A Combined Weighting for the Feature-Based Method on Topological Parameters in Semantic Taxonomy Using Social Media

Ali Muttaleb Hasan, Noorhzaimi Mohd Noor, Taha Hussein Rassem, Ahmed Muttaleb Hasan and Waleed A.Hammood
Faculty of Computing, Universiti Malaysia Pahang, 26300 Gambang, Kuantan, Pahang Darul Makmur, Malaysia

E-mail: nhuzaimi@ump.edu.my

Abstract. The textual analysis has become most important task due to the rapid increase of the number of texts that have been continuously generated in several forms such as posts and chats in social media, emails, articles, and news. The management of these texts requires efficient and effective methods, which can handle the linguistic issues that come from the complexity of natural languages. In recent years, the exploitation of semantic features from the lexical sources has been widely investigated by researchers to deal with the issues of “synonymy and ambiguity” in the tasks involved in the Social Media like document clustering. The main challenges of exploiting the lexical knowledge sources such as WordNet 3.1 in these tasks are how to integrate the various types of semantic relations for capturing additional semantic evidence, and how to settle the high dimensionality of current semantic representing approaches. In this paper, the proposed weighting of features for a new semantic feature-based method as which combined four things as which is “Synonymy, Hypernym, non-taxonomy, and Glosses”. Therefore, this research proposes a new knowledge-based semantic representation approach for text mining, which can handle the linguistic issues as well as the high dimensionality issue. Thus, the proposed approach consists of two main components: a feature-based method for incorporating the relations in the lexical sources, and a topic-based reduction method to overcome the high dimensionality issue. The proposed method approach will evaluated using WordNet 3.1 in the text clustering and text classification.

Keywords: component, text mining, text classification, sentiment analysis, semantic representation, weighting, topological parameter, Non-taxonomy.

1. Introduction

The proportion of the query and user-produced content in many websites, social media, and online services, such as Twitter, Facebook, YouTube and LinkedIn increases the quality of opinion and information obtainable from the internet. Therefore, the mining of data currently begun utilized to handle huge amounts of data all in name of defending the native land. Different kinds of information retrieval, including healthcare records, network traffic, and financial transactions being subject to mining with a view to creating profiles, constructing models of social network, and detecting communications by terrorists. This activity generates serious concerns on privacy and has creation of particularity certification data mining techniques. In text mining, the documents have clusters, for data mining is about looking for patterns in data. Furthermore, text mining is looking also for modality in the text, it’s the processing of analyzing the text to extraction the information that is useful for nominated purposes.

1 https://wordnet.princeton.edu
compared with the kind of data we have been taking about. In this paper, the motivation of this aim paper for trying to extract the vector of text mining [1]. In addition, the NLP is the aim of the techniques in AI is to give an income to computers the capacity to realize natural language processing NLP. Thus, the major problem in this is understanding how to represent natural language texts semantically in an efficient and effective manner [2]. Social networking that is Web-based (otherwise known as long-range informal communication) happens to be an online stage used by individuals for purpose of building interpersonal organizations or social relations with other individuals who are comparable and interesting, exercises, institute or veritable associations. There is some regular components among different web-based social networking accessible on the web. They are Web 2.0 based web applications, embolden the advancement of online informal platforms by associating the profile of a client with those of different people, and additional bunches. Therefore, to take the information extraction to the next level, the extracted information can be in a subsequent step to understand how to make rules for extracting the text information that is characteristic of the content of itself. These rules might predict the values of definite slot filters from the rest of the text. In certain tightly constrained situations, like internet job posting for computing-related jobs, or post from friend feed by politics to relate with the post, these all treasure based on the database as what applies to extraction of the information.

We have proposed in this paper, the weighting of feature method to combine four things as which “2 Synonymy, 3 Hypernym, 4 Non-taxonomy, and 5 Glosses”; this research proposed a new sematic representation approach of texting mining to reduce the high dimensionality relations in social Media. To capturing more relevance features between two words.

The remaining part of the paper arranged as follows: Section 2, semantic representation. While in section 3, Topological of parameter discussed. In section 4, Definition of terms. In section 5, structural-based methods discussed. In section 6, Semantic Representation Approach discussed. While section 7, result and dissection. In section 8, Conclusion.

