

# Feature Expansion using Lexical Ontology for Opinion Type Detection in Tourism Reviews Domain

Lim Jie Chen<sup>1</sup>, Gan Keng Hoon<sup>2</sup>

School of Computer Sciences, Universiti Sains Malaysia  
Penang, Malaysia

**Abstract**—Tourism reviews platform such as Trip Advisor become a major source for tourists to share their experiences and get some ideas for decision making. Since there are millions of reviews generated daily in the travel websites, tourist is often overwhelmed with huge information. This is where opinion type detection is important as it makes it easy for a tourist to obtain useful reviews for their understanding and planning processes based on the reviews' opinion type. The opinion type of texts in travel mostly involves different aspects of opinion related to the travel process, such as transportation, accommodation, price, food, entertainment, and so on. The challenge of this research is to improve this detection by proposing the lexical ontology approach to address the issue of out-of-vocabulary (OOV) keywords during a supervised detection of opinion type. Besides, there are also issues where the training data for detection has poor coverage or limited in a certain domain. In this paper, we propose a review opinion type detection approach by integrating the word (feature) expansion approach in machine learning. The suggested approach consists of two stages namely feature expansion and classification. For feature expansion, Lexical Ontology (LO) is used to expand the feature-related word to the domains such as synonyms. For classification, the expanded feature is corporate to the Machine Learning approach to detect the opinion type.

**Keywords**—Tourism domain; online review; opinion type detection; text classification; lexical ontology

## I. INTRODUCTION

Nowadays, tourists often rely on the online review when planning for their vacations such as Trip Advisor. TripAdvisor is the largest social travel website with about 500 million reviews of hotels, restaurants, attractions, and other travel-related businesses. Customer reviews provide reliable and valuable opinions about a tourist attraction such as its services, destination, and recommendations which helps tourists to understand more about a tourist attraction. However, due to the increase in the number of reviews recently, during decision making, it becomes difficult for the tourist to read all the reviews. Tourists are often overwhelmed and face difficulty in filtering relevant information from large number of reviews. Hence, it would be helpful if the opinion can be provided based on a certain type which is useful for decision making. For example, the tourism domain has Attractions, Concerts and Shows, Food & Drink, Transportation etc., job seekers domain contains Culture & Values, Work/Life Balance, Senior Management, Compensation and Benefits, and Career Opportunities. This is very important as online users trust customer reviews 12 times more than the product details provided by businesses [1]. Hence, automated identification of

reviews is important to help people to identify the opinion type of online reviews. This is the main reason of opinion type detection important. By applying opinion type detection, social network websites can structure user reviews in a fast and cost-effective way.

With this motivation, this paper focuses on opinion type detection where the problem of detection is formulated as the problem of text classification [2]. In the area of opinion analysis, classification is commonly used for topic classification or sentiment classification. In this research, our focus is on the topics or types in the context of tourism domain. For example, transportation, accommodation, food, entertainment, price and so on. Although the scope of contents of reviews is refined to tourism domain, nevertheless it is common to have various topics in the discussions. Hence, text classification in the tourism domain is an essential task to identify the topics mentioned in the text for further stage of analysis or application [3]. The outcome of this research is considered important in the decision making for both customers and the operators of tourism domain.

## II. RESEARCH PROBLEM

There are some issues in opinion type detection/classification that lead to the needs of feature expansion. It is common to encounter contents variations and word variations when a classification model is applied in a new target data, even from the same domain. A classifier trained to detect the opinion type, e.g. food, from one point of interest may not guarantee similar performance when applied in data from another point of interest. Two main reasons behind this are out-of-vocabulary (OOV) keywords and limited labeled data when training is performed.

### A. Out-Of-Vocabulary (OOV) Keywords

The first problem is caused by out-of-vocabulary (OOV) keywords during the detection involving new or unseen reviews. Once matching keywords are not captured in the model trained, this will create issue in the correct opinion type detection. For example, R1: the soup is very hot. Assumed that the dataset that we trained does not contain any keywords in R1, hence it is hard for a machine to determine its category. However, assume that there is keyword "spicy" in our training source. By using WordNet, the keyword "hot" in R1 shows the same meaning with the keyword "spicy" and they are in the same category. Expanding "hot" to "spicy" will improve the chances for machine to assign R1 to the correct opinion type. Fig. 1 shows the word "hot" has the same meaning as the word "spicy".

This paper is funded by the School of Computer Sciences, Universiti Sains Malaysia.

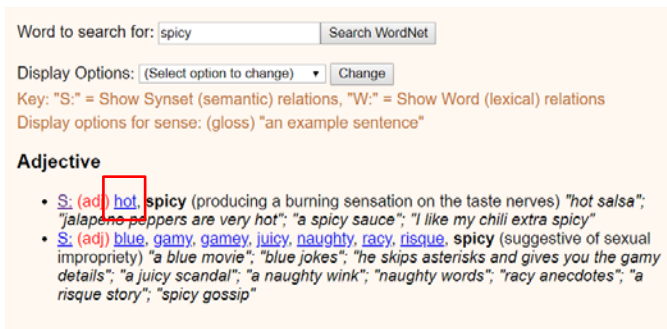


Fig. 1. A Snapshot of the Word “hot” has the same meaning as the Word “Spicy” in Princeton’s WordNet Web.

### B. Limited Labeled Data that Represents the Concept of an Opinion Type

For example, R2: soup is very spicy and R3: soup is very tasty. Since R2 and R3 belong to the same concept of taste, they should be able to be detected as the same type. However, the machine might not determine they are in the same category. Assumed R3 is the upcoming reviews and the word “spicy” is in our training, by using WordNet, the keyword “spicy” in R2 and “tasty” in R3 shows they are similar to each other and they are in the same category. Thus, this improve the chances for machine to assign R2 and R3 to the correct opinion type. Fig. 2 shows the word “tasty” is similar to the word “spicy”.



Fig. 2. A Snapshot of the Word “Tasty” is Similar to the Word “Spicy” in Princeton’s WordNet Web.

## III. LITERATURE REVIEW

This section covers literatures related to different approaches of opinion type detection, using bag of words (BOW) approach. Works related to feature expansions are also discussed.

In a BOW, the words in a matrix do not represent sentences with structure and grammar, and the semantic relationship between these words are ignored in the construction of BOW representation. Another limitation of BOW is on its semantic meaning, basic BOW approach does not consider the meaning of the word in the document. It ignores the context in which it is used. The same word can be used in multiple places based on the context of nearby words [4].

