https://doi.org/10.17576/apjitm-2018-0702-05

http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 7 No. 2, December 2018: 61 - 69
e-ISSN: 2289-2192

SENTIMENT ANALYSIS: AN ENHANCEMENT OF ONTOLOGICAL-BASED USING HYBRID MACHINE LEARNING TECHNIQUES

MUHAMMAD IQBAL ABU LATIFFI
MOHD RIDZWAN YAAKUB

ABSTRACT

With the fast development of World Wide Web 2.0 has resulted in huge number of reviews where the consumers share their opinion about a variety of products in the websites, forum and social media such as Twitter and Instagram. For the organizations, they have to analyze customer's behavior to find new market trends and insights. Sentiment analysis concept used to extract the positive, negative or neutral sentiment of the features from the unstructured data of product reviews. In this paper, we explore the techniques and tools used to enhance the ontology-based approach. Combination of ontology-based on Formal Concept Analysis (FCA) which a process of obtaining a formal ontology or a concept hierarchy from a group of objects with their properties and K-Nearest Neighbor (KNN) to classify the reviews. We believe with these techniques, we are able to view the strength and weakness of the product in more detail where the feature selection process will more be systematic and will result in the highest feature set.

Keywords: sentiment analysis, ontology, Formal Concept Analysis, K-Nearest Neighbor

INTRODUCTION

The advancement of today's information technology has made it easy for a variety of information and knowledge to be spread and access to the web. In addition, currently they are web application medium that users can interact to each other directly through various resources and services such as social media such as Facebook, Twitter, and Instagram, blogs and online commercial shops such as Amazon, eBay, and Lazada as mentioned in Haider (2012). Users can simply write their opinion whether agree or disagree with the products or services that they have purchased through online. Definitely, reviews and opinions are subjective expressions and require idea research to interpret the correct purpose of the reviewers.

Selling and buying products online now is a trend because users do not have to worry about to going out to the store, spending time choosing items and being disappointed when the items that they want is not up to what their think of, in another word, dissatisfied by them. Customers' satisfaction is the pivotal point for all business transaction. To solve this problem, data from customers' review are important, but having a huge data from different resources can be a difficult task, especially to gather right data at right time.

All these constraints require a solution to analyse the users' reviews or opinion and produce good output is a must for sellers, manufactures and customers. To figure out important data from customers' opinions and reviews that related to any fields and issues, sentiment analysis is the best answer and way to researchers (Liu 2012).

By using text mining, sentiment analysis focused on extracting features and opinionated word from customers' reviews. Current and past researchers used Natural Language Processing (NLP) and Machine Learning techniques to extract customers' review into informative knowledge. Our research is focusing mainly on NLP, especially Ontology and KNN from Machine Learning technique. The definition of ontology where the precise information and data model as a requirement of conceptualization to a selected scope. Furthermore, ontology is an object arrangement and its properties. The usage of ontology user can certainly point out the relationship between words in the text mining along the facility to determine the properties and annotations known as domain specific (Saranya, K & Jayanthy. S. 2017). In factual understanding, developing an ontology consists of defining classes in ontology, organizing the classes in the taxonomic hierarchy, defining slots and describing allowed value for these slots and filling in the value slots for instances (Noy & McGuinness 2001).

Beside the Ontology, a fundamental component of NLP where the characters of words and sentences recognized at this stage are the basic units passed to all next processing stages includes of analysis and tagging components called text pre-processing. There are phases in text pre-processing. Tokenization is the technique of breaking a flow of text into words, symbols, phrases or other important components called tokens. To identifying the meaningful keywords is the main use of tokenization. The variant forms of a word need to be merged into a common representation, which is the stem is called a stemming process. For example, the word "documentation", "documented", and "documenting" could all be reduced into a common presentation as "document". Meanwhile, stop word removal execute when numerous words in documents return routinely yet are necessary insignificant as they used to consolidate words in a sentence. Common words for stop words such as 'and', 'are', and 'this'. Normalization is the process involves transforming the text to ensure consistency. Some examples of this process include converting upper case letter to lower case and removing diacritics from letters, punctuations or numbers (Noy & McGuinness 2001).