2. Semantic representation

In cognitive psychology, semantic representation defined as a topic, which studies the techniques of using the mental lexicon to interpret the meaning of words in natural languages. In computational linguistics and artificial intelligence, semantic representation deals with how word-meaning can extracted from the lexical source database or corpus. Therefore, the most extensive model which implements distinct semantic representation is WordNet as a lexical source [1],[2] it uses a semantic representation to compute the similarity of two words, a network model for representing huge number with nouns, and verbs in English. In WordNet 3.1, “Both nouns and verbs have their own semantic relations as well as their own organization based on the role they play in constructing linguistic messages. These relations and organization that created depending on the relations, which relevant within a specific class of words. For nouns, relations, such as hierarchical relations, synonymy, and part-whole relations usually play the most significant roles. In a number of studies, they introduce the semantic representation for various methods [2]. For verbs, instead, dominant is the taxonomy (hierarchical relations related to specificity in manner), Synonym, is-a relation, holonym, meronym, and antonyms. Some evidence is compatible with a different role of relations such as hyponym and antonym for nouns and verbs.

2.1 Feature-based on semantic taxonomy

The ‘Is-A’ relations are a semantic representation formulated in every concepts/word construct on the semantic features in lexical resources like WordNet [3, 4]. To use by representing the knowledge in the lexical source as the semantic taxonomy. The semantic taxonomy is a network among concepts in the

2 https://github.com/aliMuttaleb/Ali-Muttaleb/blob/master/Synonym.txt
3 https://github.com/aliMuttaleb/Ali-Muttaleb/blob/master/Hypernym
4 https://github.com/aliMuttaleb/Ali-Muttaleb/blob/master/Non-taxonomy
5 https://github.com/aliMuttaleb/Ali-Muttaleb/blob/master/Glosses
lexicon, in which nodes of representing concepts and each edge has represented as a hypernyms and hyponyms relations. These relations are a relation that organizes nominal and verbal synsets as the inheritance semantic taxonomy. Figure 1 shows the part of feature-based of semantic taxonomy of the WordNet 3.1.

![Feature semantic-based taxonomy in WordNet 3.1](image)

**Figure 1.** Feature semantic-based taxonomy in WordNet 3.1

In the semantic taxonomy, there are several terms, which will used during this research:

- The root of the taxonomy it is a node of the entity, which superordinate all nodes in the taxonomy. In figure 1, root an entity.
- Hyponymy parent concept, it is a function, which returning the direct hypernyms “parents” in the concept, for example, hyponymy (Chap, Canine) = Organism.
- Subsumes (C) (or ancestors (C)) that will returning the hyponyms in concept C in a recursively way, for example, subsumes (Chap, Canine) = (Organism, living thing, Object, Physical Entity, entity).
- Hyponym or child concept, as which returned the direct hyponym (Children) in the concepts, for example, hyponym (Person) = (Unwelcome person, Male, Communicator, Female, Adult).
- Descendant (C) returning the hyponyms of a concept c in a recursively way, for example, (Person) = (Unwelcome person, Male, Communicator, Female, Adult, Unpleasant person, Chap, Man, Gossip, Woman, Unpleasant woman, 3 dog, 2 guy, cat, hombre, 3 cats, 2frume, dog).
- Least common subsumes LCS (C1, C2): it is a function, which takes as input two concepts (C1 and C2) and returns the least common subsumes of C1, C2, for example, LCS (Chap, Canine) = Organism.
- Path (C1, C2): It is a set containing the nodes (sometimes edges) from the concept C1 to the concept C2.
- Length (C1, C2): It is the number of nodes in the shortest path from the concept C1 to the concept C2.
- Depth (C): It is the length shortest path from the root to the concept C.

### 2.2 Pre-processing

Text pre-processing is fundamental in many NLP applications and tasks, as which depend on the unstructured knowledge sources for extracting information or manipulating the textual parts. The main aim of this step is to prepare the unstructured knowledge sources such as a collection of documents. The document in the corpus may include some non-informative and noisy terms that lead to the wrong process in natural language processing applications. For this reason, the corpus requires the pre-processing step to handle the noisy terms [5]. Since the distributional-based semantic representation approach depends on the collection of documents to extract the semantics from the text, it is a required pre-processing step. The pre-processing step is implemented to achieve three functions: tokenization, removing stop-words and stemming [6-8]. For each document, the tokenization applied to segment the text in document text into the basic linguistic units called tokens. The second function is removing stop-words, which aimed to exclude the stop-word from the document before representing the semantics [9]. The stop-words such as the, of, and for may lead to the poor semantic representation of the words [10]. For example, the text “a machine for performing calculations automatically” contains two stop-words a and for which are non-descriptive terms in text and have to be excluded. The final function of the pre-processing step is the stemming, which is responsible for converting the different forms of the words to their single source or a lexeme. This function is to avoid repeating the representation of the same lexeme.
For instance, the words books and book have to represent one time only. There are several stemmers for English texts such as the Java implementation of Porter’s stemmer.

