ISSN 2278-3091
Volume 9 No.2, March -April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse02922020.pdf
https://doi.org/10.30534/ijatcse/2020/02922020

Knowledge-Based Semantic Relatedness measure using Semantic features

Ali Muttaleb Hasan[1], Noorhuzaimi Mohd Noor[2*], Taha. H. Rassem[3], Ahmed Muttaleb Hasan[4]
Faculty of Computing (FKOM), University Malaysia Pahang 26300 Gambang, Kuantan Pahang, Malaysia.
[1]alimatlab65@yahoo.com, [2*]nhuzaimi@ump.edu.my, [3]tahahussein@ump.edu.my, [4]ahmed.matlab11@gmail.com

ABSTRACT

Measuring semantic relatedness has received much attention for uses in many fields such as information retrieval and natural language processing. For handling synonymous problem in distributional-based measures, many researchers are investigating how to exploit semantic features in lexical sources to form knowledge-based measures. In the knowledge-based measures, a hierarchy model is used to measure the relatedness between words based on only the taxonomical features extracted from a provided lexical source. In this paper, a new knowledge feature-based measure is proposed to build the semantic vector of a word construct on taxonomical and non-taxonomical feature of relation words. The proposed measure utilised the topological parameters that weight the importance of each element in the semantic vector. One of the gold dataset used to assess the proposed model and compare the findings with other related works. The results demonstrated the effectiveness of the proposed model on measuring semantic relatedness between words. In this paper, the research framework is identified based on the observations made on the previous related works that have been conducted for semantic representation and semantic relatedness measures. The required data in this research includes the semantic knowledge-based approach and the evaluation datasets. The semantic knowledge that will be used throughout of this research is extracted from English WordNet 3.1. On the other hand, the evaluation datasets covers the gold standard benchmarks which have been used for evaluating the semantic relatedness measurements and text mining tasks. Finally, the evaluation is preform to evaluate the proposed method (PM) based on approach in this research, in which obtained the result have been analyzed, to discuss and compare based on different performance measure and finding the strength and weakness in this paper, to alternative the semantic representation correlated to this research, to designing and develop the topical-based on the semantic representation method for text mining from Social media.

Keywords: semantic representation, semantic similarity, semantic measure, topological parameter, lexical source, machine learning, NLP.

1. INTRODUCTION

The semantic similarity is a metric of defined sets documents or terms of words, several metrics are used [1]WordNet 3.1, as a manually construct lexical of the source of English words. Despite the advantage of having human supervision in construction of the database, since the words are automatically to learn the database, so that is not the measure of relatedness between multi-words. The semantic relationship between units of language such as concepts or instances. The words of semantic similarity are often confused with semantic relatedness. The semantic relatedness includes any relation between two words of concept, while the semantic similarity only includes "is a" relations. E.g., "*car*" is similar to "*bus*", but is also related to "*road*" and "*driving*". Computationally, semantic similarity can be estimated by defining a topological similarity, by using ontologies to define the distance between words/concepts. Few investigations have used feature strategies to measure the semantic relatedness of words' meanings. The main goal of feature-based techniques in deciding semantic similitude depends on the features chosen to represent the semantics of concepts and the type of measurements used to measure the likeness between two delegate features. The result of few studies [1] [2] [3] [4] shows that the semantic similarity heavily relies on the features of a selected concept. In lexical sources, semantic features can be placed into two main categories: graph-based and feature-based [5].

In the graph-based method, a concept's meaning is represented as either semantic taxonomy or semantic ontology. The main idea behind this method comes from cognitive science the human brain depends on linking concepts to form the semantics of a given concept [6] [7]. When the brain receives a concept, it recalls other related concepts and links these concepts together to understand the degree of relationship between the meaning of the received concept and other concepts. A proposed method and

This work is supported by the University of Malaysia Pahang (UMP) via research grant UMP RDU1803141. Corresponding author: Noorhuzaimi Mohd Noor[2*], nhuzaimi@ump.edu.my[2*].

[1] https://wordnet.princeton.edu

classification represents concepts as features for providing the taxonomy and lexical relationships of any concept in WordNet from vocabulary sources [8] [1] [2] [9].

In the feature-based method, the semantics of a phrase or concept are represented as a set of mixed attributes from semantic relationships and [2]glosses in the knowledge sources. Semantic relationships include "is-a" hypernym-hypernym relationships, holonym, inverse glosses, and meronyms. Thus, the semantics of the "car" is represented as the set of features containing the terms "*vehicle*", "*convertible*", "*accelerator*", "*train*", and "*cable car*". The performance of the feature-based method depends on many factors related to the fidelity, continuity, and balance of the knowledge sources, such as WordNet. Although several knowledge sources have been created to contain the semantic features of the concepts in a language, they include classical semantic features only. The results presented in recent experiments seem to confirm that there is still a substantial disconnect between human judgment and what can be computed out of WordNet [10] [11]. This paper proposes the knowledge approach of feature-based method to weight the semantic representation approach to handle the constant weighting assumptions in feature-based measures by using topological parameters. The main problem is how to representing the feature semantic of words from lexicons, because the lexicon sources contain have many semantic features, such as synonym, "is-a" relationships, and textual define as a gloss. Fundamental test of this work is choosing the educational feature based that improved the semantic description of ideas. Therefore, the job of this element-based strategy is to decide the likeness between two ideas by relying on choosing the features to representing the semantic of words or connotation, and the metric using to quantify the comparability through two regular features.