Machine Learning is defined as a field of study that relates the computer science's principles and statistics. Construct statistical models, which used for upcoming predictions and identifying pattern in data is the main point of machine learning (Vijayarani, Ilamathi & Nithya 2015).

In this paper, domain ontology is used to extract the related ideas and attributes and then classify the polarity whether it is positive, negative or neutral by using K-Nearest Neighbor (KNN).

RELATED WORKS

In Pang, Lee & Vaithyanathan (2002), the authors proposed algorithms for sentiment classification using machine learning such as Support Vector Machine (SVM), Naïve Bayes and maximum entropy. They used the data from online movie reviews as their dataset. The data that collected is divided into three identical fold, controlling the placement of the class in each fold. Textual extraction was performed before the document is prepared in the initial HTML document format. This experiment show that machine learning algorithm only has unproductive result in sentiment classification's performance.

On the other hand, Sam (2013) proposed a system that has four modules where it can analysed the customer reviews from the social media. The creation of ontology, emotion and maintenance is used in an ontology management module are the basis of the other functional modules. Then the stop words for example preposition and articles is removed by the parser in the user query processing module. The purpose of soft parsing technique is to eliminate the missing words. Meanwhile, customer's review were withdrawn by the interface and placed in large local repositories occur in information foundation module.

Then, Ezhilarasi & Minu (2012) proposed an approach for emotion recognition and classification automatically. To find out the emotion, the emotional ontology is used. The usage of the parser is to find the useful and effective ontology. In this system, the text will be given as input to the NLP and it will be processed. In the beginning, the emotional ontology is to be initiated for classification. The basic emotions will be the classes in the emotion ontology. The emotion that represent by the words is to be identified from the domain knowledge. The extracted words are transfer to the WordNet to check for the proper and useful meaning and the synset is extracted. Ontology is created for all emotions from the collected synset. The usage of emotion ontology is for classifying sentences that so complicated. The text processing takes place in NLP. The word, which stands for the emotion, will be identified. Then, the term from the NLP has the access to the ontology and the exact class is identified. In the emotion classification, the ontology itself is used for the emotion classification.

A sentiment analysis using SWRL technique and fuzzy ontology to monitor the transportation actions like accidents, a polarity map and traffic rate is prepared as a result for customers proposed by Ali et al. (2017). SWRL is a language that makes results based on the rules that were described. The usage of the fuzzy ontology-based crawler is to collect data from various sources. Then, the pre-processing phase take place where it transforms the data into a section with a verb phrase and noun. Numerous queries are used to collect specific tweets and reviews from hazy social networks in the data cleaning phase. The datasets are restricted as negative, positive and neutral. The datasets are restricted through an n-gram technique to earn features from the opinionated phrase. The n-gram technique as a binary feature may extract a feature efficiently, but it may not be used to imply the overall opinion. Therefore, the proposed fuzzy ontology based semantic knowledge identifies features and the polarity is computed. There are rates used in a database called Sentiwordnet. Each opinionated word is given a positive or neutral or negative rate. A zero will be appointed to the opinion phrase if the rate of a sentiment word is absent in the SentiWordNet database. The next phase is an opinion lexicon. Sentiment analysis or opinion mining is reliant on two components. The first is the opinion lexicon and the next is computing polarity. The precision of the sentiment analysis is analyzed through the opinion lexicon. Therefore, a sentiment lexicon based on positive and negative opinion words is operated for sentiment analysis .

Doan et al. (2004) introduced a system named GLUE that operates learning techniques to semi-automatically produce a linguistic mapping between ontologies. Taxonomies are the main element of ontologies. The correlation among the taxonomies of two given ontologies are aimed and the mapping was identified. For the analysis of the system, a measure of closeness is computed.

