OPINION MINING OF PRODUCT REVIEWS USING HYBRID MACHINE LEARNING TECHNIQUES

Nivet Chirawichitchai[1*] and Pisit Charnkeitkong[2]

Abstract

This research purpose's opinion mining of product reviews using hybrid machine learning techniques based on Thai online product reviews for hotel room services, hotels, and resorts with a collection of 4,000 sample data sets. A Modeling with Genetic Algorithms and Support Vector Machine methods. It consists of traditional machine learning to compare the effectiveness of each method in analyzing opinion mining of online review products. The experiment found that the use of Genetic Algorithms with support vector machines provides better classification accuracy than using traditional vector support machines with an accuracy of 88.64% and the proposed hybrid model can effectively reduce the dimensions of the data.

Keywords: Opinion Mining, Genetic Algorithms, Support Vector Machine

Introduction

One of the key aspects of any successful business is knowing how your customers feel about your brand and your products. They often freely express their views and opinions on social media, or in product reviews, surveys, and beyond, providing a wealth of information about your customer's thoughts and feelings. But with data growing by the day, it's impossible to manually analyze this mass of information. That's where opinion mining comes in. This natural language processing technology allows you to go beyond mere numbers and statistics, to automatically understand the feelings and emotions of your customers.

Opinion mining is a text analysis technique that uses computational linguistics and natural language processing to automatically identify and extract sentiment or opinion from within text (positive, negative, neutral, etc.). It allows you to get inside your customers' heads and find out what they like and dislike, and why, so you can create products and services that meet their needs. When you have the right tools, you can perform opinion mining automatically, on almost any form of unstructured text, with very little human input needed. Opinion mining can processthousands of pages, comments, emails, or surveys in just minutes for real-time results. Or you can perform opinion mining over time to see how sentiment classification rises or falls (Pang & Lee, 2008).

Opinion mining software allows you to train models to the specific terminology and criteria of your business for a consistently accurate and objective analysis of your customers' conversations. Save time and money and leave behind the wavering subjectivity of manual human processing (Janpla & Wanapiron, 2018).

Theory and Related Literature

Opinion mining or Text mining is a process that deals with messages (commonly used with a lot of texts) to find patterns, approaches, and relationships hidden in that thread. This technique uses information

[1*]Asst. Prof. Nivet Chirawichitchai, Ph.D., Faculty of Engineering and Technology, Panyapiwat Institute of Management, Thailand. Email: nivetchi@pim.ac.th

[2]Assoc. Prof. Pisit Charnkeitkong, Ph.D., Dean of the Faculty of Engineering and Technology, Panyapiwat Institute of Management, Thailand. Email: pisitcha_pim@pim.ac.th

Received: 10/01/2022; Revised: 25/01/2022; Accepted: 25/02/2022

retrieval, data mining, machine learning, statistics, and computational linguistics.

The text mining technique is similar to data mining but the data mining technique is often used with data or structured databases. The text mining technique focuses on unstructured or semi-structured texts (Khan, Baharudin, & Khan, 2009).

Genetic Algorithms (GAs)

Genetic algorithms or GAs are a mathematical model (Mitchell, 1998) are the best method for finding answers by using natural selection and species principles. The genetic algorithm is one of the calculations that evolves in the process of finding answers. It has been classified as one in the group of Evolutionary Computing which is currently recognized for efficiency and is widely applied to solve the optimization problem. The genetic algorithm is a process of finding answers to the system. It acts as a tool in calculating. The Cycle of Genetic Algorithms basically consists of 3 major key processes, as shown in Figure 1.

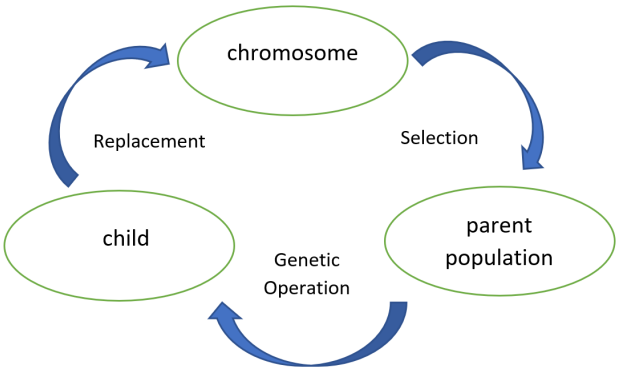

Figure 1 The Cycle of Genetic Algorithms

Genetic Algorithms (Sivanandam, 2008) process features

1. Selection: This is a step in selecting a good population in the system to be the origin of species and give birth to the next generation.