The paper is organized as follows. Section 2 explores and discusses some related work on word-sense similarity measures semantic relatedness. In section 3, the semantic representation is discussed. While in section 4, the proposed method has been discussed. In section 5, combining features is provided. In section 6, proposed method (PM) based on feature selection. While in section 7, semantic representation on text mining. In section 8, Feature selection of topical redaction. In section 9, features of evaluation. While section 10, evaluation. Section 11 is the conclusion.

2. RELATED WORK

Several methods have been proposed to measure the semantic relatedness between words based on the knowledge sources. Some of these methods exploit the features of structural and statistical-based methods. The previous literature has been reviewed, and the semantic representation approaches studied comprehensively, including distributional-based and knowledge-based approach based on semantic representation, semantic similarity measure, and knowledge-based text mining. [12], proposed Explicit Semantic Analysis (ESA), an incoming method that represented the meaning of texts on compute semantic related of NL texts presuppose access to large amounts of common sense and domain-specific of knowledge, use machine learning techniques to explicitly represent the meaning of any text as a weighting assumption vector of Wikipedia-based concepts. To constant development so its breadth and depth steadily increase over time. While, [13], proposed the new model of lexical source semantic associated that incorporating information from every explicating or implicated paths connected the two words in the completely graph our model uses a random walk over nodes and edges derived from WordNet 3.1 linking and corpus statistical. To propose a new model of lexical source of semantic relatedness that incorporating the information from every explicating or implicated path connecting the two words in the complete graph. Furthermore, up weighting invariable distribution highly infrequent words such as by TF-IDF scoring might also be inappropriate because such scarce words would get highly high scores, which is an undesirable trait in a similarity search. Other hands, [14], introduced a reduce semantic representation method that constructs the semantic interpretation of the words as the vectors over the latent topics from the original ESA representation vectors. For modeling the latent topics, the Latent Dirichlet Allocation (LDA) is adapted to the ESA vectors for extracting the topics as the probability distributions over the concepts rather than the words in the traditional model. The proposed method is applied to the wide knowledge sources used in the computational semantic analysis: WordNet and Wikipedia, the weight of a topic is updated based on it describe the idea of the *TF-IDF* technique. In Escalante [15],[16] proposed in this article a genetic program that aims at learning effective term-weighting schemes TWSs that can improve the performance of current schemes in text classification. The genetic program learns how to combine a set of basic units to give rise to discriminative TWSs. While [17] for using TFIDF to select each word from all the weighting. While, [18] [19] proposed a novel algorithm for subgroup discovery task based on genetic programming and fuzzy logic called Fuzzy Genetic Programming-based for Subgroup Discovery (FuGePSD). In addition, the weighting is different weights (w1, w2, and w3) are used in order to give a major can be also modified through external parameters in order to facilitate the expert's adaptability to real and complex problems.

[2] https://github.com/alimuttaleb/Ali-Muttaleb/blob/master/Glosses

. **Table 1:** the summary of a few related work used the semantic measure with different technique.

Authors	Method	Model	Features	Weighting	Advantages	Disadvantages
[12]	Explicit Semantic Analysis (ESA)	Vector space model	Glosses	TF-IDF	Simple Easy to implement Competitive results on measuring semantic relatedness	High-dimensionality Generally, very excessive Cannot handle ambiguity
[13]	Random Walk method	Graph-based model	All semantic relations	Stationary distribution	- Capturing implicit relations - Extendible - Handling ambiguity issue	- Computationally expensive - High-dimensionality - Unfiltered features - Many redundant concepts in the representation
[14]	Reduced Semantic representation	Probabilistic model	Glosses	LDA	- Low dimensionality - Capturing implicit and explicit relations Achieving - Competitive results	-Computationally expensive - Inference efficiency Problem. - Based on the selection of the parameters
[20]	LSI, LCA, LDA	Graph-based	Glosses	-	- The lowest common node between the paths of these two senses from the root of WordNet	- High dimensionality - High execution times - Highest-weighted features
[21]	SIMON method	The vocabulary allows the embedding-based method to capture more word variations	Lexicon	Weighted features extracted from the input text.	- Simple classifier - capturing more relevant information - low percentile	- highest feature scores - highly correlates with almost all the rest of metrics
[18]	Fuzzy genetic programming-based	SD task	features used to describe the subgroups discovery	-	- Flexibility in the learning process due to the use of populations with dynamic size and individuals with structure and size variable.	- High-quality results Wide experimental
[22]	Particle Swarm Optimization (PSO)	SVM	-	-	- PSO finds the best value of particles. - High its convergence speed becomes very slow near the global.	- High-dimensional data. - Many thousands of dimensions.

However, [23] [24, 25] proposed a novel ensemble construction method that uses PSO generated weights to create the ensemble of classifiers with better accuracy for intrusion detection used as a meta-optimizer to find better behavioral parameters for PSO, using the new approaches as well as the weighting majority algorithm (WMA) approach. Used the PSO algorithm to weight the opinion of each expert. Because the quality of the behavioral parameters inserted by the user into PSO strongly affects its effectiveness, have used the Local unimodal sampling LUS method as a met optimizer for finding high-quality parameters. Then used the improved PSO to create new weights for each expert. Table 1 shows the summary of related methods based on the previous studies.