Bhadane, Dalal & Doshi (2015) discussed the method which has the following steps, classification of aspects and polarity classification. The polarity of many phrases is domain and context particular. For figuring out the domain, the lexicons which imply a specific domain and special sentiments related to that domain are constructed. The training corpus used is such that every evaluation is annotated with an issue as well as associated sentiment. A model for element classification and a model per element for polarity classification is needed. For the element classification model, all of the lexicons in all of the opinions after pre-processing are used which tokenizes, prepends no longer and stems the sentence. Features with a usual frequency less than 4 or more than 30 and performing in opinions for more than 1 aspect are eliminated as they will be too unique and too common respectively and they may now not be beneficial. The lexicons from the opinions of a specific element after pre-processing are selected for the polarity classification models. The input is damaged around semicolon, period and comma. Next the recognition of elements will take place for every sub sentence. If adjoining sub sentences talk approximately equal element they are mixed collectively. Then the identification of polarity will take place. There are three components had to examine the final score included every element's polarity, rely on training instances and the rely on training instances of the specific element. A specific element topics greater for the users if many instances in training data talk about that element. Multiply the polarity through the quantity of instances of that element to calculate final score.

A mechanism to implement two methods, which are temporal sentiment analysis and causal rules extraction from tweets for event prediction proposed by Preethi, Uma & Kumar

(2015). From time and sentiment, transitory sentiment analysis is used to categorize the events. The plan and impact of the event are identified using sentiment causal relation and is mostly used for concluding events. The combination of these two methods to create an event prediction model, which predicts the period between the event and the sentiment of an upcoming event. The accuracy can figure out using techniques like RMSE and MAE. The connection between an event and a second event is referred to causality, where the second event is the physical consequence of the first. The first step is to extract the aspect keywords. SVM is used for sentiment classification of aspect keywords. Then the causal rule is to be identified between aspect keywords. Finally, the event is predicted from the causal rule found in the earlier step. Thus, the prediction is done using the two different techniques based on time period.

Meanwhile, the combined role of Part-of-Speech (POS) Tagging and Typed Dependency Relations (TDR) in distinguishing the relationship between features or aspects and sentiment words that proposed by (Ahmad, Yaakub & Bakar 2016) also can be highlighted as well. The researchers find the relation between words in the sentence through Stanford Typed Dependencies (STD). It supports a simple explanation of the grammatical relationship in a sentence. STD consists of 50 grammatical relations. Before that, researchers did the feature selection processed through the combination of method between Ant Colony Optimization (ACO) and KNN that produced optimum feature set. The good combination of TDR and POS Tagging method is adequate to analyse the relationship between features and sentiment words until to five layers TDR.

Choi et al. (2005) and Dave et al. (2003) used the reviews data from vehicles, movies and journey destination to find opinionated words. They classified words into two categories or classes that is positive or negative and count the polarity score for the textual content. If the document contained greater good than bad terms, they count it as positive, and vice versa. Those classifications are primarily based on document and sentence level type of Sentiment Analysis. It beneficial and boost the performance of a sentiment classification, however it cannot discover what the true intention of the viewers whether they preferred and not preferred. So, they are no longer constantly proper in some cases.

Formal Concept Analysis (FCA) is a fundamental approach of obtaining a concept hierarchy of correct ontology from a set of data and its properties. Previous researchers such as Kontopoulos et al. (2013) and Shein & Nyunt (2010) used FCA in their research. Both of the researchers combined the FCA method with different machine learning algorithms such as *OpenDover* and SVM. We proposed our new ontology based on FCA because concept hierarchy is having similarity with ontology architecture, especially in grouping the features.

A lot of surveys conducted about algorithms and computer applications that related to sentiment analysis. According to Obitko, Snášel & Smid (2004) the field that related to sentiment analysis such as emotion detection, knowledge exchange and building resources are mentioned. The currents traits and researches in the use of sentiment analysis over various applications are also mentioned.

