**International Journal of Electrical and Computer Engineering (IJECE)**
Vol. 9, No. 6, December 2019, pp. 5386∼5394
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i6.pp5386-5394 

# A computational analysis of short sentences based on ensemble similarity model

**Arifah Che Alhadi, Aziz Deraman, Masita Masila Abdul Jalil, Wan Nural Jawahir Wan Yussof, Rosmayati Mohemad**
Software Technology Research Group (SofTech), School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, Malaysia

**Article Info**

***Article history:***

Received Apr 14, 2019
Revised May 29, 2019
Accepted Jun 10, 2019

***Keywords:***

Ensemble
Text similarity
Vector space model
Edit distance
Short sentences

**ABSTRACT**

The rapid development of Internet along with the wide use of social media applications produce huge volume of unstructured data in short text form such as tweets, text snippets and instant messages. This form of data rarely contains repeated word. It presents challenge in sentences similarity analysis as the standard text similarity models merely rely on the number of word occurrence, often resulting unreliable similarity value. Besides, the use of abbreviation, acronyms, slang, smiley, jargon, symbol or non-standard short form also contributes to the difficulty in similarity analysis. Thus, an extended ensemble similarity model approach is proposed. An experimental study has been conducted using datasets of English short sentences. The findings are very encouraging in improving the similarity value for short sentences.

*Copyright © 2019 Institute of Advanced Engineering and Science.*
*All rights reserved.*

***Corresponding Author:***

Arifah Che Alhadi,
Senior Lecturer,
Software Technology Research Group (SofTech),
School of Informatics and Applied Mathematics,
Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia.
Email: arifah_hadi@umt.edu.my

## 1. INTRODUCTION

Measuring the similarity are important in various natural language processing applications and information retrieval applications such as information extraction, document clustering [1, 2], categorization or classification [3], language modeling [4] and ontology mapping [5]. Similarity measure can represent the similarity between two documents, two queries or one document and one query. It is possible to rank retrieved documents in order of presumed importance. A similarity function which computes the degree of similarity between pair of text objects.

The similarity analysis has been done between queries [6], documents [1], text snippets [7, 8], short segment [6, 9], tweets [10, 11] or question answer (QA) [12, 13]. Most of the preceding works on semantic similarity or combining the semantic and lexical model relies on additional information derived from large corpora, dictionary [14, 15] or background knowledge such as WordNet or ConceptNet [16, 17]. Thus these work highly dependence on third source information.

The increasing ease of access to the Internet is cause of recent explosion of users activity in online communication through social media. The amount of unstructured data generated from user's interaction is growing rapidly and publicly available [18]. Detecting and retrieving such information can facilitate people from overwhelming or overload information.

The nature of social media platform which has the limitation of messages which make the user tend to compact the text by using abbreviation, slangs, jargon, symbol or non standard short form [19, 20]. The existing similarity model based on lexical, semantic and ensemble based model still suffer from data sparsity and noise. It is because the dependency of this model on NLP tools or third source such as corpus, dictionary or background knowledge.

This work proposes an ensemble similarity model that uses multiple lexical-based similarity model to overcome this issue as well as to demonstrate the applicability and significant contribution of the proposed model.

## 2. RELATED WORK

A wide variety of text similarity model exists in the literature. Referring to a survey conducted by [21], words can be identical similar through lexical and semantic. Lexical-based identify the words similarity through its lexical which it has similar character sequence or word matching. Whereas semantic words similarity is based on the meaning of the word rather than character matching. Semantic similarity is computed based on corpus and knowledge information gather from large corpora or semantic networks such as WordNet [22, 23].

The ensembles of similarity model is proposed by employing ensemble or hybrid model to get a comprehensive measure which integrates different context features for similarity measuring including lexical or semantic model. In preceding work, ensemble models have been recommended by the researchers which proven give record of better performance when applying combine similarity model, compare to single model of measuring of similarity. The aim of an ensemble based similarity model is to ensemble multiple models of classifiers or features to solve various problems [24].