3. SEMANTIC REPRESENTATION

In semantic representation, the knowledge-based approach is selected to overcome synonymy and ambiguity issues in the text mining tasks.

3.1 Semantic Relation

Semantic relation is the link between two words or concepts that reveals the relevance of their semantics. Semantic relation is now utilised in a few NLP applications, such as word sense disambiguation. Semantic relation looks at the taxonomical relationships that constitute the connections between the synsets in WordNet 3.1 according to likeness to the "is-a", "has", and "lives in" aspects. It includes hyponyms, hypernyms, and meronyms. It also looks at gloss-based relationships, which are extracted from the definitions of the concepts. Finally, the important words are given more weight for the taxonomy of weighing for relevant features. Figure 3 shows the semantic relationships from WordNet 3.1 between two words.

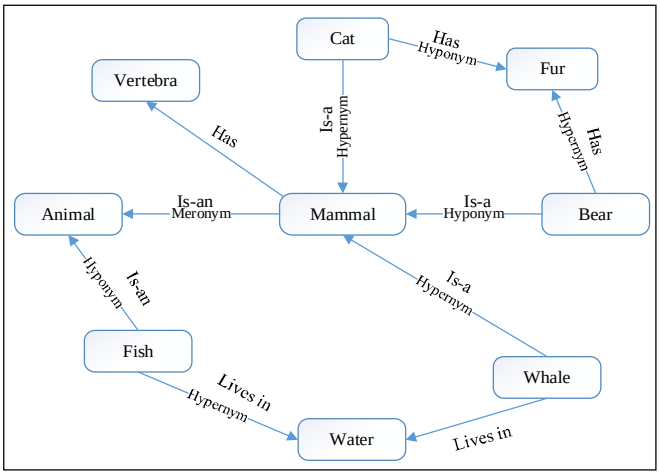

Figure 3: Semantic relationships between two words.

3.2 Semantic Taxonomy

The semantic taxonomy of a knowledge-based least subsumer "LS" is represented by the "is-a" relation. Semantic taxonomy is a network of concepts in the lexicon where nodes indicate concepts, and the edges indicate hyponyms and hypernyms. The edges are the relationships of nominal and verbal synsets of the inherent semantic taxonomy. Figure 4 shows part of the semantic taxonomy of WordNet.

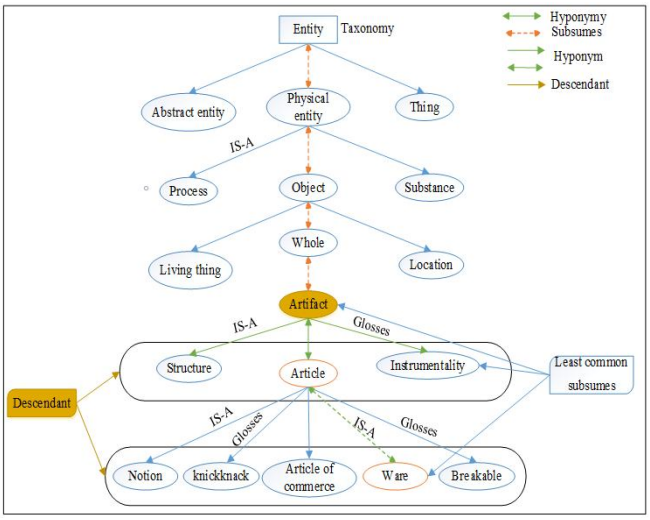

Figure 4: Semantic taxonomy relationship between two concepts.

4. PROPOSED METHOD

4.1. Feature Extraction

This step is to form the semantic features of each concept that can be used to give the details of the meaning of the head term. The features have been classified into [3]synonymous, taxonomical relations, and non-taxonomical relations. The first type of the feature is a synonymous, a "*word or phrase*" as which, lexeme (word or phrase) in the same natural language, the words that are synonym is said to be synonymous, and the status of being a synonym is called synonymy. For example, the words start, begin, set, launch, commence and initiate are all synonyms of one more. The second type of features is a taxonomical relation's feature consists of two relations: Hyponym and [4]Hypernym. Most of the previous researches have been used this feature to build the semantic representation of the synsest. This feature gives competitive results in many semantic analysis applications such as semantic similarity. The third type of features is a non-taxonomical relation which constitutes the relationships among the synsets in the WordNet 3.1 according to any

[3]https://github.com/alimuttaleb/AliMuttaleb/blob/master/Synonym.txt
[4]https://github.com/alimuttaleb/AliMuttaleb/blob/master/Hypernym

aspects of the likeness except the 'is-a' aspect. This type includes the following features: Domain-region, Domain-topic, Domain-usage, Member-holonym, and Member-meronym.