1. Building Co-occurrence for association model - In distributional of Co-occurrence model methods to build semantic representation matrix of the words. Generally, the vector space model used for the semantic representation of the concepts/words using the information in the corpus. This model is to construct the word-context matrix that contains the words as the rows and the contextual relationship in which that words occurrences as the columns. Formally, for a given corpus C, let w= wn is a group of words, which appear in whole corpus, and T= {T1, T2,…, Tm} is that set of contexts, the semantic representation of these words is represented as the co-occurrences matrix as shown in Table 1.

| Table 1. Weighting for taxonomy |
|--------------------------------|
| W1 | W2 | … | … | … | … | Tm |
| T1 | f(1,1) | f(1,2) | … | … | … | f(1,m) |
| T2 | f(2,1) | f(2,2) | … | … | … | f(2,m) |
| … | … | … | … | … | … | … |
| Wn | f(n,1) | f(n,2) | … | … | … | f(n,m) |

Each vector in this matrix reveals the representation of one word, in which the elements of this vector are the weights that correspond to each context. In this regard, the context can defined as the environment of words that uses to extract the co-occurrences of words. Different distributional methods used different contexts to represent the corpus. The context of a word can be one of the following types: direct surrounding words, -nword with +nword, sentence, paragraph or document. In all types except document, the elements t1, t2,…, tm are single words or multiword expressions. For example, the word animal can represented in the sentence “The dog and cat are a type of animals which can live with people” using the context sentence by a dog, cat, live and people, in which other words are stop-words.

2. Weighting - In this section, we weighting each element in the semantic representation of a word according to relevance, in which the element with a high weighting semantically associated with the concept of the word. The f (i, j) matrix is basic weighting (frequency of words occurs the neighbour with the context) of each context j to represent the semantics of the word i. The weight can be measured by a different formula such as TF-IDF [11], and mutual information.

3. Topological of parameter

In this phase, to proposing the novel knowledge approach method for representing semantics in textual documents in text mining tasks. It includes the main contribution of the current research which will prove that the topological parameters (such as, depth and density) are useful as an upstream dimensional reduction algorithm to the traditional representation of the textual documents in text mining tasks. This sub-phase tries to overcome the limitations in Social media, as which Facebook, LinkedIn, and tweeter this work, when people post something in politic in this case, will show many of feature in the concepts have like the same relevant ambiguity, (such as, synonymy, ambiguity, and high dimensionality) of the text mining tasks. The main aim of the present methodology is to investigate exploitation of knowledge in lexical sources for text mining tasks, the lexicon contain have a several semantic features such as: Taxonomical relation, meronym relation, synonyms, IS-A relations, textual definitions, glosses definitions, etc. The main challenge is how to select the informative features that improve semantic representation of concepts. Thus, the knowledge-based semantic representation of documents by replacing the words in the document usually leads to very huge number of features. This due to the number of concepts in the lexical sources is extremely large the number of concepts in smaller lexicon (WordNet 3.1) is around 117,659. Thus, the feature-based method is used to represent knowledge in a given lexical source as a taxonomy using “is-a” relations. Then, each concept is semantically representing by a set of feature recursively extracted from a taxonomy. In addition, the feature-based method is how to incorporate the features to capturing more semantic evidence, in the lexical source, by using topological parameter to reduce these features from the weighting of words such as depth, density, descendants and ancestors. All features of concepts have the same relevance based on feature semantic
method that most promising topological parameters that could be exploiting by constant weighting assumption. Figure 2 show the processing of semantic representation based on the feature-based method.

**Figure 2.** The methodology of illustration of the research issues

### 3.1 Depth of synset

The weighting of depth is a mechanism used to reduce the features from the concepts in a semantic taxonomy the weighting of the taxonomy. The depth of synset and taxonomy to the concepts indicate, by depth (synset) is commonly the numbering of nodes straight the protracted path between synset and root of the entity. The depth is considerable in taxonomy-based on semantic representations. However, the words in the synset will reduce by using the depth to find the nearest words in the concepts to give specific words/concepts. The depth of concepts in WordNet 3.1 ranged between 1 and 20. The weight concepts or synset will determine by taxonomy like how many of descendant’s node and how many of ancestor’s node awhile arrange the depth of all words. In Fig 3 shows how the depth accounts the words to find the relevance features by hyponym and hypernym.