CLASSIFICATION OF SENTIMENT ANALYSIS

Sentiment classification is a mechanism for extracting opinion or word that has sentiment value from text reviews then it will group into a number of categories or classes whether it positive, negative or neutral in its sentiment. It is also known as Sentiment Polarity. For the product review, text comments or emoticons in the comments or rating part is frequently used as data. Sentiment classification consists of:

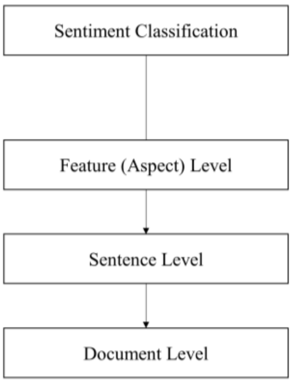

FIGURE 1. Level of Sentiment Classification

One of the approached used in sentiment analysis is lexicon based with information regarding which words or phrases have the polarity. The process of word's annotations is generally prepared manually and large sets of features are trained by the classifiers to classify new words or phrases. On other hand, sentiment analysis is a concern on mining the sentence or the entire document compared to depend on the type of the words itself. The main dilemma with document classification where the orientation of sentiment is calculated by overall sentiment properties of the whole document, while intended sentiment can be contained in just one word or sentence. We concentrate on the aspect level of sentiment classification to identify together with extract features that users commented, which produced a more accurate result than document and sentence level (Sam 2013) and (Yaakub et al. 2012).

PROPOSED SYSTEM

We proposed ontology based on FCA design where KNN classifier (Tavish Srivastava 2014) is employed to instruct the aspect based on supervised learning and the classifier will be classify the sentiments from the reviews whether the polarity positive or negative. We have several phases for the methodology as shown in Figure 2.

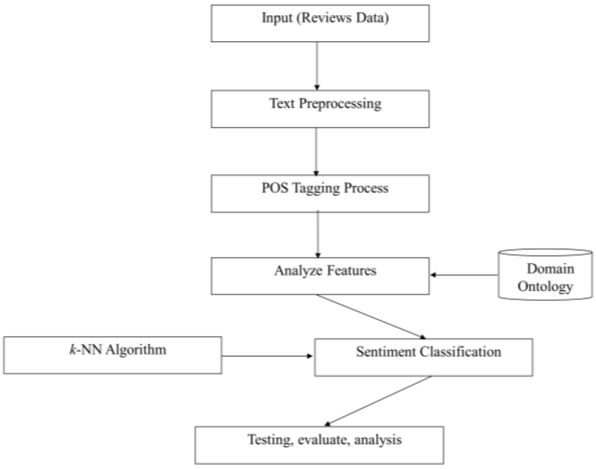

FIGURE 2. Phases in the Proposed System

Our works are in an initial stage. The idea of our works is shown in Figure 2.

Phase 1: Pre-processing of text: We implement the research over five datasets where five different types of the dataset used from different electronic products gather from the Amazon website. We choose this dataset because it is a benchmark dataset and it also used in Yaakub et al. (2012) and Hu & Liu (2004) which will be our baseline model in further works later. First, we get rid of the noisy data from the reviews before the next phase can take place. Most of the users or customers do not have complete understanding towards the language use; a spell in wrong way and ungrammatical error are detectable in the statistics. In addition, also carry observation out when comments that might be written with punctuation mistakes, in short form words, informal spellings, structures and words without capitalizing the phrases and others accurately and correctly.

Phase 2: Performing the POS Tags: POS Tag take place when to figuring out sentiment phrases. To implement this method Standard Parser is used (Ahmad, Yaakub & Bakar 2016). Phrases which include tags for adverbs, nouns, adjectives, also verbs in a sentence will be a tag by the POS Tagger. The capabilities related to the proposed system review domain will be extracted by the tags. Table 1 lists some of these POS tags.

TABLE 1. POS Tags

POS Tags	Description
NN	Noun
NNS	Noun plural
NNP	Proper plural, singular
NNPS	Proper noun, plural
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
VB	Verb, base form
RB	adverb

Phase 3: Creating the domain ontology: The use of domain ontology is to get the features that related to the certain domain and to provide knowledge about the certain domain, which can be comprehensible by way of both computers and creators and essential for expertise knowledge change and representation. The use of arrangement order, the concept do not have any various attributes due to insufficient for knowledge exchange of the information retrieval (Kontopoulos et al. 2013).

Thus, Formal Concept Analysis (FCA) is introduced. The function of FCA is to analyse data and create semantic structures, which are the formal idea of principles human opinions and pick out the theoretical systems among data set. As the current evolution of the Semantic Web and the organizing of ontologies since its predominant method for information representation, FCA has been considered as a precious engineering tool for developing an ontology from a set of objects and its properties (Obitko, Snášel & Smid 2004).