4.2. Weighting

This step depends on the TP in the semantic taxonomy for weighting each feature according to its importance in the semantic representation. Although there are several topological parameters in the semantic taxonomy, we used the descendants and the depth of the concept since these parameters yield better evaluation results in the previous studies. The weight of the concept c in the semantic taxonomy is defined as the following formula:

$$w(c) = \left(\frac{depth(c)}{Max_{depth}}\right) \times \left(\log\left(\frac{|T|}{|descendant(c)|+1}\right)\right), \quad (1)$$

Where $|T|$ the number of concepts is in the semantic taxonomy, Max_{depth} is the depth of the semantic taxonomy.

4.3. Similarity Measure

The semantic similarity measure is a formula that computes the likeness of two words based on the semantic taxonomy, which has been constructed using hypernyms and hyponyms. Although there are many techniques through which semantic similarity can be measured for two words or concepts using different relationships, semantic similarity is limited to hypernym-hyponym relationships only, [26]. Figure 6 shows the similarity measure between words extracted from a sentence.

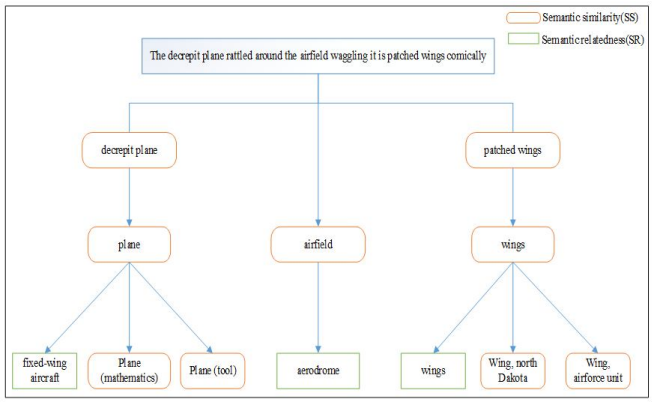

Figure 6: The semantic similarity measure between words in a sentence.

After the semantic signature representations for a pair of concepts have been obtained, the semantic similarity of two concepts is calculated by comparing their corresponding semantic representation vectors. In comparing the semantic representation vectors, the method that has been used extensively in previous work to compare the vectors and the *cosine* is used in this research. The cosine metric is defined to measure the similar among of two concepts c_1 and c_2 as following:

$$Sim(c_1, c_2) = \frac{\sum_{c' \in R(c_1) \cap R(c_2)} w_1(c') \times w_2(c')}{\sqrt{\sum_{c' \in R(c_1)} w_1(c')^2} \sqrt{\sum_{c' \in R(c_2)} w_2(c')^2}}, \quad (2)$$

Where $w_1(c)$ and $w_2(c)$ are the weighting of the concept c in semantic representation (R) of the concepts c_1 and c_2 respectively.

Given the ambiguity issue, polysemy words in WordNet 3.1 that represented by several concepts (at least two concepts). Thus, the semantic similarity between two words depends on the similarity between their corresponding concepts. The semantic similarity between two words is computed in this work based on the maximum technique to ensure a fair comparison of the results. Firstly, all of the concepts that corresponding to the senses of each word are extracted from WordNet 3.1. Secondly, the semantic similarity is computed for each pair of concepts in the combination of candidates. Finally, the maximum semantic similarity score is selected from the candidates as the final similarity between two words. Formally, given two potentially polysemous words w_1 and w_2, the semantic similarity between them is computed as follows:

$$Sim(t_1, t_2) = \max_{i,j} sim\ (c_{1i}, c_{2j}), \quad (3)$$

Where c_{1i} and c_{2j} are the concepts extracted from WordNet 3.1 for the words t_1 and t_2, respectively.

5. COMBINING FEATURES

This step is to combine two types of features of each concept to form the semantics of the concept from different aspects. For measuring the semantic relatedness between two concepts, each concept is represented by two sets of features for the types of semantic features. Then, each set from the semantic representation of the first concept is compared to the corresponding setting from the semantic representation of the second concept using the cosine measure. Finally, the maximum value of measuring the semantic relatedness from the combination of features is selected to be the relatedness score between two concepts.

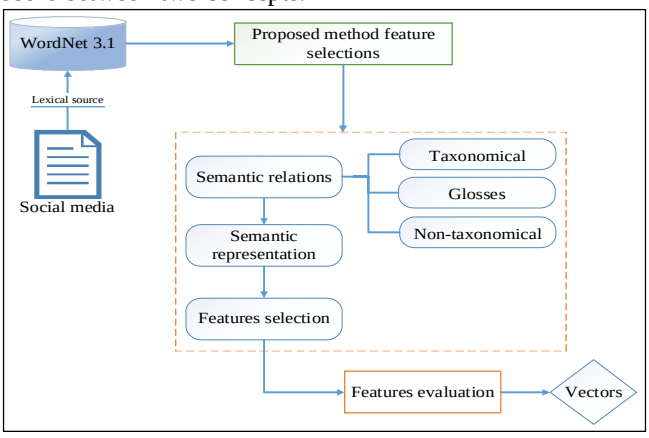

Figure 7: The proposed method of semantic representation.

6. PROPOSED METHOD BASED ON FEATURE SELECTION

In order to evaluate the proposed methods, the manual golden standard dataset is required to compare the performance of the machine-based methods against the human annotations.