Existing combination models can be categorized into different types such as combine classifier or models [16, 25, 26] combine features [11, 27, 28, 29] and combine feature and classifier or models [7, 30, 31]. M. A. Sultan, S. Bethard, and T. Sumner in [28] measured the similarities of text by combining a vector similarity feature derived from the word embedding with alignment based similarity. The other model for the semantic measurement between short texts was proposed by [32] using combinations of two different features: (1) distributed word representation, (2) corpus and knowledge-based metrics. Later, the presented method was tested and evaluated by using datasets of Microsoft Research Paraphrase Corpus and SemEval2015.

This work is different with [30] where the hybrid model were used to reduce the impact of less informative terms. The hybrid model is based on *tf-idf* information and semantic word embedding. These combination leads to a greater model for short text semantic content. Combinations of different lexical, syntactic and semantic similarity measures have been carried out by [16, 31] in their research to tackle problem in identifying paraphrase and assessing sentence similarity. R. Ferreira et al in [16] also combined with statistical model where syntactic similarity between sentences was measured using the relation of the syntactic layer calculated by matching the vertices of the RDF triples. Semantic Role Annotation (SRA) was used to identify the semantic function of each RDF graph entities and to identify the meaning in the sentence.

The focus of the work by [25] used knowledge and corpus-based similarity model to estimate the similarities of the corresponding words. They presented two systems for automated measuring of semantic similarity of short texts that is submitted to SemEval-2012 Task 6. However, their work was relied on semantic network (WordNet) and the Latent Semantic Analysis (LSA) over a large corpus to estimate the distribution. Dataset from WordSim353 was later used to compared and evaluate the model.

J. Oliva et al in [33] introduced new method namely SyMSS that used a combination of syntactic and semantic information from WordNet to compute sentence semantic similarity of two sentences. The sentences were represented as syntactic dependence tree which the idea was the meaning of terms can be seen through its syntactic connections among terms. The semantic information was obtained using WordNet.

The statistical information within the short text snippets pair contained in the corpus involved is combined with semantic information by [7]. The WordNet was employed as lexical database to compare with word similarity measure. CMU newsgroup dataset was used to simulate short text clustering scenarios. Through the proposed method, initial similarities were established between words via lexical database. Later, the method calculates iteratively words similarities and short text similarities and finally, the proximity metric was constructed and used to convert the raw text snippets into vectors. However, in this short text modelling method, the corpus that provides the context for understanding the particular meanings of the words usually needs to contain several thousands of text snippets However, the corpus need to contain thousands of text snippet to provides the meaning of particular context of the words.

The combine model also been used to analyze Malay sentences by Noah et al. [26] where the word order similarity was ensemble with semantic similarity. The word order similarities was relied on the calculation of vector space model (cosine), however base on the experimental results, this model is not effectively to be used alone. Open dictionary was use as a corpus for semantic analysis and the its shows the dependency of another third resource.

Although, significant progress has been made by some well-known text similarity model, such as the lexical and semantic similarity model, they are not truly supporting the short text retrieval with respect to the limitation of the length and changes the way of user's writing the sentence to compact with the length which limit their applicability. In an effort to address this limitation, this work has focused on developing an extended ensemble similarity model that consist of both lexical and spelling error model as briefly discussed in the following section.

## 3. LEXICAL-BASED SIMILARITY MODEL

Lexical-based similarity model were originally applied to identify the words with similar string sequences and character composition [34]. In contrast on relying on a single measure, our model relies on combination of lexical similarity models which is combination of VSM (cosine) and edit distance (Damerau-Leveinstein Distance) model.

The proposed methodology is applied to benchmark dataset by [27]. This dataset has been used in many studies [16, 22, 35]. This dataset consist of 65 pairs or noun definitions of terms. However, the previous work by [22, 35] and [36] generally considered only a subset of 30 pairs.

### 3.1. Vector space model

Figure 1. shows the process for computing the sentence similarity between two short text using cosine similarity model. In adopting the cosine model, each sentences is tokenized and a set of joint distinct word if formed. A joint distinct word set of $W$ is formed between $s$ and $t$ as $W = s \cup t$, where $s = w_1, w_2, ......., w_n$; and $t = v_1, v_2, ......., v_n$. The joint distinct word set is used to construct term-document matrix.