**Figure 3.** Code of features by hyponym and hypernym

### 3.2 Density of synset

In a semantic taxonomy, the domestic density of a synset node indicate to the numbering of connexion nodes or the number of its child nodes. Based on the assumption that two concepts are semantically close if they densely connected locally. However, density used as a topological parameter to locate the specific of a concept. In the WordNet 3.1 the description of the semantic relations between words in synset, one such as a relationship in is-a relationship, at which connect as a Hyponym and Hypernym in synset to give the more specific concept of the words.
4. Definition of terms

The terms of deflation, it is worth defining some of the terms, which related to the topic of the current research. The following sub-sections provide the definitions of the terms relating to computational semantic analysis.

4.1 Semantic Relation

Semantic relation is a link between two words/concepts reveals the relevance of their semantics. Semantic relation has been exploiting in natural language processing applications such as WSD (Word Sense Disambiguation) [12] Query expansion and suggestion. There are many types of semantic relations such as:

- Synonyms: it is equivalence relations. Two words are synonyms if their meaning is similar such as car and automobile.

- Is-A relation: includes hypernyms and hyponyms. It has called also a hierarchical relation. For example, fruit is a hypernym of apple; apple is a hyponym of fruit.

- Holonym: the concept C1 is a holonym of the concept C2 if C2 is a part of C1. For example, the computer is holonym of RAM.

- Meronym: the concept C1 is a meronym of the concept C2 if c1 is a part of C2. For example, the finger is meronym of hand. The Cheese, beef, tomato, and bread are meronyms of the burger.

To clarity, more based on our episode of our implementation, as which show the synonyms of words.

```java
public class Synonyms {
    private final FileProcess fp = new FileProcess();
    private String pth = "data/en/synonyms";
    private Map<Integer, Set<Integer>> sid = new HashMap<>();
    private Map<String, Set<Integer>> WordToSynset = new HashMap<>();
    private void BuildWordTosynset(String s) {
        String trimmedWord = s.toLowerCase().trim();
        Set<Integer> sx = WordToSynset.get(trimmedWord);
        if (sx == null) {
            sx = new HashSet<>();
            sid.put(trimmedWord, sx);
            sx.add(1);
            WordToSynset.put(trimmedWord, sx);
        }
    }
}
```

4.2 Semantic “Is-a” relation

The semantic of “is-a” relation in the reviewing previous of literature studies of the current research, the recent works about using the lexical sources for representing the textual documents in text mining task focus on three topics. The first topic focus on how the semantic information can be organized to provide the efficient semantic measures by introducing new approach for representation [13, 14], proposing new measure [15, 16] and combining the semantic relations [17] to ameliorate the performance of measuring in semantic relatedness SR. The second topic focuses on using the existing approaches and measures to examine them by using the comparative evaluation among approaches or measures or between lexical sources. The third topic focuses on handling the semantic knowledge acquisition issues to enhance the quality and quantity of the lexical resources by integrating two resources. In this paper, the feature-based method based on representing the semantic of concepts as a set of mixed attributes from the semantic relations and glosses in the knowledge sources. The semantic relations include the “is-a” hypernym-hypernym relations, holonym, inverse gloss, and meronym. Thus, the semantic of the car represented as the set of features containing the following terms: vehicle, convertible, accelerator, train, and cable car. The performance of the feature-based method depends on the factors, which related to the fidelity, continuity, and balance of the knowledge sources such as WordNet 3.1. Although there are a number of knowledge sources that have created to contain the semantic features of the concepts in the language, they include only the classical semantic features.
5. Structural-based methods

The structural method refers to the “semantic similarity” SS measures, which are depend on the features of the semantic taxonomy (ST), of the knowledge sources to weighting the features. There are several features in the semantic taxonomy that can used for measuring semantic similarity between two concepts, such as, depth, ancestors and path. In the previous study, many ways have used to classify these types of semantic measures. In this work, these measurements are clustering depending using weighting features to measures the semantic similarity (SS) into two main types: direct and indirect.

5.1 Direct Measures

The direct measures refer to the semantic measures, which directly exploited the structural weighting of features. The structural features are length; depth of concepts, depth of semantic taxonomy, and least common subsumer (LCS). Table 2 shows a summary of direct semantic measures.