6.1 Semantic Relation

The basic of semantic relation is of the hierarchical of feature-based semantic relatedness between words is that tacitly by supposes that each unit in the representative of features of a given concepts has the same semantic relatedness to other units. Although, the units in the semantic representation is taxonomically relatedness to extract the computing of the semantic similarity. To combine a number of semantic features for measuring semantic relatedness, text mining tasks require special investigation to select features. To classify these features as follow:

6.1.1 Taxonomical Relation

The goal of taxonomical is overcome the limitation that are observed from other related work. However, the important words have more weight, for taxonomy of weight relevant features. In addition, to a chive the goal of this paper and answer the key research question, an experimental methodology is adopted. In table 1 the relation shows the taxonomy of relation between these types are (Synonym, hypernym, glosses and non-taxonomy) from the type will take the relation between words/concepts. Furthermore, each of these words class is forms of the sysnset of the nodes and relation of concepts of filed as which "Offset_1, Offset_2" and the type of relationship in the WordNet 3.1, to represent them as a hypernym, related form, and hyponym relations.

Table 1: The relation of consists of the three fields "Offset_1, Offset_2" and the type of relation.

Offset_1	Offset_2	Relation_Type
00002137	00001740	HYPERNYM
00002137	00694095	RELATED_FORM
00002137	00023280	HYPONYM
00001930	14604577	HYPERNYM

In each of these words class is forms as a graph sysnest of node and relation of the consists of three of filed "Offset_1, Offset_2" and the type of relationship in the WordNet 3.1, the Hypernym, Related form, and Hyponym relations are central to the organization of the nouns in WordNet 3.1 figure 4.3 show the relations of the consists of three types of relations fields [27, 28].

6.1.2 Glosses

In a certain of knowledge source (KS), let the C= $\{c_1, c_2 \ldots, c_n\}$ is the set of distinguished concepts that are expressed by semantic representation. G= $\{g_1, g_2 \ldots, g_n\}$ is the set of glosses textual definitions of corresponding concepts, such as gloss can be considered as the document. T= $\{t_1, t_2 \ldots, t_m\}$ is the set of vocabularies in collection of the glosses. The semantic representation of a given term t_i is defined as a vector.

6.1.3 Non-taxonomy Relation

In our own particular case, the evaluation suggests that a hypothesis semantic representation have use, taking the best case methods for each task and combining them appropriately, the semantic representation will combined a new method in the feature-based methods to select the accurate feature from the new methods as which, Taxonomy Relations, Non Taxonomy Relations, Glosses and Proposed Method to get each method the accurate result, for a given corpus, the distributional methods relying on the distributional hypothesis for extracting the semantic for each word. As we mentioned in propose a weighting semantic representation approach for building semantic vectors of the textual data.

7. SEAMNTIC REPRESENTATION ON TEXT MINING

The cosine similarity equation in the semantic representation of the angle between the two data points of the documents, whereas Euclidean distance is the square root of entity straight line differences between data points. The cosine similarity equation will result in a value between 0 and 1, the smaller cosine angle results in a bigger cosine value, indicating higher similarity, in this case Euclidean distance will be zero. That's what have to find the similarity between two pair's words. The equation of these distance as follow:

$$D(\chi, \gamma) = \sum_{i=1}^{n} (\chi_i, \gamma_i)^2 \quad (4)$$

$$sim(A, B) = \cos(0) = \frac{x * y}{\|x\| \|y\|} \quad (5)$$

8. FEATURE SELECTION OF TOPICAL REDACTION

The topical feature reduction method of design and development is to propose a novel knowledge-based method for representing the semantics of the textual documents in the text mining tasks. It includes the main contribution of the current research which will prove that the knowledge-based measure used 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 such as synonymy, ambiguity, and high dimensionality of the text mining tasks.

9. FEATURES OF EVALUATION

The features are focus on handling the relatedness, to challenge is how to select the informative features that improve the semantic representation based on the proposed method of concepts. To introduce the proposed method for semantic representation based on the lexical and semantical features of the well-known knowledge source (called WordNet 3.1). The proposed method will be referred to as the feature-based semantic representation method. The input of this approach is the semantic features (semantic and lexical relations, and glosses) in certain lexical resources. The corpus of semantic representation is collection the documents/texts in the structured of information form, as which useful for extracting many of features of language. In this work, the domain is limited to the social media, as which Facebook, Tweeter, LinkedIn and YouTube, this research focus on the politic dataset.

10. EVALUATION

We compared four measures for each measure, compared the findings with a dataset of [5]MC30, and took the results from the coefficient of Pearson, Spearman, M, and Nonzero correlations. The Nonzero correlation was the final accuracy of the result. To perceive how well they reflect human decisions about semantic relations, we contrasted our semantic measure and the other nine techniques as mentioned in Figure 6 [29], [30], [31], [32], [33], [34], [35], [36], [37]. After that, we used the results obtained in this section to evaluate the features-based measure.

10.1 Evaluation Measure

The evaluation of the proposed method is carried out by computing the Pearson correlation coefficient between the Human Judgments and the scores of the proposed method. There are two correlation coefficients, which have been used extensively in the direct evaluation technique: Pearson product-moment correlation r and Spearman's rank correlation ρ. This correlation coefficient is computed between the list of human judgments (X) and the list of values for the semantic relatedness measure (Y) as shown in the following formula:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (6)$$

Where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y respectively, n is the size of X and Y.