2. Genetic Operation: This process is a chromosome transforming process by means of species. It creates offspring which is obtained from combining the origin of species that has combined from parents or by changing the genes of parents in order to get new offspring.

3. Replacement: This process is bringing the new offspring to replace the old generations. It is a process of choosing which group of new generation and how many of them will be used to replace which group of the old generation.

The genetic algorithm simulates the evolution of life in natural systems, that is to say, the process inside a genetic algorithm makes the answer that existed in the system evolve in itself. This process leads to adaptation and becomes a better and best answer. From figure 1, it can be seen that the components in the Genetic cycle, the algorithm consists of population, origin, species, new species, and details of each element.

General procedures of genetic algorithm and connecting to real-world systems to search for the desired answer. The answer that the system requires genetic algorithms to search is in chromosome forms in the population group (the desired answer must be the best chromosome in the group). Therefore, the system will be able to know whether the answers contained in the genetic algorithm at a given time are good or bad by evaluating the chromosomes. The system will connect to the genetic algorithm through the objective function to evaluate the chromosomes for each procedure shown in the pseudocode below.

Comment P is Population
Initialize P
While not terminate
Evaluate P for fitness
P' = Selection.Crossover.Mutation of P
P = P'
Terminate

Answers close to the target (less than £; referring to the minimum amount before a new cycle of the order amount is refilled). The number of cycles exceeded the limit from the pseudocode, you will see that the iteration will begin repeatedly from step 2 (Population Evaluation) until the desired answer is obtained. The answer will come from the best chromosome in the group of population. The values from objective functions can be used in order to assess whether the answer is needed.

Genetic Algorithms features of each procedure

1. Population: Generate initial population which is usually created randomly.

2. Population Evaluation: Evaluate the population or chromosome of the entire population using an objective function due to the system is unable to understand the value of the chromosome within the genetic algorithm. Therefore, the chromosome must be decoded before being calculated with the objective function.

3. Fitness: Calculate the suitability then return to the genetic algorithm.

4. Selection: Select by the suitability of certain chromosome groups to be used as the origin of the species. These species will then be used as a substitute for the relay breed for the next generation.

5. Genetic Operation: Operate the crossover and mutation process by bringing the origin of species to create offspring with species operations. The chromosome obtained at this stage is the offspring chromosome.

6. Replacement: Replace the original population's chromosome with the offspring obtained from item 5. Some of the population will be replaced by a specific strategy. The replacement process has used the appropriateness of the decision.

Support Vector Machine

The theory is presented by Cortes and Vapnik (1995) to minimize errors from the prediction process. It is a technique used to solve data pattern recognition problems based on data classification by finding the decision plane and dividing the information into 2 parts. It tries to create the midline between the groups to have the most distance optimal separating hyperplane (Joachims, 1998) to find the decision plane for data dividing. It tries to create a dividing line between the groups to have the most distance between the boundaries of both groups by using the mapping function to move data from the input space to the feature space and creating a similarity measure function called the kernel function. In the media about SVM, we will variables used to make a decision as attribute and variable. They are used to determine a multi-dimensional plane called the feature. The most suitable selection is called a feature selection. The number of sets of features that are described in one instance (such as the row of predictive values) is called a vector. Therefore, the purpose of the SVM model (Joachims,1998) is to get the most out of the multi-dimensional plane that separates groups of vectors. In this case, the feature space is suitable for data that has a given dimension of data the most where (x_i, y_i), ..., (x_n, y_n) is an example used for teaching, n is the amount of data, m sample is the number of input dimensions and y is the result with + 1 or -1 as in equation. For linear problems, high dimensions are divided into 2 groups by the decision plane which can be calculated as in equation 1.

$$(W*X)+b=0 \quad (1)$$

Where w is the weight vaue an b is the ias value. The quation is used o classify the dat as in equation.

$$(w*x)+>0 \text{ if } yi = +1 \text{ and } (w*x) + b <0 \text{ if } y\_i= -1 \quad (2)$$

In the opinion mining of product reviews using hybrid machine learning

techniques to compare the effectiveness of opinion mining regarding online product reviews.