Table 2. The summary of direct semantic measures

| Author   | Semantic measures | Features                              |
|----------|-------------------|---------------------------------------|
| [18]     | \( Sim_{Rada}(c_1, c_2) = \frac{1}{\delta(c_1, c_2) + 1} \) | Length                                |
| [19]     | \( Sim_{LCH}(c_1, c_2) = -\log \left( \frac{\delta(c_1, c_2) + 1}{2 \times D} \right) \) | Length, depth of semantic taxonomy    |
| [20]     | \( Sim_{Wup}(c_1, c_2) = \frac{2 \times d(LCS(c_1, c_2))}{\delta(c_1, c_2) + 2 \times d(LCS(c_1, c_2))} \) | LCS, length, depth                    |
| [21]     | \( Sim_{Li}(c_1, c_2) = e^{-\alpha \delta(c_1, c_2)} \times \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} ; h = d(LCS(c_1, c_2)) \) | LCS, length, depth                    |
| [22]     | \( Sim_{Liu}(c_1, c_2) = \frac{\beta \times d(LCS(c_1, c_2)) + \alpha \times \delta(c_1, c_2)}{\beta \times d(LCS(c_1, c_2))} \) | LCS, length, depth                    |
| [23]     | \( Dist_{Mubaid}(c_1, c_2) = \log(\delta(c_1, c_2) - 1)^{\alpha CSpec(c_1, c_2)^{\beta} + k} \) | LCS, length, depth, depth of semantic taxonomy |
| [24]     | \( Sim(c_1, c_2) = -\log \left( \frac{|T(c_1) \cup T(c_2)| - |T(c_1) \cap T(c_2)|}{|T(c_1) \cup T(c_2)|} \right) \) | Subsumes                              |

5.2 Indirect Measures

The indirect measures refer to the semantic measures, which exploit the structural features to quantify the information content (IC) of a given concept and then use the IC to measure the semantic similarity between two concepts. These measures have classified in the previous researches as the IC-based measures. The IC based on previous researchers shown in Table 3.

Table 3. Information content

| Author   | IC metrics                                      | Features                                                                 |
|----------|------------------------------------------------|---------------------------------------------------------------------------|
| [25]     | \( IC_{Seco}(c) = 1 - \frac{\log(desc(c) + 1)}{\log(Max_{nodes})} \) | Descendants, between two nodes in the semantic relation                   |
| [26]     | \( IC_{Zhou}(c) = k \left( 1 - \frac{\log(desc(c) + 1)}{\log(Max_{nodes})} + 1 \right) - k \left( \frac{\log(depth(c))}{\log(Max_{depth})} \right) \) | Descendants, number of nodes in the semantic taxonomy, depth of concept, and depth of taxonomy |
| [27]     | \( IC_{Sanchez}(c) = -\log \left( \frac{Leaves(c)}{\frac{subsumers(c)}{Max_{Leaves}} + 1} \right) \) | Leaves of concept subsumes, and number of leaves in the semantic taxonomy |
In the semantic representation approach, we proposed a weighted function for a system based on supervised hashing method. We proposed weighting on a feature extremel performance with agreeable statistical to challenge problems. Based on the observation the features that will select the depth and destiny depends on the taxonomy of weighting the features to output a good have been developed to learn the dataset depending on hash fuction in recent year, the hash mapping will select the depth and destiny depends on the taxonomy of weighting the features to output a good performance with agreeable statistical to challenge problems. Based on the observation the features that extremely related to the KNN classification accuracy, this paper a novel the feature- based on KNN-based on supervised hashing method. We proposed weighting on a feature-based method to the assumption the semantic representation to method the proximity relationships inveterate in training data.

| Equation | Description |
|----------|-------------|
| $IC_{SanBatet}(c) = -\log \left( \frac{\text{commonness}(c)}{\text{commonness}(\text{root})} \right)$ | Leaves of concept subsumes, and number of leaves in the semantic taxonomy |
| $IC_{Taiibe}(c) = \left( \sum_{c' \in E} \text{Score}(c') \right) \times \text{AvgDepth}(c)$ | Descendants, depth of concept, hypernyms, ancestors |
| $\text{Score}(c) = \left( \sum_{a \in \text{hyper}(c)} \frac{\text{depth}(a)}{\text{Desc}(a)} \right) \times \text{Hypo}(c)$ | hyponyms of the concept c, depth of concept, number of nodes, depth of semantic taxonomy |
| $IC_{Aouicha}(c) = \sum_{c' \in \text{Subgraph}(c)} \text{Score}(c')$ | depth, ancestors, hypernyms, descendants, and hyponyms |
| $\text{Score}(c) = \left( \sum_{a \in \text{hypon}(c)} -\log \left( \frac{1}{\text{depth}(c')} \times \text{Term}(c') \right) \right) \times \text{Term}(c)$ | |
| $\text{Term}(c) = 1 - \frac{\log(\text{HypoInfo}(c) + 1)}{\log(\text{maxHypoInfo})}$ | |
| $\text{HypoInfo}(c) = \sum_{c' \in \text{descendants}(c)} \frac{1}{\text{depth}(c')}$ | |