FCA has the following advantages such as:

1. The domain ontology is steadily developed, depend on the data set. Therefore, that it does not contain unnecessary concept or properties.
2. Concepts and concept hierarchies implicitly described, however are dramatically unique through the properties that have been detected. Enhanced ontology design and more extra classification of concepts are performed.
3. When the properties of various concepts are identical, then the concepts are the same.

A formal context K: = (*O*, *A*, *R*) where *O* is a set of objects, *A* is a set of attributes (properties), and *R* is an occurrence which shows the relationship between *O* and *R*. o*Ra* is a two relationship where (*o*,*a*)R, then the "object *o* has attributed *a*" or "the attribute a applies to the object o". A formal context K is represented as a cross table in which the column denotes *A*, the rows denote *O* and the existence relation *R* is represented by a series of crosses (Priya & Kumar 2015).

FCA design is developed based on the OWL (Web Ontology Language) and implementation of domain ontology used protégé 2000. KNN then will extract features that can be part of the domain.

Phase 4: Sentiment Classification: We chose a KNN for feature sentiment classification. Both classification and regression predictive problems can be solved by using KNN. However, it is more commonly used in the industry for classification problems. To figure out any technique we usually look at three important aspects:

1. Ease to interpret the output,
2. Calculation time,
3. Predictive power.

We primarily focus on how does the algorithm function and whether the output is affected by the input perimeter. KNN algorithm fair across all parameters of considerations. It is commonly used for its ease of interpretation and low calculation time as in the scale (Tavish Srivastava 2014).

TABLE 2. Comparison of Algorithms in Scale

	Logistic regression	CART	Random forest	KNN
Ease to interpret the output	2	3	1	3
Calculation time	3	2	1	3
Predictive power	2	2	3	2

KNN algorithm is choose because it is one of the simplest classification algorithms as showed in Table 2. Nevertheless, this algorithm we believe can give high competitive results (Tavish Srivastava 2014). Besides that, for regression problems this algorithm can be an approach.

Phase 5: Analysis: In this section, experimental expected results for the proposed algorithms are discussed. The researchers focus on identifying the relation between features and sentiment words in customers' comments. The sentence that contains the opinion on the product features is interested in the researchers. All the reviews will be checked by the proposed algorithm.

Calculation and analysis of the algorithm's overall performance, precision (P), recall (R) and F-score (F) is used in order to know the capability. The precision, recall and F-score may be determined via the method below.

$$\text{Precision} = \frac{TP}{TP+FP}, \qquad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN}, \qquad (2)$$

$$\text{F-score} = \frac{(2 * Precision * Recall)}{Precision + Recall}. \qquad (3)$$

True positive (*TP*) is the number of opinions from which the algorithms. The exact relations between sentiment words and features need to be extracted. Then, the amount of opinions from the algorithm that extracted the relations among features and sentiment phrase in a wrong way called false positive (*FP*). Finally, false negative (*FN*) where the algorithm cannot detect the connection between sentiment words and features in the variety of opinions (Yaakub et al. 2012).

CONCLUSION AND DISCUSSION

We proposed the combination technique of text pre-processing, POS Tagging, domain ontology based on FCA design and KNN classifier intend to decorate the sentiment classification. In knowledge representation, Ontology and FCA play an important role.

We believe this technique is expected to be the most convenient for sentiment analysis. The classification task will be enhance by the usage machine learning algorithms and the domain ontology to extract the reviews, plus we are able to view the strength and weakness of the product more detail where feature selection process will more systematic and will result the highest feature set. For future research, we expect to improving, refining and investigating for implied sentiment classification.

ACKNOWLEDGEMENT

We thank to Geran Galakan Penyelidik Muda for supporting this research under grant scheme GGPM-2016-006 and Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi for the opportunity to conduct this research.

REFERENCES

Ahmad, S.R., Yaakub, M.R. & Bakar, A.A. 2016. Detecting Relationship between Features and Sentiment Words using Hybrid of Typed Dependency Relations Layer and POS Tagging (TDR Layer POS Tags) Algorithm. *International Journal on Advanced Science, Engineering and Information Technology* 6(6): 1120.