According to E. Rudkowsky *et al.* [5], in the domain of social science, the use of word embeddings introduces a new approach to the field of sentiment analysis in the social

sciences that offers potential to improve on current bag-of-words approaches. The main advantage of using word embeddings is its potential to detect and classify unseen or out-of-context words that are not included in the training data. Vector representations of text that allocate similar words closer to each other, such approach can supplement training data, which is promising in improving the results of machine learning tasks.

According to Sneha [6], the very first step of sentiment classification is to extract the phrases containing adverbs and adjectives in the review because they are good indicators of subjectivity. However, single-word adjectives and adverbs may have different meanings in different contexts where they modify the meaning of other words quickly. It is not sufficient to reply on single adjectives and adverbs as potential opinion word, noun and verb may represent aspects or its attributes in the review. Therefore, rather than selecting single word adjective or adverb, bigrams which contain noun phases with adjective and adverb are better choice.

From an e-commerce perspective, M. Hu *et al.* [7] and K. Vivekanandan *et al.* [8] have proposed a frequency-based method for aspect extraction. In this approach most frequent words in reviews usually, nouns and pronouns are considered to be candidate of aspects. However, S. Abeyasinghe *et al.* [9], then improves the method by applying part-of-speech patterns to filter the terms added to the frequency terms as well. Another approach is using syntactic relations in words to determine the aspects.

For feature expansion techniques, several investigations have attempted to improve the out of vocabulary keywords problem. S. M. Rezaeinia *et al.* [10] conducts research in word embedding method, and they found that their Improved Word Vectors (IWV) which is based on the combination of natural language processing techniques, lexicon-based approaches and Word2Vec/GloVe methods which increased the accuracy of pre-trained vectors in sentiment analysis. They proposed a method that gets a sentence and returns improved word vectors of the sentence. They used Word2Vec which is based on continuous Bag-of-Words (CBOW) and Skip-gram architectures which can provide high quality word embedding vectors.

In the tourism domain, Muhammad Afzaal *et al.* [11] presented an aspect-based sentiment classification framework using tree-based aspects extraction method that classifies opinions/reviews of aspects into positive or negative. The opinion-less and irrelevant sentences are first removed by employing Stanford Basic Dependency on each sentence and the features are extracted from the remaining sentences with N-Grams and POS Tags to train the classifiers. Therefore, the limitation is that some opinion-less/irrelevant texts might be an important source/text for opinion type in the classification process. Removing them may result in OOV issue.

K. Soo-Min *et al.* [12] develops an automatic algorithm to produce opinion-bearing words by hybridizing two methods. First method is a small set of human-annotated data that shows that productive synonyms and antonyms of an opinion-bearing word can be found through automatic expansion in WordNet and use them as feature sets of a classifier. They also use all

synonyms of a given word as well as the word itself to determine a word's closeness to opinion-bearing or non-opinion-bearing synonym set.

TABLE I. SUMMARY ON GAP ANALYSIS ON OPINION TYPE DETECTION

|                             | Paper     | Method  | Problem Domain   | Dataset   |
|-----------------------------|-----------|---|--|---|
| Bag-of-Word                 | [6]       | Words representation of text content, ignoring grammar and order of appearance are ignored. The limitation of <b>Bags of Words (BOW)</b> is on its semantic & contextual meaning.   | Automatic topic identification of health-related messages in online health community | Health-related messages (4041) in three message boards: treatment board, emotional support board and survivorship board.                |
|                             | [7]       | Word embedding method outperform <b>Bags of Words (BOW)</b> improving detection and classification of unseen or out-of-context words that are not included in the training data.  | Sentiment analysis in social sciences  | 20580 sentences from party press releases, transcripts of parliamentary speeches, and media reports.                                    |
| Natural Language Processing | [8]       | <b>Single-word adjectives and adverbs</b> may have different meanings. Rather than selecting single word adjective or adverb, <b>bigrams</b> which contain <b>noun phrases with adjective and adverb</b> are used.                                  | Sentiment classification   | Multi-Domain Sentiment Dataset (667 reviews) - Book, DVDs, Kitchen Appliances, and Electronics Appliances                               |
|                             | [9], [10] | <b>Frequency-based method</b> for aspect extraction. Most frequent words in reviews comprising of nouns and pronouns are considered to be aspects.  | Mining opinion features in customer reviews  | Customer reviews of five electronics products - from Amazon.com and C net.com.  |
| Lexical Ontology            | [14]      | Automatic <b>expansion using WordNet</b> and use it as feature sets of a classifier. Used all <b>synonyms</b> of a given word.  | Sentence-level opinion detection system  | Collections of opinion-bearing (2683) and non-opinion-bearing words (2548) manually from Columbia University                            |
|                             | [13]      | Extracts <b>frequent nouns and noun phrases</b> from reviews text. Groups similar nouns using <b>WordNet</b> . <b>Decision tree</b> is employed on reviews where review words are used as internal nodes and extracted nouns as the leaf of a tree. | Aspect-based sentiment classification for tourist reviews                            | Restaurant (2000 reviews) and hotel (4000 reviews) domains dataset-collected from popular social media websites using crawler and APIs. |

To wrap up the literatures for this research, we have summarized some of the relevant works in three aspects, i.e. Bag-of-Word, Natural Language Processing and Lexical Ontology in Table I.

IV. RESEARCH METHODOLOGY

In this section, we present our proposed work on the Opinion Type Detection framework for training and detecting review sentence's opinion type as shown in Fig. 1. In this framework, there are three main steps which are text pre-processing, feature expansion, and classification. The input to this framework is reviews sentence with its corresponding opinion type and the output is the accuracy of the model.

Fig. 3 shows the pipeline for opinion type detection using a supervised learning algorithm. In text pre-processing task, the datasets are collected, categorized, cleaned, and sorted based on some filtering task. In feature expansion, the expansion can be applied on features obtained from methods such as bags-of-words (BOW) and Natural Language Processing (NLP). In this study, we propose a Lexical Ontology (LO) approach to improve the opinion type detection in the tourism review. After feature expansion, machine learning approach such as Naïve Bayes (NB) classifier, Support Vector Machine (SVM), and Decision Tree (DT) is applied.

A. Tourism Review

Tourism review is a review made by a consumer who has experienced gain from travelling (see Fig. 4). Customer reviews provide true and valuable opinions about a tourist attraction which helps tourist to understand more about tourist attraction when making a decision. In tourism domain, there are some important opinion types in which a review can be categorized such as "Attraction", "Fee", "Time", "Weather", "Transport", "Service" and "Food".

