorithm we have several advantages and it can give the good result and the accuracy can be improved. Experimental results show that LR, SVM and NN are the best among the training methods. This study is valuable as a reference for applications of opinion mining in socioeconomic fields. Opinion mining at the document and sentence level is used to determine whether a statement is positive or negative. Algorithms: Logistic Regression (LR), Support Vector Machines (SVM) and Neural Network (NN).

VI. PROCESSING ARCHITECTURE

Opinion mining is a type of text mining which classify the text into several classes. Sentiment analysis which also known as Opinion mining use some algorithm techniques to categorize the user opinions into positive, negative and neutral classes. This categorization of text is called polarity of text. The main objective of Sentiment analysis is classification of sentiment. It classifies

the given text into three level document level, sentence level, and entity/aspect
level. In document level classification, a single review about a single topic is considered (i.e.) either positive or negative opinion. Gathering the data in which the analysis is to be drawn out is the prior step to be done and continuing to visualization of the numerical attributes present and figuring out the dependencies of various attributes using correlations. Here, as there is no strong correlation among the attributes present in the dataset, there cannot be a picturization of dependent attributes. Continuing with classification tasks by setting a target attribute, user defined function basing upon the ratings given to a product by the customer. Figuring out the compatibility between this ratings and review is the main method to be followed for classification. The target variable can be set on the values of rating: Positive (4, 5), neutral (3), negative
(1, 2).



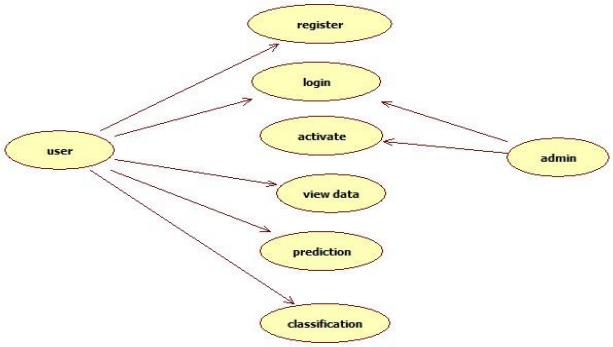
It is a simple graphical formalism that can be used to represent in a data flow diagram also, is is a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system. The data flow diagram is one of the most important modelling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

VII. Logistic Regression Model

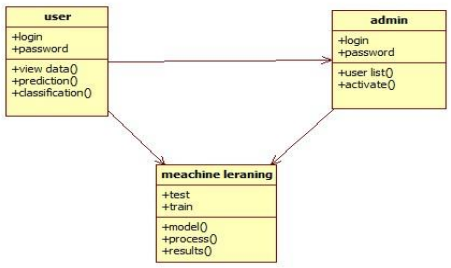
Logistic Regression is basically a supervised learning algorithm which is used to predict the probability of the class label. This is not only used for classification among the labels but also used in the case of estimation of probability, that a particular data tuple is related or is belonging to the class. Estimating probabilities: It is an easy method, calculating the sum of input attributes including their bias term. Vectorised form of the probabilities [24].
In the below cost function it is so clear that the positive probability i.e.; -log (P) grows very large when P approaches to 0, concluding the
model estimating a low probability which is approaching 0. Similarly, in the negative probability case i.e.; -log (1-P) the cost function grows high whenever P approaches 1. Cost function in logistic regression is as shown, C (θ) = -log (P) if y=1, -log (1-P) if y=0. The entire functionalities is also represented using the UML diagrams it is Unified Modelling Language. UML is a standardized general-purpose modelling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.
The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: A Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.
The Unified Modelling Language is a standard language for specifying, Visualization, Constructing and documenting the artefacts of software system, as well as for business modelling and other non-software systems.
The UML represents a collection of best engineering practices that have proven successful in the modelling of large and complex systems. The UML is a very important part of developing object-oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects. The Primary goals in the design of the UML are as follows:
Provide users a ready-to-use, expressive visual modelling Language so that they can develop and exchange meaningful models.
Provide extendibility and specialization mechanisms to extend the core concepts.
Be independent of particular programming languages and development process.
Provide a formal basis for understanding the modelling language. Encourage the growth of OO tools market. Support higher level development concepts such as collaborations, frameworks, patterns and components to integrate best practices.

Parsing: In the parsing task, a parser develops the parse tree given a sentence. A few parsers accept the presence of an arrangement of language structure rules so as to parse however late parsers are sufficiently keen to conclude the
parse trees straightforwardly from the given information utilizing complex measurable models[12].