10.2 Results

To explore the connection between the similarity of context and similarity of meaning in one of the highest quality level datasets, we picked the taxonomy of WordNet and got 30 sets of synonymy decisions from 30 human subjects on 30 sets of words as shown in Table 2. The sets went from "highly synonymous" to "semantically random", and the subjects were asked to the similarity of significance rate them on a scale of 0 to 4.

[5]https://github.com/alimuttaleb/semantictaxonomy/blob/master/mc30.txt

Table 2: Human Decisions and Computer Appraisals for the MC30 Set of Word Sets.

Word Pair	Human Judgment	Taxonomy Relation	Non-Taxonomy	Glosses	Proposed Method
“Car” + “Automobile”	3.92	1.000	1.000	1.00	1.000
“Gem” + “Jewel”	3.84	1.000	1.000	1.00	1.000
“Journey” +“Voyage”	3.84	0.760	0.640	0.00	0.760
“Boy” +“Lad”	3.76	0.680	0.710	0.00	0.710
“Coast” + “Shore”	3.70	0.720	0.470	0.00	0.720
“Asylum” +“Madhouse”	3.61	0.830	0.010	0.00	0.830
“Magician” +“Wizard”	3.50	1.000	1.000	1.00	1.000
“Midday” +“Noon”	3.42	1.000	1.000	1.00	1.000
“Furnace” +“Stove”	3.11	0.017	0.020	0.63	0.020
“Food” +“Fruit”	3.08	0.008	0.010	0.00	0.010
“Bird” + “Cock”	3.05	0.550	0.050	0.00	0.550
“Bird” + “Crane”	2.97	0.440	0.009	0.00	0.440
“Tool” + “Implement”	2.95	0.710	0.940	0.00	0.940
“Brother” +“Monk”	2.82	0.750	0.210	0.00	0.750
“Crane”+ “Implement”	1.68	0.120	0.003	0.00	0.120
“Lad” +“Brother”	1.66	0.053	0.460	0.00	0.460
“Journey” +“Car”	1.16	0.000	0.010	0.00	0.010
“Monk +“Oracle”	1.10	0.030	0.040	0.00	0.040
“Cemetery” + “Woodland”	0.95	1.800	0.007	0.00	0.007
“Food” +“Rooster”	0.89	7.500	0.000	0.00	7.500
“Coast” +“Hill”	0.87	0.150	0.080	0.00	0.150
“Forest” + “Graveyard”	0.84	0.001	0.010	0.00	0.010
“Shore” + “Woodland”	0.63	0.003	0.350	0.00	0.350
“Monk” + “Slave”	0.55	0.074	0.000	0.00	0.070
“Coast +“Forest”	0.42	0.002	0.008	0.00	0.008
“Lad” +“Wizard”	0.42	0.057	0.017	0.00	0.050
“Chord” + “Smile”	0.13	0.014	0.008	0.00	0.014
“Glass” +“Magician”	0.11	6.900	0.000	0.00	0.006
“Noon +“String”	0.08	5.900	0.000	0.00	5.900
“Rooster” + “Voyage”	0.08	0.000	0.000	0.00	0.000
Correlation Averages					
Pearson	N/A	0.840	0.660	0.50	0.820
Spearman	N/A	0.780	0.740	0.62	0.800
M	N/A	0.810	0.700	0.56	0.810
Nonzero	N/A	**0.930**	**0.830**	**0.16**	**0.960**

10.3 Compare with the stat of art measures

Table 3 and Figure 8 show that our proposed technique had strong outcome among the nine strategies compared to the coefficient of connection between human decisions and evaluations of semantic comparability measure. These results demonstrate the great execution of our measure. The contrasted aftereffects of our semantic relations and the other nine techniques are displayed in Figure 8.

Table 3: Coefficients of Connection between Human Judgment Evaluations of Semantic Measure.

Coefficient of Correlation with Human Judgement	Semantic Measure (MC30)
Rada	0.70
Wu & Palmer	0.75
Li	0.77
Leacock & Chodorow	0.75
Sebti & Barfroush	0.8
Sanchez [29]	0.78
Meng & Gu	0.81
Hadj Taieb	0.76
Aouicha & Ezzeddine	0.79
Our Proposed Method (PM)	**0.82**