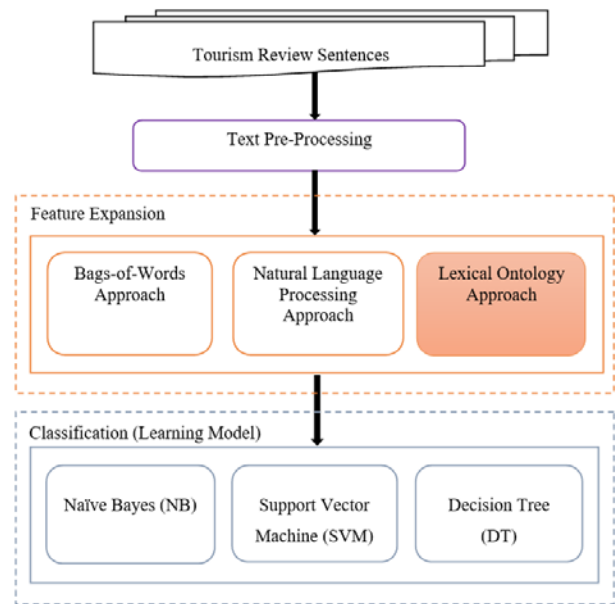


Fig. 3. The Design Framework of Opinion type Detection.

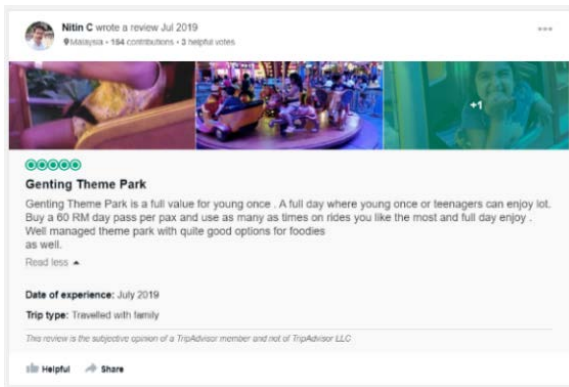


Fig. 4. Example of a Tourism review at Trip Advisor Website.

### B. Text Pre-Processing

Text pre-processing is the process the cleans and prepare the text prior to classification process. Since real world data often contain noise and formatting errors, in pre-processing step, these unnecessary data will be removed to improve the quality of the input data. In our case, input that will be pre-processed are sentences of review text. Each sentence will go through the following steps:

#### Removing punctuation and convert text to lowercase:

Each review sentence is converted to lowercase and has its punctuation removed.

Example: “Genting Theme Park is a full value for young once.”

After this step, the sentence will become “genting theme park is a full value for young once”.

**Tokenization:** A review sentence is treated as a string and split into a list of tokens.

Example: “genting theme park is a full value for young once”

After tokenization step, the sentence is divided into tokens, “genting”, “theme”, “park”, “is”, “a”, “full”, “value”, “for”, “young”, “once”.

**Removing stop words:** Stop words such as “the”, “a”, “and” etc. occur frequently, but do have significant role in the semantic/context of the text. Removing stopwords can potentially help improve the performance as there are fewer and only meaningful words retained. Thus, it could increase classification accuracy.

Example: “genting theme park is a full value for young once”

After removing stop words the tokens will be “genting”, “theme”, “park”, “full”, “value”, “young”.

**Lemmatization:** Lemmatization reduces a token in inflected form to the root form, called Lemma. A lemma is the canonical or dictionary form of a word.

Example: “runs”, “running”, “ran” will be transformed to its root form “run”.

### C. Feature Expansion

#### Bag-of-words (BoW)

Bag-of-words method is a simple representation of features (e.g. in word token form) obtained from the text documents. The model consists of bag, i.e. multiset, of words, where grammar rules are disregards. Word counts are represented in this model [13]. This method is often used in

1. Natural Language Analysis.
2. Document Classification.
3. Information Retrieval.

An illustration of how word vectors are generated in bag-of-words model is shown below. Given two sentences,

1. Dad likes to watch movies. Mum likes movies too.
2. Dad also likes to watch indoor games.

These two sentences can be represented as follows as a collection of words.

1. [“Dad”, “likes”, “to”, “watch”, “movies”, “Mum”, “likes”, “movies”, “too”]
2. [“Dad”, “also”, “likes”, “to”, “watch”, “indoor”, “games”]

Remove the duplicate words and use the word count for the representation.

1. {“Dad”:1, “likes”:2, “to”:1, “watch”:1, “movies”:2, “Mum”:1, “too”:1}
2. {“Dad”:1, “also”:1, “likes”:1, “to”:1, “watch”:1, “indoor”:1, “games”:1}

Combine both sentences with their word frequency.

{“Dad”:2, “likes”:3, “to”:2, “watch”:2, “movies”:2, “Mum”:1, “too”:1, “also”:1, “indoor”:1, “games”:1}

Hence, we obtain a vocabulary for this small collection and by using this vocabulary we can create vectors for our sentences.

The length of the vector must be equal to the vocabulary size. Here the length of the vector is ten. Then, by comparing the sentences with the vocabulary and we get the vectors as follows.

Dad likes to watch movies. Mum likes movies too.

[1, 2, 1, 1, 2, 1, 1, 0, 0, 0]

Dad also likes to watch indoor games.

[1, 1, 1, 1, 0, 0, 0, 1, 1, 1]

The size of vector is proportionate to the size of vocabulary. Hence, for document with long texts, the size of the vocabulary is high. This also cause the vector to contain a greater number of zeros. It is called sparse matrix and the sparse matrix require more memory and high computational power [13].

### Natural Language Processing (NLP)

Useful features can be identified using natural language analysis process like Part-of-Speech (POS) tagging. By tagging each word in terms of its POS, such as noun, pronoun, adverb, adjective, verb, etc. the syntactical meaning of the word can be used as reference to select relevant features. Table II shows the POS tags and their related meanings while Table III shows the example of the generation of lexical elements for one of the review sentences.

### Lexical Ontology (LO)

To address the OOV words and limited labelled data issues, we propose to include Lexical Ontology in the opinion type detection task by expanding the features for each review sentences. Given a review, features are extracted from the reviews. Then, the features/keywords will be expanded by synonyms using WordNet, a well-known Lexical Ontology. We make assumption that these additional features can improve the accuracy of opinion type detection. Based on the two basic methods of features identification (BoW and NLP), we perform the expansion on four variants of feature sets: feature set  $F_{BOW}$ ,  $F_{NLP}$ ,  $F_{BOW+LO}$  and  $F_{NLP+LO}$ . Feature set  $F_{BOW}$  and  $F_{NLP}$  are used as the baseline feature set to assess the performance of the proposed expanded feature sets ( $F_{BOW+LO}$  and  $F_{NLP+LO}$ ).