**6. Semantic representation approach**

After compiling the literature survey on semantic representation approaches, semantic measures, and knowledge-based text mining tasks, to develop and improve the semantic representation approach for text mining tasks/applications. In the semantic representation approach, we proposed a weighted semantic representation approach to handle the constant weighting assumption in feature-based methods using topological parameters. The researchers showed the numbers of depending on the techniques for Words/concepts on the classification. During our research, we have tried to consider the most recently confirmed efficient approaches that we have found the result of our investigation. In this section, we have tried to presenting a study on some research papers to give a view of three mechanization to grade the feature of concepts.

**6.1 Text Classification Application**

In text classification, we proposed a new weighting to extract the important features from the concepts by the taxonomy of semantic representation. Instead of using texting or word relation, we proposed using weighting to extract the features from concepts in an incorporate the features for capturing more semantic evidence all features of concepts have the same relevance constant weighting assumption The concept of semantic representation classify features by the feature-based method. A system based on the PM has been implemented and tested. The experimental results show that the proposed method works as an effective text classify.

1) **KNN hash mapping** - In this section, the hash mapping in KNN is very efficient in many tasks to reduction high dimensionality and pressuring the database. In spite of the fact that a lot of approaches have been developed to learn the dataset depending on hash functions in recent year, the hash mapping will select the depth and destiny depends on the taxonomy of weighting the features to output a good performance with agreeable statistical to challenge problems. Based on the observation the features that extremely related to the KNN classification accuracy, this paper a novel the feature- based on KNN-based on supervised hashing method. We proposed weighting on a feature-based method to the assumption the semantic representation to method the proximity relationships inveterate in training data.
2) Naïve Bayes - In this section, there are a few approaches that based on machine learning has presented. We will consider the method to be machine learning based on features selection. The interesting in the machine learning communication on the semantic representation based on semantic relation problem. Starting from simple statistical Naive Bayes-based methods, we describe methods using semantic representation fields.

6.2 Sentiment analysis
The sentiment analysis over the Social Media offers organizations a quick and efficient way to monitor the publics’ feelings towards their brand, business, directors, etc. A wide range of features and methods for training sentiment classify for Social Media datasets have researched in recent years with varying results. In this paper, we introduce a new weighting feature-based approach to sentiment the analysis words in Social Media based on the measure of semantic taxonomy to reduce the noisy and irrelevant features. For each extracted features from the concept, we add its semantic concept (e.g. “politic”) as an additional feature and measure the correlation of the representative concept with negative/positive sentiment.

6.3 Clustering Unsupervised
In the clustering of unsupervised feature selection using a combined weighting of a feature-based method to measure the similarity as measures described. The structural-based methods refer to the semantic, which based on the features of semantic taxonomy of the knowledge source. Although the units in the semantic representation that taxonomically extracted, the topological metrics (e.g. depth and density) ignored when computing the semantic similarity.

1. K-means - This method of vectors, originally from features based method, which is prevalent for clustering the features after made from semantic taxonomy. K-means clustering aims to root M documents perception into k clusters. The algorithm has a loose relationship to the k-nearest neighbour classifier, a prevalent machine learning technique for classification that is used with k-means. Applying the 1-nearest neighbour classifier the cluster centres obtained by k-means classifies new data into the existing clusters.

2. Hierarchical clustering - In data mining, the statistical of hierarchical clustering features method to combine the relations between two words on concepts, also called hierarchical cluster or hierarchical clustering analysis.