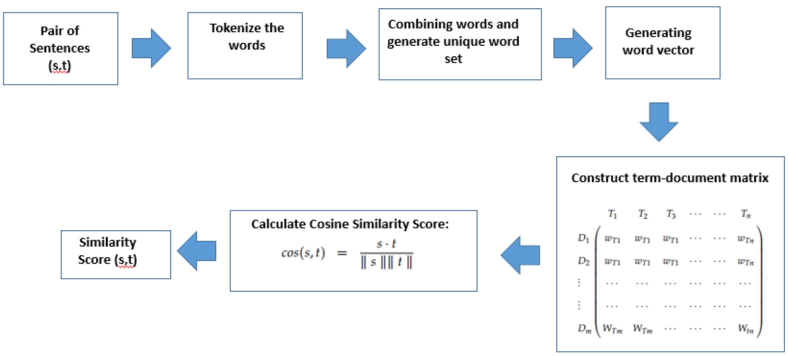


Figure 1. The process for Cosine similarity between sentences

In a term-document matrix, a document vectors is represent the frequency of terms occurrence in document collections. The rows of the matrix represent to the documents in the collections and columns corresponds to term. The matrix value indicate the terms appearance in pairwise short text i.e zero value are set for unavailable terms and non-zero value for the occurrence terms in a pairwise short text. The non-zero values of each entry in the matrix are set based on frequency of terms occurrence within a document using *tf-idf* scheme. The frequencies of terms in a document tend to indicate the relevance or similarity of the document to each other.

A pair of sentences midday:noon, $s$ and $t$ from [27] dataset which:

(a) *s : Midday is 12 o'clock in the middle of the day*

(b) *t : Noon is 12 o'clock in the midle of the dey* where *t* is used as the input sentences and *s* as the pair sentences from the database. The *s* is represent as sentence with correct spelling paired with the spelling errors of the sentence *t*. From this pair of sentences, the join set is generate st = [midday, noon, is, 12, o'clock, in, the, middle, midle, of, the, day, dey]. With the given of two vectors, the similarity value of *s* and *t* can be measure by calculating their cosine product based on matrix constructed. Thus the cosine similarity value for aforementioned sentences are 0.67. However the cosine similarity value is much lower than the mutual agreement of 32 raters which indicate that those similarity value of that particular pair of sentences is 0.96.

### 3.2. Edit distance

The edit distance between sentence *s* and *t* is defined as the minimum number of edit required to transforming *s* into *t* or vise versa. The DLD allows insertion, deletion, substitution and transposition for two adjacent characters. Damerau stated that these four operations correspond to more than 80% of all human misspellings [37].

Figure 2 shows the process of calculating the DLD between two sentences. It involves three main steps which are constructing the matrix, mapping the character to the suitable conditions and finally calculating the similarity score.

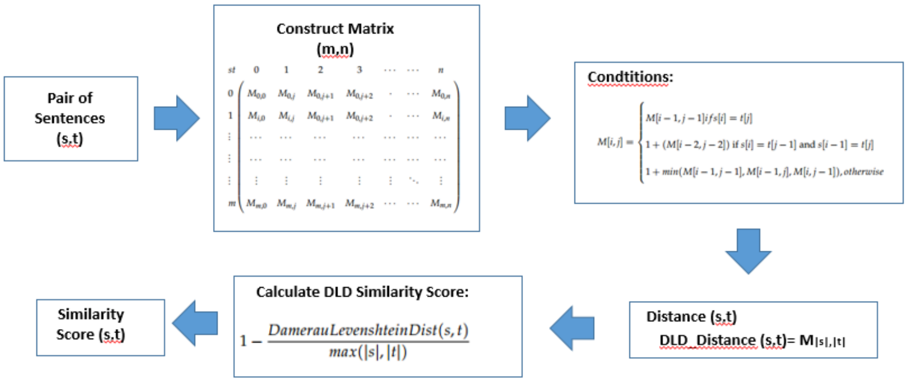


Figure 2. The process for DLD similarity between sentences

The aforementioned pair of sentences was again used to calculate the DLD. The bigger edit distance value implies that the sentences *s* and *t* are not similar. Then the similarity value is 0.82 was calculated as follow:

$$Sim_{DamerauLevenshtein}(s,t) = 1 - \frac{DLD(s,t)}{max(|s|,|t|)} \tag{1}$$

## 4. ENSEMBLE MODEL

Ensemble based models have recently gain attention due to their reported results which is better than using single model alone [6, 15, 26, 31]. The aim of these ensemble model is to combine multiple models to solve the short text similarity problems containing noise (spelling error, short form, jargon etc). The models can be combined by several approaches such averaging [27, 26], majority vote, weighted majority vote [38] and boosting [39]. The proposed ensemble model is based on averaging of VSM and edit-distance, where both uncertainty and reliability of each single model are taken into account.

The proposed ensemble models is a combinations of VSM (cosine similarity) and edit-distance similarity (DLD) model for analysing the text similarity contents of social media platform. The proposed ensemble model will calculate the overall similarity between two sentences by a linear combinations as follows:

$$Sim_{combinr} = \delta Sim_{cosine} + (1-\delta)Sim_{DLD} \tag{2}$$

The ensemble averaging model is same as used by [27] and [26].

## 5. RESULT AND DISCUSSION

The experimental result is based on dataset by [27]. The lack of suitable evaluation dataset is greatest obstacle that hinders for evaluating our proposed EXSIMO. Y. Li et al in [27] used this benchmark dataset for evaluating algorithms to measure the semantic similarity of the short text. The dataset was modified as shown in Table 1 for pair number two and three which contain spelling errors in the sentence and the changes of the sentence structure.

The testing results of the selective sentences are as illustrated in Table 1. Table shows the similarity value of cosine similarity, DLD and combination of both model for pairwise analysis on short sentence (*s*) with sentence in database (*t*). The $\overline{X}$ denote the mean value of 32 raters agreement. The different pair of sentences was tested to show the applicability of proposed combine model with single model. The full result was represented by graph as shown in Figure 3.

In general, the proposed ensemble model gives superior results and its satisfactory as compared to the single baseline model. The result of the proposed ensemble model in Table 1 indicates a consistent similarity value compared with $\overline{X}$ with very minimal differences.

The similarity value produced by the proposed ensemble model is significant under dissimilar pair. In the case of most selection sentences pair, the similarity value of dissimilar sentence is decreased. For example for the pair of *cord:smile*, the proposed ensemble model producing 0.13 of similarity value compared with cosine 0.11 and DLD are 0.15. The DLD give higher similarity value than cosine, even though the sentences are totally dissimilar in word and meaning. Then the proposed model overcome these problems by averaging the similarity value and automatically decrease the similarity value for dissimilar pair of sentences.

Table 1. Comparison of Similarity Models using Dataset by [27]

| Words pair | Sentence pair | $\overline{X}$ | $Cos\theta$ | DLD | Ensemble |
|---|---|---|---|---|---|
| *1. midday:noon* | | | | | |
| | - Midday is 12 o'clock in the middle of the day.<br>- Noon is 12 o'clock in the middle of the day. | 0.95 | 0.89 | 0.87 | 0.88 |
| *2. midday:noon* | | | | | |
| | - Midday is 12 o'clock in the middle of the day.<br>- Noon is 12 o'clock in the midle of the dey. | - | 0.67 | 0.82 | 0.75 |
| *3. midday:noon* | | | | | |
| | - Midday is 12 o'clock in the middle of the day.<br>- 12 o'clock in the midle of the dey is noon. | - | 0.67 | 0.56 | 0.62 |
| *4. cord:smile* | | | | | |
| | - Cord is strong, thick string.<br>- A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly. | 0.01 | 0.11 | 0.15 | 0.13 |
| *5. rooster:voyage* | | | | | |
| | - A rooster is an adult male chicken.<br>- A voyage is a long journey on a ship or in a spacecraft. | 0.01 | 0.24 | 0.29 | 0.27 |
| *6. gem: jewel* | | | | | |
| | - A gem is a jewel or stone that is used in jewellery.<br>- A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces. | 0.65 | 0.52 | 0.32 | 0.42 |
| *7. cock:rooster* | | | | | |
| | - A cock is an adult male chicken<br>- A rooster is an adult male chicken. | 0.86 | 0.86 | 0.83 | 0.85 |
| *8. cemetery: graveyard* | | | | | |
| | - A cemetery is a place where dead people's bodies or their ashes are buried.<br>- A graveyard is an area of land, sometimes near a church, where dead people are buried. | 0.77 | 0.36 | 0.39 | 0.38 |
| *9. automobile:wizard* | | | | | |
| | - An automobile is a car.<br>- In legends and fairy stories, a wizard is a man who has magic powers. | 0.02 | 0.25 | 0.23 | 0.24 |
| *10. mound:stove* | | | | | |
| | - A mound of something is a large rounded pile of it.<br>- A stove is a piece of equipment which provides heat, either for cooking or for heating a room. | 0.01 | 0.26 | 0.29 | 0.28 |