TABLE II. POS TAGS AND MEANING

| POS TAGGING            | MEANING   |
|------------------------|-----------|
| JJ                     | ADJECTIVE |
| RB, RBR, or RBS        | ADVERB    |
| NN or NNS              | NOUN      |
| VB, VBD, VBN, VBZ, VBG | VERB      |

TABLE III. THE GENERATION OF LEXICAL ELEMENTS FOR A REVIEW SENTENCE

| STEPS                    | EXAMPLE   |
|--------------------------|---|
| A review sentence        | Genting Theme Park is a full value for young once.  |
| POS Tagging and Stemming | Genting/NNP<br>Theme/NNP<br>Park/NNP<br>is/VBZ<br>a/DT<br>full/JJ<br>value/NN<br>for/IN<br>young/JJ<br>once/RB<br>./. |
| Nouns                    | Genting, Theme, Park, value   |

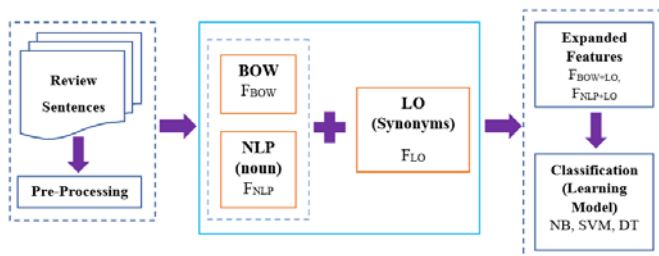


Fig. 5. General Process of Feature Expansion using Lexical Ontology.

### Base Features Extraction

Fig. 5 shows the process of feature expansion. A review sentence will first go through the pre-processing step, followed by Base Feature Extraction. The outcome for BOW method will be the feature of each review,  $F_{BOW} = \{f_{b1}, f_{b2}, \dots, f_{bn}\}$ . After extraction, a list of features for each review  $F_{BOW} = \{f_{b1}, f_{b2}, \dots, f_{bn}\}$  will be stored for feature expansion.

Similar feature set will be extracted using the NLP method, resulting in  $F_{NLP} = \{f_{n1}, f_{n2}, \dots, f_{nn}\}$ .

### Feature Expansion

```

Input: Features for each review,  $F_W = \{f_{w1}, f_{w2}, \dots, f_{wn}\}$  with its opinion type, where  $F_W$  can be  $F_{BOW}$  or  $F_{NLP}$ 
Output:  $F_{W+LO} \leftarrow \{f_{w1}, f_{w2}, \dots, f_{wn}, f_{w1+LO(1)}, f_{w1+LO(2)}, \dots, f_{w1+LO(n)}, f_{w2+LO(1)}, f_{w2+LO(2)}, \dots, f_{w2+LO(n)}, f_{wn+LO(1)}, f_{wn+LO(2)}, \dots, f_{wn+LO(n)}\}$ 

for all  $F_W$  do
     $F_{LO} \leftarrow$  expand  $F_{BOW}$  with WordNet (Synonyms)
     $F_{LO} \leftarrow \{f_{w1+LO(1)}, f_{w1+LO(2)}, \dots, f_{w1+LO(n)}, f_{w2+LO(1)}, f_{w2+LO(2)}, \dots, f_{w2+LO(n)}, f_{wn+LO(1)}, f_{wn+LO(2)}, \dots, f_{wn+LO(n)}\}$ 
     $F_{W+LO} = F_W + F_{LO}$ 
     $F_{W+LO} \leftarrow \{f_{w1}, f_{w2}, \dots, f_{wn}, f_{w1+LO(1)}, f_{w1+LO(2)}, \dots, f_{w1+LO(n)}, f_{w2+LO(1)}, f_{w2+LO(2)}, \dots, f_{w2+LO(n)}, f_{wn+LO(1)}, f_{wn+LO(2)}, \dots, f_{wn+LO(n)}\}$ 
     $F_{W+LO} \leftarrow$  Remove Duplicate feature
end for
return  $F_{W+LO}$ 
    
```

Fig. 6. Pseudocode for Process of BOW+LO/NLP+LO Approach.

Fig. 6 shows the pseudocode for applying LO approach to BOW method and NLP method. By expanding the feature in  $F_W$ , the output will be stored in  $F_{LO}$  where  $F_{LO}$  is the feature obtained by expanding  $F_W$  with WordNet (Synonyms). Then, both  $F_W$  and  $F_{LO}$  are combined,  $F_{W+LO}$ . Using a sample review "There is an area for arcade games, again, maybe more suitable for children." from "Genting Highlands Theme Park" POI, the features/keywords selected for four variant approaches is shown in Table IV. The steps of how the features are expanded are described in the subsequent paragraphs.

For  $F_{BOW+LO}$ , the review sentences go through pre-processing, which includes remove punctuation, stemming, tokenization, and remove stop words. This resulted the base features from the review, i.e. ['area', 'arcade', 'game', 'maybe', 'suitable', 'child'] as  $F_{BOW}$ .

TABLE IV. FEATURES/KEYWORD SELECTED FROM ONE OF THE SAMPLE REVIEWS

| Symbol       | Feature/Keywords   | Approach |
|--------------|--|----------|
| $F_{BOW}$    | ['area', 'arcade', 'game', 'maybe', 'suitable', 'child']                   | BOW      |
| $F_{NLP}$    | ['area', 'game', 'child', 'maybe', 'suitable']                             | NLP      |
| $F_{BOW+LO}$ | ['area', 'arcade', 'game', 'maybe', 'suitable', 'child', 'unfit', 'brave'] | BOW+LO   |
| $F_{NLP+LO}$ | ['area', 'game', 'child', 'unfit', 'brave', 'suitable', 'maybe']           | NLP+LO   |

Similarly, for  $F_{NLP+LO}$ , all review sentences will be pre-processed as well, followed by POS tagging. Feature which is noun and other main tag types are selected as  $F_{NLP+LO}$  features, i.e. ['area', 'game', 'child', 'maybe', 'suitable'].