7. Result & discussion
In the semantic of the relation is described the exploited the computing the relatedness degree the taxonomy of the WordNet 3.1 for the comparison, the WordNet (Christiane Fellbaum, 1998), is a lexical of the database as a development from the University of the Princeton in 1985. Is the most of resource used to computing the semantic relations of the task, it’s free and the document resource that is structure the semantics of the lexical contents of the English language. The WordNet 3.1 has organized the set of the synonyms called the synset, each of synset corresponds to the concepts are noun and verbs are organized in “is-a” of the taxonomy. The version is 3.1, the WordNet (Is-A) taxonomy assigned the nouns contains are 82,192 synset. In our comparison show the result between the proposed method and the previous methods using three of method as which taxonomy relation, non-taxonomy relation and Glosses. In the significance of the measuring as using the ρ of the null of the hypothesis when it become true. In the most of the empirical analysis, is a result of the yielding the r value and the ρ are considered the statistical significance, in these results the considered statistically significant when the μ is a value of these result, to consider the highly statistically significant. In measuring the semantic similarity, there are two cases of the statistical of the significance are used to compare the similarity of values. The statistical of the significance of the feature-based method is different between the proposed method and human judgment in the three of the measures are (taxonomy relations, Non-taxonomy relations and Glosses), each one has the hypothesis is that the scores of the similarity. Table IV show the statistical of
the significance feature-based method based on the semantic representation of the proposed method in the taxonomical of the relations.

| Features          | Rada | Wu | Li | LC | Sebti | Sánchez | Meng | Taieb | Aouicha | Our result |
|-------------------|------|----|----|----|-------|---------|------|-------|---------|------------|
| Taxonomy Relations | r    | 0.66 | 0.75 | 0.81 | 0.75 | 0.82 | 0.82 | 0.86 | 0.76 | 0.82 | 0.82 |
|                   | ρ    | 0.75 | 0.76 | 0.74 | 0.75 | 0.78 | 0.74 | 0.77 | 0.76 | 0.76 | 0.80 |
|                   | μ    | 0.70 | 0.75 | 0.77 | 0.75 | 0.80 | 0.78 | 0.81 | 0.76 | 0.79 | 0.81 |
| Non-Taxonomy Relations | Our PM | 0.82 |      |     |      |        |      |       |        |            |
| Glosses           | Non-zero | 96% |      |     |      |        |      |       |        |            |

8. Conclusions

In this paper, we proposed a new weighting of features based method, in this paper has described the weighting of features selection that is followed to locate the research objective and to coincide the scope of the feature based on the proposed method. Next, the details about the research phases are given, which consist of four phases starting with compiling of the literature review and defining the research problem followed by the collection of the text mining in social media and preparing it for the next phase. Based on these findings, using WordNet 3.1 as background knowledge for text clustering: First, experimenters should consider the nature of the dataset in hand and the diversity of topics before decide to use WordNet 3.1 for measuring similarity of relations. Second, the WordNet 3.1 structure does not seem to support the application of similarity measures. However, WordNet 3.1 can be better exploited by capturing specific types of relations such as “Synonymy, Hypernym, non-taxonomy, and Glosses”. The statistical of the significance of feature-based method shows the proposed method 0.82 and non-zero 0.96%, is better from the other human judgment based on our methods as which (Taxonomy relations, Non-taxonomy relations and Glosses). To concerns, this feature-based method, which aims to achieve the research objectives and finally the evaluation measurements are highlighted. Each sub-phase of the design and development phase is constructed to achieve one objective of the research, which in turn tries to overcome some challenges and problems that are observed in the related works.

Acknowledgment

This work is supported by the Universiti Malaysia Pahang (UMP) in Research Grant UMP RDU1803141, PGRS190398.

References

[1] K. M. Hermann, D. Das, J. Weston, and K. Ganchev, "Semantic frame identification with distributed word representations," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, vol. 1, pp. 1448-1458.
[2] A. Saif, M. J. Ab Aziz, and N. Omar, "Reducing explicit semantic representation vectors using Latent Dirichlet Allocation," Knowledge-Based Systems, vol. 100, pp. 145-159, 2016.
[3] M. B. Aouicha, M. A. Hadj Taieb, and A. B. Hamadou, "Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness," Applied Intelligence, journal article pp. 1-37, 2016.
[4] M. B. Aouicha, M. A. H. Taieb, and M. Ezzeddine, "Derivation of “is a” taxonomy from Wikipedia Category Graph," Engineering Applications of Artificial Intelligence, vol. 50, pp. 265-286, 2016.
[5] T. H. Rassem, N. M. Makbol, A. M. Hasan, S. S. M. Zaki, and P. Girija, "Restoring the missing features of the corrupted speech using linear interpolation methods," in AIP Conference Proceedings, 2017, vol. 1891, no. 1, p. 020119: AIP Publishing.
[6] A. M. Hasan, T. H. Rassem, and M. Karimah, "Pattern-Matching Based for Arabic Question Answering: A Challenge Perspective," Advanced Science Letters, vol. 24, no. 10, pp. 7655-7661, 2018.
[7] A. M. Hasan and L. Q. Zakaria, "QUESTION CLASSIFICATION USING SUPPORT VECTOR MACHINE AND PATTERN MATCHING." Journal of Theoretical & Applied Information Technology, vol. 87, no. 2, 2016.