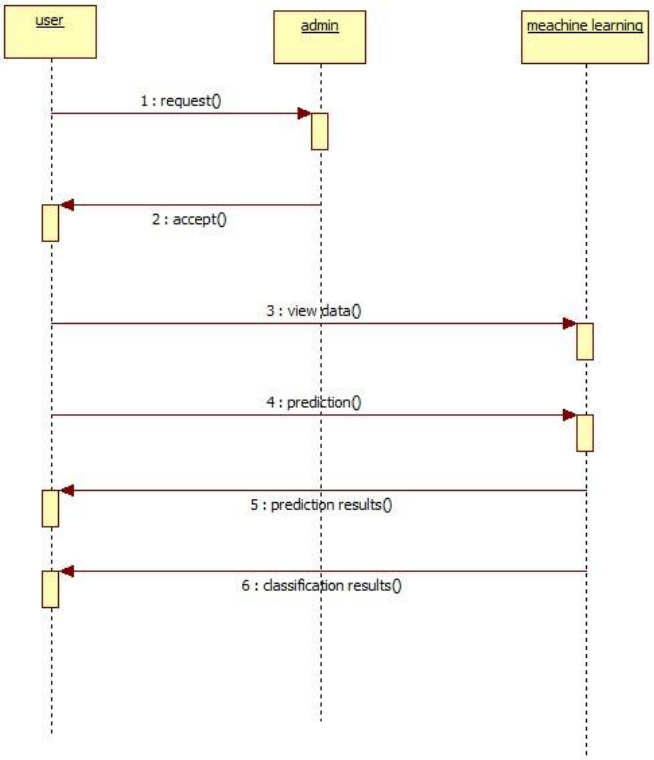
Use Case: A use case diagram in the Unified Modelling Language (UML) is a type of behavioural diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



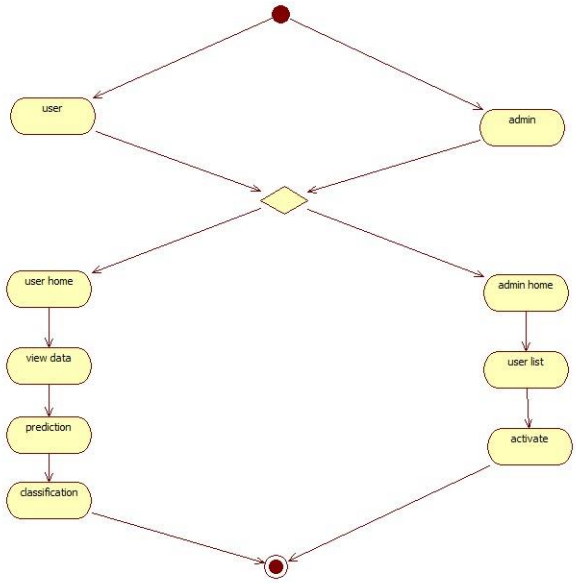
Class representation: In software engineering, a class diagram in the Unified Modelling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.



Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modelling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



Modules are 1. User 2. Admin 3. Data Pre-processing 4. Machine Learning

User: The User can register the first. While registering he required a valid user email and mobile for further communications. Once the user register then admin can activate the user. Once admin activated the user then user can login into

our system. User can upload the dataset based on our dataset column matched. For algorithm execution data must be in float format. Here we took land resource online customer reviews dataset for testing purpose. User can also add the new data for existing dataset based on our Django application. User can click the Classification in the web page so that the data calculated Accuracy, precision, recall and F1-Score based on the algorithms. User can click Prediction in the web page so that user can write the review after predict the review That will display results depends upon review like positive, negative or neutral.

Admin: Admin can login with his login details. Admin can activate the registered users. Once he activates then only the user can login into our system. Admin can view the overall data in the browser. Admin can click the Results in the web page so calculated Accuracy, precision, recall and f1 score based on the algorithms is displayed. All algorithms execution complete then admin can see the overall accuracy in web page.

Data Pre-processing: A dataset can be viewed as a collection of data objects, which are often also called as a records, points, vectors, patterns, events, cases, samples, observations, or entities. Data objects are described by a number of features that capture the basic characteristics of an object, such as the mass of a physical object or the time at which an event occurred, etc. Features are often called as variables, characteristics, fields, attributes, or dimensions. The data pre-processing in this forecast uses techniques like removal of noise in the data, the expulsion of missing information, modifying default values if relevant and grouping of attributes for prediction at various levels.

Machine learning: Based on the split criterion, the cleansed data is split into 60% training and 40% test, then the dataset is subjected to four machine learning classifiers such as Naive bayes (NB), Support Vector Machine (SVM), Decision Tree, logical regression, (LR), Neural network, Random forest (RF). The accuracy and f1 score, precision, recall of the classifiers was calculated and displayed in my results. The classifier which bags up the highest accuracy could be determined as the best classifier.

VIII.CONCLUSION

This study has conducted a theoretical background on opinion mining methods, opinion classification techniques and proposed the application of supervised machine learning method for automatic opinion mining. Experimental results show that LR, SVM and NN are the best among the training methods. This study is valuable as a reference for applications of opinion mining in socioeconomic fields. However, this study still has some limits that can be adjusted in future studies. Firstly, in terms of data collection, this study only collects customer reviews about hotels on Agoda.com. The study may expand to collect reviews about any products or services on ecommerce websites or social networks.