Then, we expand the base features with synonyms from LO, that results in  $F_{BOW+LO}$  ['area', 'arcade', 'game', 'maybe', 'suitable', 'child', 'unfit', 'brave'] and  $F_{NLP+LO}$  approach ['area', 'game', 'child', 'unfit', 'brave', 'suitable', 'maybe'].

Fig. 7 shows the process of expanding the keywords from the baseline methods (NLP and BOW). First, the features/keywords of BOW method reviews (training source) are generated,  $(f_{b1}, f_{b2}, \dots, f_{bn})$ . Then the expansion goes through each feature, e.g.  $f_{b1}$  means that the first feature is selected from  $T_1$  training review (and the rest can be done in the same manner). Then, the expanded features are obtained using LO's synonym,  $(f_{b1+LO(1)}, f_{b1+LO(2)}, \dots, f_{b1+LO(n)}, f_{b2+LO(1)}, f_{b2+LO(2)}, \dots, f_{b2+LO(n)}, f_{bn+LO(1)}, f_{bn+LO(2)}, \dots, f_{bn+LO(n)})$ .  $f_{b1+LO(1)}$  refers to the new feature that is expanded from the first feature. Finally, the BOW+LO features set can be generated by combining the features from the baseline and new features, which is called "BOW+LO approach". The same process applies to NLP method.

#### D. Classification

An important step in the opinion type detection pipeline is choosing a good classifier. This can be done by adopting different type of classifiers and measure their performances to serve as a guideline in the selection. Supervised machine learning models such as Naïve Bayes (NB) classifier, Support Vector Machine (SVM), and Decision Tree (DT) are chosen due to their popularities. Review data set for each opinion type is also split into training and test datasets to train and test the model.

1) *Naïve Bayes (NB)*: Naïve Bayes classifier has been widely used for document categorization tasks [14]. It is theoretically based on Bayes theorem, which was developed by Thomas Bayes [15]. Recent studies show that NB is commonly used in information retrieval [16]. Naïve Bayes classifier is a generative model, which is a traditional method of text categorization. This classifier is chosen as the since it is the common base for classification task.

If the number of documents ( $n$ ) fits into  $k$  categories where  $k \in \{c_1, c_2, \dots, c_k\}$ , the predicted class as output is  $c \in C$ . The Naïve Bayes algorithm can be described as follows [17], [18]:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

$$P(d|X) = P(x_1|d) \times P(x_2|d) \times \dots \times P(x_n|d) \times P(d)$$

where  $d$  is document and  $c$  indicate classes.

where:

$$P(c|d) = \text{posterior probability. The probability of hypothesis } h \text{ being true, given the data } x, \text{ where } P(c|d) = P(d_1|c)P(d_2|c) \dots P(d_n|c)P(c)$$

$P(d|c)$  = Likelihood. The probability of data  $d$  given that the hypothesis  $h$  was true.

$P(c)$  = Class prior probability. The probability of hypothesis  $h$  is true

$P(d)$  = Predictor prior probability. The probability of the data (irrespective of the hypothesis)

$$C_{MAP} = \arg \max_{c \in C} P(d|c) P(c) \\ = \arg \max_{c \in C} P(x_1, x_2, \dots, x_n|c) p(c)$$

2) *Support Vector Machine (SVM)*: The original version of the Support Vector Machine (SVM) was formulated by Vapnik and Chervonenkis and adapted this version into a nonlinear formulation. SVM is a powerful statistical machine-learning technique [20]. Due to the ability to handle millions of inputs and good performance, SVM was widely used in text classification studies. SVM was originally designed for binary classification tasks. However, many researchers work on multi-class problems using this technique [19].

Since SVMs are traditionally used for the binary classification, a Multiple-SVM (MSVM) for multi-class problems is proposed by [20]. One-vs-One is a technique for multi-class SVM that builds  $N(N - 1)$  classifiers as follows [21]:

$$f(x) = \arg \max_i \left( \sum_j f_{ij}(x) \right)$$

3) *Decision Tree (DT)*: A decision tree is a tree whose internal nodes are tests and leaf nodes are categories. Each internal node test one attribute and each branch from a node selects one value for the attribute. The attribute used to make the decision is not defined. So, attribute which gives maximum information can be used and the leaf node predicts a category or class. The decision trees are not limited to Boolean functions, but they can be extended for general categorically values functions [22].

$$H\left(\frac{p}{n+p}, \frac{n}{n+p}\right) = -\frac{p}{n+p} \log_2 \frac{p}{n+p} - \frac{n}{n+p} \log_2 \frac{n}{n+p}$$

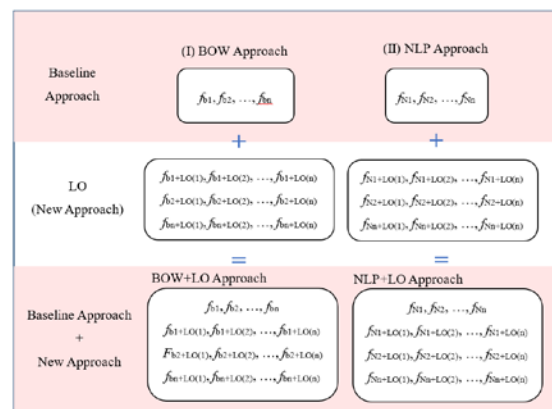


Fig. 7. Process of Feature Expansion by Lexical Ontology Approach.

V. EXPERIMENTS

A. Dataset

The data collection for this research is tourism review data about a tourism place, i.e. point of interest. The dataset is collected from reviews written by users regarding a point of interest, such as Penang Hill. The data is collected from Trip Advisor website. In this evaluation, a total of five point of interest are identified, and their reviews are collected. These five POIs are Genting Highlands Theme Park, Cameron Highlands Boh’s Tea Centre, Club Med Cherating Beach, Escape Penang, and Penang Hill. For each POI, we have collected 50 reviews within a certain date range. If a point of interest has more reviews, the date range will be shorter and vice versa. Table V lists the Point of Interests and Reviews Data Range.

In this dataset, each review will be stored at the sentence level. Fig. 8 shows the output of the sentence segmentation of reviews. Table VI shows a listing of sentences for one example POI.

B. Data Benchmarking

To prepare the golden standard data collection, the reviews sentences are annotated based on the opinion types, i.e. “Attraction”, “Fee”, “Time”, “Weather”, “Transport”, “Service” or “Food”. A total of three annotators are recruited to perform the annotation on the sentences.