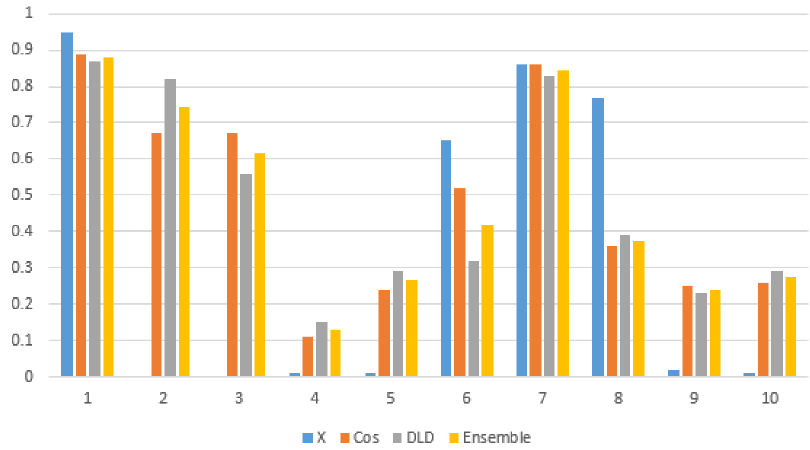


Figure 3. Comparison Similarity Value by Cosine, DLD and Proposed Model

## 6. CONCLUSION

This paper demonstrates the combination of lexical-based model shows the significant in analysing the short text. The experimental results also have shown the significant use of the proposed ensemble model in analysing the short sentences. This research is to overcome the limitations of both lexical-based models used. The experimental results also have indicated the potential use of our combination similarity model in overcome the weakness of both models by resolving the spelling error for the input query

Both techniques are only consider the similarities in lexical terms, without taking into account the semantic context of the terms in sentences. However, the advantage of both single models also influence the attaining better similarity value and show convincing results. Comparison with single similarity model indicates that the proposed model produces a marked improvement in short text similarity retrieval.

## 7. ACKNOWLEDGMENT

This research is supported by the Fundamental Research Grant Scheme (FRGS) funded by Malaysian Ministry of Higher Education, under grant number 59467.

## REFERENCES

[1] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand*, 2008, pp. 49–56.

[2] M. John Basha and K. Kaliyamurthie, "An improved similarity matching based clustering framework for short and sentence level text," *International Journal of Electrical and Computer Engineering*, vol. 7, pp. 551–558, 02 2017.

[3] Y. Y. Andreas Lianos, "Classifying unstructured text using structured training instances and an ensemble of classifiers," *Journal of Intelligent Learning Systems and Applications*, vol. 07, pp. 58–73, 2015.

[4] I. Dagan, L. Lee, and F. C. N. Pereira, "Similarity-based models of word cooccurrence probabilities," *Machine Learning*, vol. 34, no. 1, pp. 43–69, Feb 1999.

[5] Q. Ji, P. Haase, and G. Qi, *Combination of Similarity Measures in Ontology Matching Using the OWA Operator*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 281–295.