Ali, F., Kwak, D., Khan, P., Islam, S.M.R., Kim, K.H. & Kwak, K.S. (2017). Fuzzy Ontology-based Sentiment Analysis of Transportation and City Feature Reviews for Safe Traveling. *Transportation Research Part C: Emerging Technologies* 7 (2017): pp. 33-48.

Bhadane, C., Dalal, H. & Doshi, H. 2015. Sentiment analysis: Measuring opinions. *Procedia Computer Science* 45(C): pp. 808–814.

Choi, Y., Cardie, C., Riloff, E. & Patwardhan, S. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*(2003): pp. 355–362.

Dave, K., Dave, K., Lawrence, S., Lawrence, S., Pennock, D.M. & Pennock, D.M. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of the 12th international conference on World Wide Web*: pp. 519–528.

Doan, A., Madhavan, J., Domingos, P. & Halevy, A. 2004. Ontology matching: A machine learning approach. *Science* Vol. 13: pp. 1–20.

Ezhilarasi, R. & Minu, R.I. 2012. Automatic emotion recognition and classification. *Procedia Engineering* 38: pp. 21–26.

Haider, S. 2012. An Ontology Based Sentiment Analysis. *A Case Study*. Univesity of Skovde.

Hu, M. & Liu, B. 2004. Mining Opinion Features in Customer Reviews. *19th national conference on Artifical intelligence*: pp. 755–760.

Kontopoulos, E., Berberidis, C., Dergiades, T. & Bassiliades, N. 2013. Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications* 40(10): 4065–4074.

Liu, B. 2012. Sentiment Analysis and Opinion Mining(May): pp. 1–108.

Noy, N.F. & McGuinness, D.L. 2001. What is an ontology and why we need it.

http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html [18 June 2018]

Obitko, M., Snášel, V. & Smid, J. 2004. Ontology Design With Formal Concept Analysis. *Proceedings of the International Workshop on Concept Lattices and their Applications (CLA 2004)*: pp. 111–119.

Pang, B., Lee, L. & Vaithyanathan, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*: pp. 79–86.

Preethi, P.G., Uma, V. & Kumar, A. 2015. Temporal sentiment analysis and causal rules extraction from tweets for event prediction. *Procedia Computer Science* 48(C): pp. 84–89.

Priya, M. & Kumar, C.A. 2015. A survey of state of the art of ontology construction and merging using Formal Concept Analysis. *Indian Journal of Science and Technology* 8(24): pp. 1-7.

Sam, K.M. 2013. Ontology-Based Sentiment Analysis Model of Customer Reviews for Electronic Products. *International Journal of e-Education, e-Business, e-Management and e-Learning* 3(6): pp. 477-482.

Saranya, K & Jayanthy. S. 2017. Learning Techniques. *International Conference of Innovations in information Embedded and Communication System*: pp. 1–18.

Shein, K.P.P. & Nyunt, T.T.S. 2010. Sentiment Classification Based on Ontology and SVM Classifier. *2010 Second International Conference on Communication Software and Networks*: pp. 169–172.

Tavish Srivastava. 2014. Introduction to KNN, K-Nearest Neighbors : Simplified. https://www.analyticsvidhya.com/blog/2014/10/introduction-k-neighbours-algorithm-clustering/ [25 June 2018]

Vijayarani, S., Ilamathi, J. & Nithya, M. 2015. Preprocessing Techniques for Text Mining - An Overview. *International Journal of Computer Science & Communication Networks* 5(1): pp. 7–16.

Yaakub, M.R., Li, Y., Algarni, A. & Peng, B. 2012. Integration of opinion into customer analysis model. *Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, WI-IAT 2012*: pp. 164–168.

Muhammad Iqbal Abu Latiffi
Dr. Mohd Ridzwan Yaakub
Center for Artificial Intelligence Technology,
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia.
eiqbal.latiffi@gmail.com, ridzwanyaakub@ukm.edu.my

Received: 2 Mei 2018
Accepted: 13 August 2018
Published: 18 December 2018