TABLE V. LIST OF POINT OF INTEREST AND REVIEWS DATA RANGE

| Point of interest (POI)            | Date       |            | Number of Reviews Collected |
|------------------------------------|------------|------------|-----------------------------|
|                                    | From       | To         |                             |
| Genting Highlands Theme Park       | 19/2/2019  | 17/8/2019  | 50                          |
| Cameron Highlands Boh’s Tea Centre | 12/5/2019  | 8/11/2019  | 50                          |
| Club Med Cherating Beach           | 2/12/2019  | 21/1/2020  | 50                          |
| Escape Penang                      | 26/7/2019  | 16/8/2019  | 50                          |
| Penang Hill                        | 26/11/2019 | 30/12/2019 | 50                          |

TABLE VI. LISTING OF SENTENCES FOR ONE POI

| Review Sentence  |
|--|
| lots of indoor fun activities for kids and adults.       |
| Pocket friendly as well.                                 |
| We had a great time and 5 nights flew by so quickly.     |
| Totally rested despite the rain.                         |
| went there by cable car.                                 |
| Self Service check-in was a breeze.                      |
| There must be hundreds of food and drink options around. |

Before an annotator starts his task, he is given the guideline and a set of data for the annotation task. Using the guideline of the definition of opinion type, the annotator is required to manually annotate each review as one of the seven opinion types, for example, “Attraction”, “Fee” based on his judgement on the opinion type that best match the content of the sentence. If the review sentence does not fit any of the seven opinion types, the review will be annotated with “N/A”.

C. Data Statistics

From the 250 reviews data, 1691 sentences were obtained from the sentence segmentation. From a total of 1691 sentences, a total of 1576 review sentences are annotated with the opinion type. If there is more than an opinion type for a sentence, the best type will be chosen. Table VII shows the number of review sentences collected for each point of interest. Since 50 reviews will be collected is based on POI, and each review will have a different number of sentences depending on the length of the review, hence, it is natural to have difference numbers of review sentences for different POI.

As the data collection was made based on POI, there are also distinct differences in the number of review sentences related to each opinion type. Some opinions tend to be mentioned more in the reviews, e.g. Attraction, compared to Weather in this data collection. The number of review sentences for each opinion type is shown in Table VIII and Table IX (with POI).

TABLE VII. NUMBER OF REVIEW SENTENCES FOR EACH POI

| Point of Interest (POI)            | Number of Reviews Sentences |
|------------------------------------|-----------------------------|
| Genting Highlands Theme Park       | 275                         |
| Cameron Highlands Boh’s Tea Centre | 410                         |
| Club Med Cherating Beach           | 408                         |
| Escape Penang                      | 300                         |
| Penang Hill                        | 298                         |
| Total                              | 1691                        |

TABLE VIII. NUMBER OF REVIEW SENTENCES FOR EACH OPINION TYPE

| Opinion Type | Number of Reviews Sentences |
|--------------|-----------------------------|
| Attraction   | 631                         |
| Fee          | 86                          |
| Time         | 183                         |
| Weather      | 33                          |
| Transport    | 74                          |
| Service      | 229                         |
| Food         | 154                         |

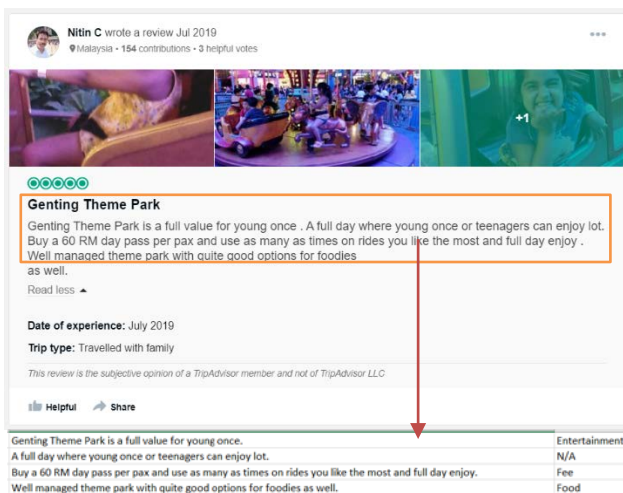


Fig. 8. Output of Sentence Segmentation of Reviews.

TABLE IX. NUMBER OF REVIEW SENTENCES COLLECTED FOR EACH POI FOR ITS CORRESPONDING OPINION TYPE

| POI \ Opinion Type | Genting Highlands Theme Park | Cameron Highlands Boh's Tea Centre | Club Med Cherating Beach | Escape Penang | Penang Hill |
|--------------------|------------------------------|------------------------------------|--------------------------|---------------|-------------|
| Attraction         | 134                          | 120                                | 123                      | 143           | 111         |
| Fee                | 16                           | 21                                 | 4                        | 5             | 40          |
| Food               | 23                           | 44                                 | 64                       | 12            | 11          |
| Service            | 18                           | 3                                  | 119                      | 82            | 7           |
| Time               | 24                           | 41                                 | 29                       | 26            | 63          |
| Transport          | 23                           | 31                                 | 0                        | 1             | 19          |
| Weather            | 5                            | 3                                  | 24                       | 2             | 9           |
| Total              | 275                          | 410                                | 408                      | 300           | 298         |

#### D. Evaluation Setting

In the experiment, each POI is selected in turns to be used as training data, with the remaining POIs used as testing. The training and testing approach used in this experiment is similar to cross-domain learning, where a source domain (i.e. source POI) is used as training and target domain (i.e. target POIs) are used as testing. As for the baseline comparison, in our experiment, we compare opinion type detection using the proposed approach with its baselines. Two experiment settings conducted are listed as below.

1) Classifier Selection for Opinion Identification. This experiment is carried out to select a classifier that will be used in our experiment. The experiment compares three classifiers, i.e. SVM, NB, and DT in their classification accuracy in opinion type detection. In this experiment, features are extracted using NLP and BOW approaches.

2) Feature Expansion for Opinion Identification. This experiment is carried out to compare the proposed approach with its baselines, i.e. LO+NLP vs NLP and LO+BOW vs BOW. The experiment compares opinion type identification under two settings, i. using Source Target with low number of training, i.e. SOURCE<sub>LOW</sub> and ii. using Source Target with high number of training, i.e. SOURCE<sub>HIGH</sub>.

All evaluation will be performed based on the seven opinion types, and five POI of review sentences which are Genting Highlands Theme Park, Cameron Highlands Boh's Tea Centre, Club Med Cherating Beach, Escape Penang, and Penang Hill.