Secondly, in terms of the scale, this study only classifies customer reviews on a 2-level scale: positive and negative. More level scales may be applied in the next study (for example, on a 5-level Likert scale). Thirdly, in terms of opinion classification technique, this study only uses supervised machine learning method. It will give better results with a hybrid method of supervised machine learning and lexicon based. However, currently, there are not many tools that support processing Vietnamese as well as English. Finally, this research is just limited to the classification of opinions. The

extended research's directions will focus on the application of opinion mining in behaviour, sentiment, and shopping preference analysis as well as products and services quality assessment, which has more practical implications for entrepreneurs and customers.

REFERENCES

[1] A. Dhokrat, S. Khillare, and C. N. Mahender, "Review on techniques and tools used for opinion mining,"International Journal of Computer Applications Technology and Research, vol. 4, no. 6, pp. 419-424,2015.

[2] B. Liu, "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies,vol. 5, no. 1, pp. 1-167, 2012.

[3] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and trends in information retrieval, vol 2., no. 12, pp. 1-135, 2008.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, pp. 993-1022, 2003.

[5] F. Wanner, C. Rohrdantz, F. Mansmann, D. Oelke, and D. A. Keim, "Visual sentiment analysis of RSS news feeds featuring the us presidential election in 2008," In VISSW, 2009.

[6] G. Stylios, D. Christodoulakis, J. Besharat, M.Vonitsanou, I. Kotrotsos, A. Koumpouri, and S.Stamou, "Public opinion mining for governmental decisions," Electronic Journal of e-Government, vol. 8, no. 2, pp. 203-214, 2010.

[7] H. Binali, V. Potdar, and C. Wu, "A state of the artopinion mining and its application domains,"International Conference on Industrial Technology,pp. 1-6, 2009.
[8] J. Lee, D. H. Park, and I. Han, "The different effects ofonline consumer reviews on consumers' purchaseintentions depending on trust in online shopping malls: An advertising perspective," Internet Research, vol. 21, no. 2, pp. 187–206, 2011.
[9] J. Shawe-Taylor and S. Sun, "A review of optimizationmethodologies in support vector machines,"Neurocomputing, vol. 74, no. 17, pp. 3609-3618, 2011.
[10] S. K. Yadav, "Sentiment analysis and classification: A survey," International Journal of Advance Research in Computer Science and Management Studies, vol. 3,no. 3, pp. 113-121, 2015.
[11] S. M. Mudambi and D. Schuff, "What makes a helpfulreview? A study of customer reviews on Amazon.com," MIS quarterly, vol. 34, no. 1, pp. 185-200, 2010.
[12] S. Sun, C. Luo, and J. Chen, "A review of naturallanguage processing techniques for opinion mining systems," Information Fusion, vol. 36, pp. 10-25, 2017.
[13] W. Medhat, A. Hassan, and H. Korashy, "Sentimentanalysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093-1113, 2014.246
[14] M. Haenlein, A. Kaplan, "A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence", California Management Review, 61(4), 2019, pp 5-14
[15] V. Kaul, S. Enslin, S.A. Gross, "The history of artificial intelligence in medicine", Gastrointestinal Endoscopy, In Press, 2020.
[16] M. Minsky, S.A. Papert, "Perceptrons: An Introduction to Computational Geometry", MIT Press, Cambridge, MA, 1969.
[17] A. Esteva et al., "A guide to deep learning in healthcare", Nature Medicine, 25(1), 2019, pp 24-29.
[18] D. Ravi et al., "Deep Learning for Health Informatics,", IEEE Journal of Biomedical and Health Informatics, 21(1), 2017, pp. 4-21.
[19] Y.J. Yang, C.S. Bang, "Application of artificial intelligence in gastroenterology", World Journal of Gastroenterolology, 2019(25), 2019, pp 1666-1683.
[20] A. Hosny, C. Parmar, J. Quackenbush, L.H. Schwartz, H.J.W.L. Aerts,"Artificial intelligence in radiology", Nature Reviews. Cancer, 18(8), 2018, pp 500-510.
[21].K.W. Johnson, et al., "Artificial Intelligence in Cardiology", Journal of the American College of Cardiology, 71(23), 2018, pp 2668-2679.
[22]. V.N. Perisic, B. Jankovic, "Pedrijatrija za studente medicine", Medicinski fakultet, Univerziteta u Beogradu, Beograd, 2010.
[23]. N. Chen et al., "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study", The Lancet, 395(10223), 2020, pp 507-513.
[24]. D. Kermany, K. Zhang, M. Goldbaum, "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification", Mendeley Data, v2, 2018. Accessed July 2020. [Online].
[25] Havinash P.H, Jeril Johnson N, Glen Thomas, Emily Stephen, Mining Opinion Features in Customer Reviews, IJCERT, Vol-3, Issue-9, September-2016, pp. 535-539.