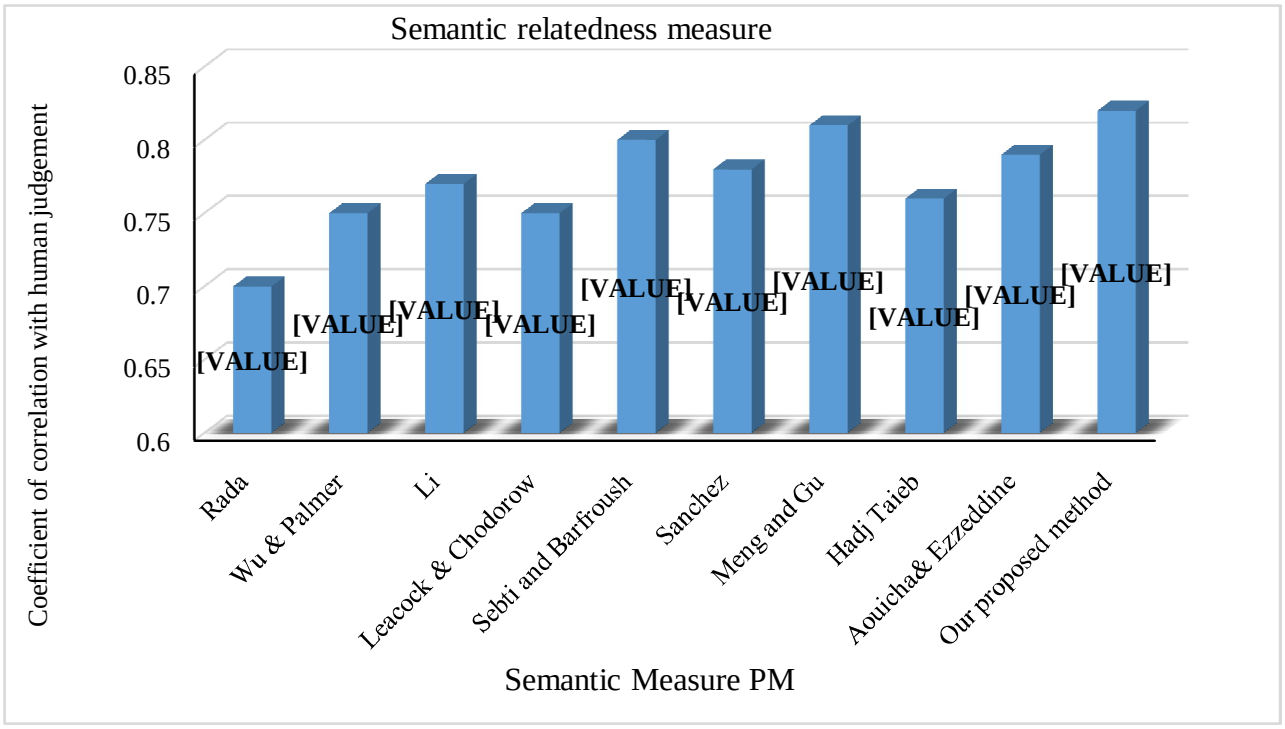

Figure 8: Benchmarking with another model.

11. CONCLUSION

In this paper, we proposed a knowledge feature-based method for semantic representation to weight semantic relationships that overcomes the primary confinement of the research question. The proposed semantic-based model found a similarity measure of 0.82, which shows that the proposed model performed well. The dataset used for this study was the gold standard dataset of Taxonomy by using WordNet 3.1. The proposed model showed that the feature-based measure is more accurate than the proposed measure in various assessment datasets. In future work, more datasets will be used to evaluate the proposed model.

ACKNOWLEDGMENTS

This research has been supported by University Malaysia Pahang (UMP), grant number (RDU1803141, PGRS190398).

REFERENCES

1. Jiang, Y., et al., *Feature-based approaches to semantic similarity assessment of concepts using Wikipedia.* Information Processing & Management, 2015. **51**(3): p. 215-234. https://doi.org/10.1016/j.ipm.2015.01.001
2. Saif, A., M.J. Ab Aziz, and N. Omar, *Mapping Arabic WordNet synsets to Wikipedia articles using monolingual and bilingual features.* Natural Language Engineering, 2017. **23**(1): p. 53-91.
3. Jiang, X., et al. *Deep compositional cross-modal learning to rank via local-global alignment.* in *Proceedings of the 23rd ACM international conference on Multimedia*. 2015. ACM.
4. Al-Tashi, Q. and A.M. Hasan, *WORD SENSE DISAMBIGUATION: A REVIEW.* Southern Connecticut State University, Hilton C. Buley Library, 2019. **1,2**: p. pp. 20-458.
5. Maghrebi, W., et al., *An Intelligent mutli-object retrieval system for historical mosaics.* Editorial Preface, 2013. **4**(4).
6. Griffiths, T.L., M. Steyvers, and J.B. Tenenbaum, *Topics in semantic representation.* Psychological review, 2007. **114**(2): p. 211. https://doi.org/10.1037/0033-295X.114.2.211
7. Rassem, T.H., et al. *Restoring the missing features of the corrupted speech using linear interpolation methods.* in *AIP Conference Proceedings.* 2017. AIP Publishing.
8. Hasan, A.M., T.H. Rassem, and M. Karimah, *Pattern-Matching Based for Arabic Question Answering: A Challenge Perspective.* Advanced Science Letters, 2018. **24**(10): p. 7655-7661.
9. Saif, A., M.J. Ab Aziz, and N. Omar, *Evaluating knowledge-based semantic measures on Arabic.* International Journal on Communications Antenna and Propagation, 2014. **4**(5): p. 180-194.
10. Hasan, A.M., T.H. Rassem, and M. Noorhuzaimi. *Combined Support Vector Machine and Pattern Matching for Arabic Islamic Hadith Question*

Classification System. in *International Conference of Reliable Information and Communication Technology*. 2018. Springer.