Research Methods

The method of this research aims to develop opinion mining of product reviews using hybrid machine learning techniques based on Thai online product reviews from Thai public resources. The conceptual framework for the development of sentiment analysis of Thai online product reviews as in Figure 2.

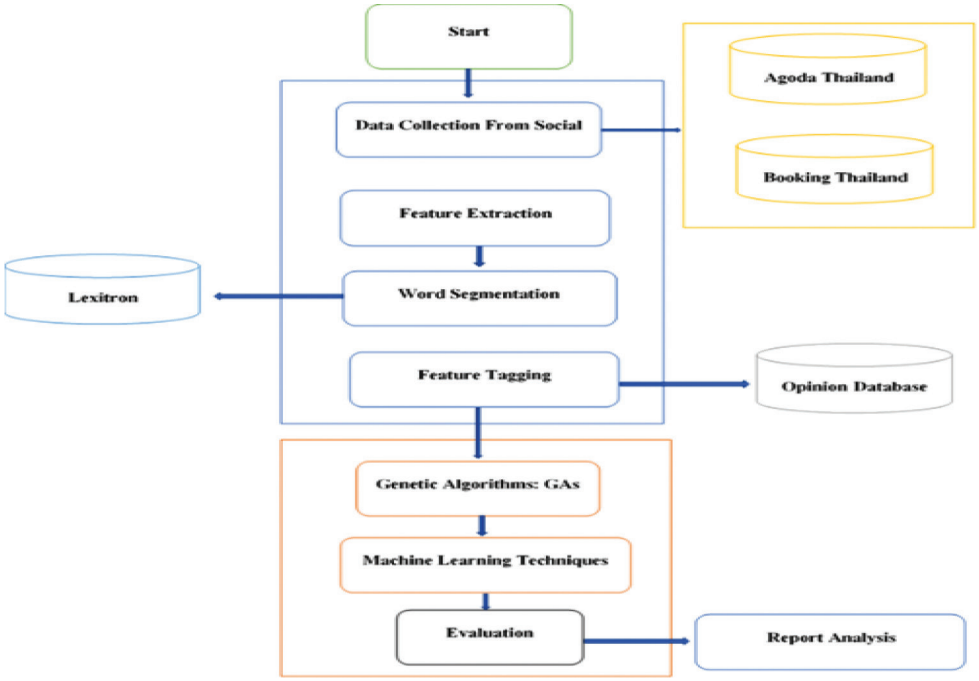

Figure 2 Conceptual framework of Opinion Mining of Product Reviews Using Hybrid Machine Learning Techniques

Data Preparation

This is a process of collecting sentiment analysis of Thai online product reviews that are used to create 4000 learning models from Agoda Thailand and Booking.com. The public news sources studied are shown in Table 1.

Table 1 Volume of opinion mining of Thai online product reviews

Thai public news sources	Number of messages
positive	2000
negative	2000

Document Parser

The process of extracting data from data sources then filtering the data by eliminating duplication. The datum was conversed into the appropriate format for modeling by using text to break words through the Thai word processing program called Thai Lexeme Tokenizer: LexTo (Chirawichitchai, 2013). Then tag the information to use for the training set and specify the categories of text by hand as in Table 2.

Table 2 Data tagging for the Training Set

Detail	Class
การ | บริการ | ดี | พนักงาน | ดู | น้อย | ไป | หน่อย | ส่วน |โรงแรม | ดู | สะอาด | แต่ | ก็ | ยังมี |สิ่ง อำนวยความสะดวก | ใน |ห้อง | ยัง | น้อย	Negative
บริการ | แย่มาก | จอง | เตียง | ใหญ่ | เตียง | เล็ก | พนักงาน | เช็คอิน | ช้า | ไม่มี | เตียง | จองไว้ | พนักงาน | พูดจา | แย่มาก | ลูกค้า | เสีย | แรง | จอง	Negative
เป็น | โรงแรม | ที่ | ดีเยี่ยม | มากๆ | คะ | สระ | ว่ายน้ำ | สะอาด สวยงาม | ห้องพัก | สะอาด | กว้างขวาง	Positive
บรรยากาศ | ดีมาก | วิว | ทะเล | สวย | มาก | สระน้ำ | ดีมาก คน | ไม่ | เยอะ | ห้อง | หรู | ไฟ |สว่าง	Positive

Feature extraction

The purpose of the feature extraction procedure is to extract the feature of the comments. The feature extracting should determine what is the representative of the comment feature and which value will represent that particular comment feature. Then use the value representing that comment feature by creating the Tokenize words and filter the tokenized words by specifying words that are between 2 to 25 characters long and filling in some stop words which is the removal of insignificant words without changing the meaning of the news. At this stage, the Rapidminer program is used to help extract the feature of opinion mining of product reviews (Chirawichitchai, 2015).