#### E. Experiment Results

1) Classifier selection for opinion identification: In this experiment, Genting POI is used as source training POI, while other POIs are used for testing. This experiment is carried out to select a classifier that will be used in the experiment for feature expansion evaluation.

#### Natural Language Processing (NLP)

By analyzing the results in Table X, the overall accuracy for each POI is higher when using SVM classifier. For

“Cameron”, the overall accuracy by using SVM is higher at 61.59%, the percentage is higher by 14.82% compared with NB classifier (46.77%), and 24.33% compared to DT classifier (37.26). The results also can be seen for “Penang Hill”, the overall accuracy by using SVM is higher at 51.54%, the percentage is higher by 8.46% compared with NB classifier (43.08%) and 16.92% compared to DT classifier (34.62). For “Escape”, the overall accuracy by using SVM is higher at 63.47%, the percentage is higher by 9.6% compared with NB classifier (53.87%), and 7.38% compared to DT classifier (56.09%). For “Cherating”, the overall accuracy by using SVM is higher at 52.69%, the percentage is higher by 14.73% compared with NB classifier (37.96%), and 1.12% compared to DT classifier (51.84%). From the results, it is observed that SVM is higher than NB and DT among all the four POI, hence, SVM will be chosen for further discussion for NLP approach.

#### Bag-of-Words (BOW)

By analyzing the results in Table XI, the overall accuracy for each POI is higher when using SVM and DT classifier. For “Cameron”, the overall accuracy by using SVM is 57.41%, the percentage is higher by 11.78% compared with NB classifier (45.63%), and 1.14% compared to DT classifier (56.27). The results also can be seen for “Escape”, the overall accuracy by using SVM is higher at 62.36%, the percentage is higher by 9.59% compared with NB classifier (52.77%) and 10.7% compared to DT classifier (51.66%).

For “Penang Hill”, the overall accuracy by using DT is higher at 60.38%, the percentage is higher by 4.23% compared with SVM classifier (56.15%), and 17.69% compared to NB classifier (42.69%). For “Cherating”, the overall accuracy by using DT is higher at 53.26%, the percentage is higher by 1.42% compared with SVM classifier (51.84%), and 17.57% compared to NB classifier (35.69%). Since only SVM perform better in NLP approach, hence for consistency purpose, only SVM classifier will be chosen for next evaluation.

TABLE X. OVERALL ACCURACY OF EACH POI BY EACH CLASSIFIER IN NLP APPROACH (NOUN FEATURE)

| Classifier \ POI | SVM   | NB    | DT    |
|------------------|-------|-------|-------|
| Cameron          | 61.59 | 46.77 | 37.26 |
| Penang Hill      | 51.54 | 43.08 | 34.62 |
| Escape           | 63.47 | 53.87 | 56.09 |
| Cherating        | 52.69 | 37.96 | 51.84 |

TABLE XI. OVERALL ACCURACY OF EACH POI BY EACH CLASSIFIER IN BOW APPROACH

| Classifier \ POI | SVM   | NB    | DT    |
|------------------|-------|-------|-------|
| Cameron          | 57.41 | 45.63 | 56.27 |
| Penang Hill      | 56.15 | 42.69 | 60.38 |
| Escape           | 62.36 | 52.77 | 51.66 |
| Cherating        | 51.84 | 35.69 | 53.26 |



2) *Feature expansion for opinion type identification*: This experiment is carried out to compare the proposed approach with its baselines, i.e. LO+NLP vs NLP and LO+BOW vs BOW under two settings, one is using Source Target with low number of training set, SOURCE<sub>LOW</sub> (Genting as training data, remaining POIs as testing data), second is using Source Target with high number of training set, SOURCE<sub>HIGH</sub> (Cameron as training data, remaining POIs as testing data).

**SOURCE<sub>LOW</sub> (Genting as training data, remaining POIs as testing data)**

Table XII illustrates the results for the overall accuracy for “Genting” as a training dataset. By studying the table, we can see that there is no difference in the overall accuracy as all the approach resulted in 0.57.

**SOURCE<sub>HIGH</sub> (Cameron as training data, remaining POIs as testing data)**

Table XIII shows the results for the overall accuracy for “Cameron” as a training dataset. BOW+LO. From the results, there is an accuracy of 0.60 by using LO+BOW which is 3% higher compared to BOW approach (0.57), with 6% higher compare to NLP (Noun) approach (0.54) and NLP+LO approach (0.54). In addition, NLP+LO approach does not perform well in NLP approach where the same results are presented for NLP and NLP+LO which is 0.54.

Table XIV shows that BOW+LO perform the best for SOURCE<sub>HIGH</sub> over SOURCE<sub>LOW</sub> in overall accuracy and over other approaches (BOW, NLP, NLP+LO). This proves that the number of training data has a significant impact on the classification accuracy. From the results, the highest accuracy of 0.60 is achieved by using BOW+LO. Therefore, we can conclude that LO can be potentially used with BOW (BOW+LO) in achieving a better overall accuracy.

TABLE XII. OVERALL ACCURACY FOR BOW VS NLP VS LO (GENTING AS TRAINING DATASET)

| Approach \ POI   | BOW  | NLP  | BOW+LO | NLP+LO |
|------------------|------|------|--------|--------|
| Cameron          | 0.57 | 0.62 | 0.57   | 0.61   |
| Penang Hill      | 0.56 | 0.52 | 0.55   | 0.52   |
| Genting          | 0.62 | 0.63 | 0.62   | 0.62   |
| Cherating        | 0.52 | 0.53 | 0.53   | 0.53   |
| Overall Accuracy | 0.57 | 0.57 | 0.57   | 0.57   |

TABLE XIII. OVERALL ACCURACY FOR BOW VS NLP VS LO (CAMERON AS TRAINING DATASET)

| Approach \ POI   | BOW  | NLP  | BOW+LO | NLP+LO |
|------------------|------|------|--------|--------|
| Escape           | 0.54 | 0.53 | 0.54   | 0.53   |
| Penang Hill      | 0.62 | 0.57 | 0.58   | 0.57   |
| Genting          | 0.62 | 0.59 | 0.68   | 0.59   |
| Cherating        | 0.50 | 0.48 | 0.59   | 0.48   |
| Overall Accuracy | 0.57 | 0.54 | 0.60   | 0.54   |

TABLE XIV. OVERALL ACCURACY FOR SOURCE<sub>LOW</sub> VS SOURCE<sub>HIGH</sub>

| Approach \ POI         | BOW  | NLP  | BOW+LO | NLP+LO |
|------------------------|------|------|--------|--------|
| SOURCE <sub>LOW</sub>  | 0.57 | 0.57 | 0.57   | 0.57   |
| SOURCE <sub>HIGH</sub> | 0.57 | 0.54 | 0.60   | 0.54   |

Note: SOURCE<sub>LOW</sub> (Genting as training source) SOURCE<sub>HIGH</sub> (Cameron as training source)

VI. CONCLUSION

This research aims to help the tourist to easily digest the vast availability of opinion by categorizing the reviews. Specifically, we improve of opinion type detection via keyword expansion to address the out-of-vocabulary (OOV) and limited labeled data issues.