11. Lastra-Díaz, J.J. and A. García-Serrano, *A novel family of IC-based similarity measures with a detailed experimental survey on WordNet.* Engineering Applications of Artificial Intelligence, 2015. **46**: p. 140-153. https://doi.org/10.1016/j.engappai.2015.09.006
12. Gabrilovich, E. and S. Markovitch. *Computing semantic relatedness using wikipedia-based explicit semantic analysis*. in *IJcAI*. 2007.
13. Hughes, T. and D. Ramage. *Lexical semantic relatedness with random graph walks*. in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*. 2007.
14. Saif, A., M.J. Ab Aziz, and N. Omar, *Reducing explicit semantic representation vectors using Latent Dirichlet Allocation.* Knowledge-Based Systems, 2016. **100**: p. 145-159.
15. Escalante, H.J., et al., *Term-weighting learning via genetic programming for text classification.* Knowledge-Based Systems, 2015. **83**: p. 176-189.
16. Al-Tashi, Q., et al., *Binary Optimization Using Hybrid Grey Wolf Optimization for Feature Selection.* IEEE Access, 2019.
17. Hasan, A.M. and L.Q. Zakaria, *QUESTION CLASSIFICATION USING SUPPORT VECTOR MACHINE AND PATTERN MATCHING.* Journal of Theoretical & Applied Information Technology, 2016. **87**(2).
18. Carmona, C.J., et al., *A fuzzy genetic programming-based algorithm for subgroup discovery and the application to one problem of pathogenesis of acute sore throat conditions in humans.* Information Sciences, 2015. **298**: p. 180-197.
19. Al-Tashi, Q., H. Rais, and S. Jadid. *Feature Selection Method Based on Grey Wolf Optimization for Coronary Artery Disease Classification.* in *International Conference of Reliable Information and Communication Technology*. 2018. Springer.
20. Wei, T., et al., *A semantic approach for text clustering using WordNet and lexical chains.* Expert Systems with Applications, 2015. **42**(4): p. 2264-2275. https://doi.org/10.1016/j.eswa.2014.10.023
21. Araque, O., G. Zhu, and C.A. Iglesias, *A semantic similarity-based perspective of affect lexicons for sentiment analysis.* Knowledge-Based Systems, 2019. **165**: p. 346-359.
22. Esmin, A.A., R.A. Coelho, and S. Matwin, *A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data.* Artificial Intelligence Review, 2015. **44**(1): p. 23-45.
23. Aburomman, A.A. and M.B.I. Reaz, *A novel SVM-kNN-PSO ensemble method for intrusion detection system.* Applied Soft Computing, 2016. **38**: p. 360-372. https://doi.org/10.1016/j.asoc.2015.10.011
24. Al-Tashi, Q., H. Rais, and S.J. Abdulkadir. *Hybrid Swarm Intelligence Algorithms with Ensemble Machine Learning for Medical Diagnosis.* in *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*. 2018. IEEE.
25. Omar, N. and Q. Al-Tashi, *Arabic nested noun compound extraction based on linguistic features and statistical measures.* GEMA Online® Journal of Language Studies, 2018. **18**(2).
26. Budanitsky, A. and G. Hirst, *Evaluating wordnet-based measures of lexical semantic relatedness.* Computational Linguistics, 2006. **32**(1): p. 13-47. https://doi.org/10.1162/coli.2006.32.1.13
27. Hasan, A.M., et al. *A Semantic Taxonomy for Weighting Assumptions to Reduce Feature Selection from Social Media and Forum Posts*. in *International Conference of Reliable Information and Communication Technology*. 2019. Springer.
28. Hasan, A.M., et al., *A Proposed Method Using the Semantic Similarity of WordNet 3.1 to Handle the Ambiguity to Apply in Social Media Text*, in *Information Science and Applications*. 2020, Springer. p. 471-483.
29. Sánchez, D., D. Isern, and M. Millan, *Content annotation for the semantic web: an automatic web-based approach.* Knowledge and Information Systems, 2011. **27**(3): p. 393-418.
30. Rada, R., et al., *Development and application of a metric on semantic nets.* IEEE transactions on systems, man, and cybernetics, 1989. **19**(1): p. 17-30. https://doi.org/10.1109/21.24528
31. Wu, Z. and M. Palmer. *Verbs semantics and lexical selection.* in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. 1994. Association for Computational Linguistics.
32. Li, Y., 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, 2003. **15**(4): p. 871-882.
33. Leacock, C. and M. Chodorow, *Combining local context and WordNet similarity for word sense identification.* WordNet: An electronic lexical database, 1998. **49**(2): p. 265-283.
34. Sebti, A. and A.A. Barfroush. *A new word sense similarity measure in WordNet.* in *Computer Science and Information Technology, 2008. IMCSIT 2008. International Multiconference on*. 2008. IEEE. https://doi.org/10.1109/IMCSIT.2008.4747267
35. Meng, L., 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, 2012. **5**(3): p. 81-94.

36. Taieb, M.A.H., M.B. Aouicha, and A.B. Hamadou, *A new semantic relatedness measurement using WordNet features.* Knowledge and information systems, 2014. **41**(2): p. 467-497. https://doi.org/10.1007/s10115-013-0672-4
37. Aouicha, M.B., M.A.H. Taieb, and M. Ezzeddine, *Derivation of "is a" taxonomy from Wikipedia Category Graph.* Engineering Applications of Artificial Intelligence, 2016. **50**: p. 265-286.