Stop-Word List Removal

This is to remove insignificant words without changing the meaning of the document. The insignificant words are words commonly used that has no significant meaning to the document when they are being removed from the document, they do not change the meaning. For example, prepositions or words that connect words or groups of words together, conjunctions or words that connect words with other words, and pronouns or words that are used in place of nouns that have already been mentioned in the sentence. Therefore, stop-words are considered insignificant to classify (Chirawichitchai, 2013).

Indexing

Due to computers are not able to directly classify documents in natural language, the documents should be converted to suit the computers' capability. The document's converting process is called indexing. This process creates a document representative to use in the learning process. The purpose of creating the indexing is to calculate values that will be used as document attributes. The indexing process is commonly used to begin creating a representative vector document and matrix of the document groups created from all of the document vectors in the group This research uses an experiment to assign weight values to the following index of Term Frequency–Inverse Document Frequency (TFIDF). This value is determined by the word frequency in the document multiplied by the log function of all documents and divided by the number of documents of that word occurring (Chirawichitchai, 2013).

Testing and Evaluation

Modeling of data classification test sentiment analysis of Thai online product reviews from Thai public resources. We considered the accuracy by using the ability

assessment of the model to measure the effectiveness of data classification according to the concept of information retrieval which is the measurement of the accuracy.

From the results of the opinion mining of product reviews using hybrid machine learning techniques to reduce the size of the data dimension. The result from the data dimension being reduced was being sent into the machine with effective learning solutions and conducted a comparative test of accuracy performance. When substituting the index values using the TFIDF method by 4 learning algorithms, it was found that the use of a genetic algorithm together with the support vector machine provides the most classification efficiency of 88.64%, followed by using the support vector machine only gave the classification efficiency of 87.17%.

Conclusion

In this research, the researchers present an opinion mining of product reviews using hybrid machine learning techniques based on Thai online product reviews from Agoda Thailand and Booking.com and create a model by using hybrid machine learning techniques in order to compare the accuracy values. We found that to reduce the characteristics with the genetic algorithm and a learning machine using the genetic algorithm together with the Support Vector Machine gave the best classification efficiency. It gave the highest accuracy of 88.64%, which will provide a good classification of feedback written from the feelings and emotions of users. This type of feedback can reduce the data dimension using the genetic algorithm. From the proposed process, we found that the data dimension reduction does not affect the efficiency of the data classification in any way and it can also be applied to other services.

References

Chirawichitchai, N. (2013). Automatic Thai Document Classification Model. *The Journal of Industrial Technology*, *9*(1), 142-149.

Chirawichitchai, N. (2013). Sentiment classification by a hybrid method of greedy search and multinomial naïve bayes algorithm. In *2013 Eleventh International Conference on ICT and Knowledge Engineering 2013* (pp. 1-4). Bangkok: Computer Science.

Chirawichitchai, N. (2015). *Developing term weighting scheme based on term occurrence ratio for sentiment analysis*. Berlin, Heidelberg: Springer.

Cortes, C., & Vapnik, V. (1995). Supportvector networks. *Machine Learning, 20,* 273-297.

Janpla, S., & Wanapiron, P. (2018). System framework for an intelligent question bank and examination system. *International Journal of Machine Learning and Computing*, *8*(5), 488-498.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98,* 137-142.

Khan K, Baharudin, B. B., & Khan, A. (2009). Mining opinion from text documents: A survey. In *2009 3rd IEEE International Conference on Digital Ecosystems and Technologies* (pp. 217-222). Istanbul: IEEE.

Mitchell, M. (1998). *An introduction to Genetic Algorithms*. Massachusetts, USA.: MIT Press.

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

Sivanandam, S. N. (2008). *Introduction to Genetic Algorithm.* Switzerland: Springer Science & Business Media Publisher.