[8] Q. Al-Tashi and A. M. Hasan, "WORD SENSE DISAMBIGUATION: A REVIEW," Southern Connecticut State University, Hilton C. Buley Library, vol. 1, 2, pp. 20-458, 2019, google scholar.

[9] S. Fodeh, B. Punch, and P.-N. Tan, "On ontology-driven document clustering using core semantic features," Knowledge and information systems, vol. 28, no. 2, pp. 395-421, 2011.

[10] A. M. Hasan, T. H. Rassem, and M. Noorhuzaaim, "Combined Support Vector Machine and Pattern Matching for Arabic Islamic Hadith Question Classification System," in International Conference of Reliable Information and Communication Technology, 2018, pp. 278-290: Springer.

[11] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," Expert Systems with Applications, vol. 66, pp. 245-260, 2016.

[12] X. Chen, Z. Liu, and M. Sun, "A unified model for word sense representation and disambiguation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1025-1035.

[13] S. Hassan, Measuring semantic relatedness using salient encyclopedic concepts. University of North Texas, 2011.

[14] M. A. H. Taieb, M. B. Aouicha, and A. B. Hamadou, "Computing semantic relatedness using Wikipedia features," Knowledge-Based Systems, vol. 50, pp. 260-278, 2013.

[15] L. Meng, J. Gu, and Z. Zhou, "A new model of information content based on concept’s topology for measuring semantic similarity in WordNet," International Journal of Grid and Distributed Computing, vol. 5, no. 3, pp. 81-94, 2012.

[16] D. Sánchez, M. Batet, and D. Isern, "Ontology-based information content computation," Knowledge-Based Systems, vol. 24, no. 2, pp. 297-303, 2011.

[17] G. Pirró, "A semantic similarity metric combining features and intrinsic information content," Data & Knowledge Engineering, vol. 68, no. 11, pp. 1289-1308, 2009.

[18] R. Rada, H. Milí, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," IEEE transactions on systems, man, and cybernetics, vol. 19, no. 1, pp. 17-30, 1989.

[19] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification," WordNet: An electronic lexical database, vol. 49, no. 2, pp. 265-283, 1998.

[20] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in Proceedings of the 32nd annual meeting on Association for Computational Linguistics, 1994, pp. 133-138: Association for Computational Linguistics.

[21] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," IEEE Transactions on knowledge and data engineering, vol. 15, no. 4, pp. 871-882, 2003.

[22] H. Liu, H. Bao, and D. Xu, "Concept vector for semantic similarity and relatedness based on wordnet structure," Journal of Systems and software, vol. 85, no. 2, pp. 370-381, 2012.

[23] H. Al-Mubaid and H. A. Nguyen, "A cluster-based approach for semantic similarity in the biomedical domain," in Engineering in Medicine and Biology Society, 2006. EMBS’06. 28th Annual International Conference of the IEEE, 2006, pp. 2713-2717: IEEE.

[24] M. Batet, S. Harispe, S. Ranwez, D. Sánchez, and V. Ranwez, "An information theoretic approach to improve semantic similarity assessments across multiple ontologies," Information Sciences, vol. 283, pp. 197-210, 2014.

[25] N. Seco, T. Veale, and J. Hayes, "An intrinsic information content metric for semantic similarity in WordNet," in ECAI, 2004, vol. 16, p. 1089.

[26] Z. Zhou, Y. Wang, and J. Gu, "A new model of information content for semantic similarity in WordNet," in Future Generation Communication and Networking Symposium, 2008. FGCS’08. Second International Conference on, 2008, vol. 3, pp. 85-89: IEEE.

[27] D. Sánchez and M. Batet, "A semantic similarity method based on information content exploiting multiple ontologies," Expert Systems with Applications, vol. 40, no. 4, pp. 1393-1399, 2013.

[28] M. A. H. Taieb, M. B. Aouicha, and A. B. Hamadou, "Ontology-based approach for measuring semantic similarity," Engineering Applications of Artificial Intelligence, vol. 36, pp. 238-261, 2014.

[29] M. A. H. Taieb, M. B. Aouicha, and A. B. Hamadou, "A new semantic relatedness measurement using WordNet features," Knowledge and information systems, vol. 41, no. 2, pp. 467-497, 2014.