From this study, we found that WordNet’s labels of semantic relations are useful for the research of feature expansion. This is validated from the experiment that shown that our proposed feature expansion approach is able to improve opinion type detection with reasonable accuracy. Better accuracy can be seen for BOW+LO (in Table XIII) as well as when SOURCE<sub>HIGH</sub>, i.e. when larger sentences, compared to the one with lesser sentences, SOURCE<sub>LOW</sub> are used as training data. This result suggests that the former could yield more keywords/features to be expanded and trained.

In summary, opinion type detection is important as it helps to automatically categorize customer review according to opinion type. This is convenient for customer and it could improve the way of how information can be selected to reach its users by filter the information they need. Designing a good opinion type detection framework is challenging as it involves solving problems at various stages ranging from training reviews collection, features selection, classification of reviews, and building model. In order to verify all these stages, the proposed feature expansion has been evaluated with real user reviews and data collection. A positive outcome in terms of performance accuracy is achieved from the evaluation and this motivates us to move forward to further investigation the potential of other semantic relationships in the adapted LO as future work.

REFERENCES

- [1] Z. Kavasoglu, “Personalized Summarization Of Customer Reviews Based On User’s Browsing History,” International Journal on Computer Science and Information Systems, vol 8(2), pp. 147–158, 2013.
- [2] A. Khan, B. Baharudin, L. H. Lee, & K. Khan, “A review of machine learning algorithms for text-documents classification,” Journal of advances in information technology, vol 1(1), pp. 4–20, Feb. 2010.
- [3] Q. Li, S. Li, S. Zhang, & J. Hu, “A Review of Text Corpus-Based Tourism Big Data Mining,” Applied Sciences, vol. 9, pp 1–27, Aug 2019.
- [4] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes and Donald Brown, “Text Classification Algorithms: A Survey,” In Information 2019, vol 10, pp. 1-68, 2019.
- [5] E. Rudkowsky, M. Haselmayer, M. Wastian, M. Jenny, S. Emrich, & M. Sedlmair, “More than Bags of Words: Sentiment Analysis with Word Embeddings,” Communication Methods and Measures, pp. 1931-2458, Apr. 2018.
- [6] Sneha M Nakade & Sachin N Deshmukh, “Finding Semantic Orientation of Reviews Using Unsupervised PMI Algorithm,” International Journal of Science and Research (IJSR), vol 5(2), pp. 2107–2110, Feb. 2016.

- [7] M. Hu and B. Liu, "Mining Opinion Features in Customer Reviews," Proceedings of the 19th National Conference on Artificial Intelligence, vol. 4, no. 4, pp. 755–760, No2004.
- [8] K. Vivekanandan, & J. S. Aravindan, "Aspect-based opinion mining: A survey," International Journal of Computer Applications, vol 106(3), pp. 21–26, Nov. 2014.
- [9] S. Abeysinghe, I. Manchanayake, C. Samarajeewa, P. Rathnayaka, M. J. Walpola, R. Nawaratne, T. Bandaragoda & D. Alahakoon, "Enhancing Decision Making Capacity in Tourism Domain Using Social Media Analytics," 2018 18th International Conference on Advances in ICT for Emerging Regions, ICTer 2018 – Proceedings, pp. 369–375, Sept. 2018
- [10] S. M. Rezaeinia, A. Ghodsi, & R. Rahmani, "Improving the accuracy of pre-trained word embeddings for sentiment analysis," arXiv 2017, arXiv:1711.08609.
- [11] Muhammad Afzaal, Muhammad Usman, and Alvis Fong, "Tourism Mobile App With Aspect-Based Sentiment Classification Framework for Tourist Reviews," IEEE Transactions on Consumer Electronics, vol. 65, pp 233–241, 2019.
- [12] K. Soo-Min and H. Eduard, "Automatic Detection of Opinion Bearing Words and Sentences," Information Sciences Institute University of Southern California, vol 10, pp. 60–66, Apr. 2005.
- [13] L. Wu, S. C. H. Hoi and N. Yu, "Semantics-Preserving Bag-of-Words Models and Applications," in IEEE Transactions on Image Processing, vol. 19, no. 7, pp. 1908-1920, July 2010.
- [14] M.F. Porter, "An algorithm for suffix stripping," Program, vol 14(3), pp. 130–137, July 1980.
- [15] Bruce M. Hill, "Posterior distribution of percentiles: Bayes' theorem for sampling from a population," Journal of the American Statistical Association, vol 63, pp. 677–691, Nov. 2014.
- [16] Z. Qu, X. Song, S. Zheng, X. Wang, X. Song, & Z. Li, "Improved Bayes Method Based on TF-IDF Feature and Grade Factor Feature for Chinese Information Classification," Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, China. pp. 677–680, 2018.
- [17] L. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, "Deep Learning," MIT Press: Cambridge, MA, USA, vol 1, 2016.
- [18] W. Wang, Y. Huang, Y. Wang, and L. Wang, "Generalized autoencoder: A neural network framework for dimensionality reduction," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, pp. 490-497, 2014.
- [19] G. Bo, and H. Xianwu, "SVM Multi-Class Classification," Journal of Data Acquisition & Processing, vol 3, pp. 47–52, 2006.
- [20] M. Mohri, A. Rostamizadeh, and A. Talwalkar, "Foundations of Machine Learning," MIT Press: Cambridge, MA, USA, 2012.
- [21] R. Farzi, and V. Bolandi, "Estimation of organic facies using ensemble methods in comparison with conventional intelligent approaches: A case study of the South Pars Gas Field, Persian Gulf, Iran," In Modeling Earth Systems and Environment, vol 2(105), pp. 1-13, 2016.
- [22] A. Genkin, D. D. Lewis, & D. Madigan, "Large-scale Bayesian logistic regression for text categorization," Technometrics, vol 49(3), pp. 291–304, Aug 2007.