[6] D. Metzler, S. Dumais, and C. Meek, "Similarity measures for short segments of text," in *Proceedings of the 29th European Conference on IR Research*, ser. ECIR'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 16–27. [Online]. Available: http://dl.acm.org/citation.cfm?id=1763653.1763660

[7] L. Wenyin, X. Quan, M. Feng, and B. Qiu, "A short text modeling method combining semantic and statistical information," *Inf. Sci.*, vol. 180, no. 20, pp. 4031–4041, Oct. 2010.

[8] M. Sahami and T. D. Heilman, "A web-based kernel function for measuring the similarity of short text snippets," in *Proceedings of the 15th International Conference on World Wide Web*, ser. WWW '06. New

York, NY, USA: ACM, 2006, pp. 377–386.

[9] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, ser. AAAI'06. AAAI Press, 2006, pp. 775–780.

[10] M. AL-Smadi, Z. Jaradat, M. AL-Ayyoub, and Y. Jararweh, "Paraphrase identification and semantic text similarity analysis in arabic news tweets using lexical, syntactic, and semantic features," *Inf. Process. Manage.*, vol. 53, no. 3, pp. 640–652, May 2017.

[11] M. Rizzo Irfan, M. Fauzi, T. Tibyani, and N. Dyah Mentari, "Twitter sentiment analysis on 2013 curriculum using ensemble features and k-nearest neighbor," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, p. 5409, 12 2018.

[12] A. Severyn and A. Moschitti, "Learning to rank short text pairs with convolutional deep neural networks," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '15. New York, NY, USA: ACM, 2015, pp. 373–382.

[13] A. Shtok, G. Dror, Y. Maarek, and I. Szpektor, "Learning from the past: answering new questions with past answers," in *Proceedings of the 21st international conference on World Wide Web*, ser. WWW '12, 2012, pp. 759–768.

[14] Y. Gu, Z. Yang, J. Zhou, W. Qu, J. Wei, and X. Shi, "A fast approach for semantic similar short texts retrieval," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*, 2016.

[15] S. A. Noah, A. Y. Amruddin, and N. Omar, "Semantic similarity measures for malay sentences," in *Proceedings of the 10th International Conference on Asian Digital Libraries: Looking Back 10 Years and Forging New Frontiers*, ser. ICADL'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 117–126.

[16] R. Ferreira, R. D. Lins, S. J. Simske, F. Freitas, , and M. Riss, "Assessing sentence similarity through lexical, syntactic and semantic analysis," *Comput. Speech Lang.*, vol. 39, no. C, pp. 1–28, Sep. 2016.

[17] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, "A semantic approach for text clustering using wordnet and lexical chains," *Expert Systems with Applications*, vol. 42, no. 4, pp. 2264–2275, 2015.

[18] P. Nakov, S. Rosenthal, S. Kiritchenko, S. M. Mohammad, Z. Kozareva, A. Ritter, V. Stoyanov, and X. Zhu, "Developing a successful semeval task in sentiment analysis of twitter and other social media texts," *Language Resources and Evaluation*, vol. 50, no. 1, pp. 35–65, Mar 2016.

[19] S. Anson, H. Watson, K. Wadhwa, and K. Metz, "Analysing social media data for disaster preparedness: Understanding the opportunities and barriers faced by humanitarian actors," *International Journal of Disaster Risk Reduction*, vol. 21, pp. 131 – 139, 2017.

[20] J. V. Lochter, R. F. Zanetti, D. Reller, and T. A. Almeida, "Short text opinion detection using ensemble of classifiers and semantic indexing," *Expert Systems with Applications*, vol. 62, pp. 243 – 249, 2016.

[21] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, no. 13, pp. 13–18, April 2013.

[22] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, "Semantic similarity from natural language and ontology analysis," *CoRR*, vol. abs/1704.05295, 2017.

[23] K. Abdalgader and A. Skabar, "Short-text similarity measurement using word sense disambiguation and synonym expansion," in *AI 2010: Advances in Artificial Intelligence*, J. Li, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 435–444.

[24] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine Learning*, vol. 40, no. 2, pp. 139–157, Aug 2000.

[25] F. Šarić, G. Glavaš, M. Karan, J. Šnajder, and B. D. Bašić, "Takelab: Systems for measuring semantic text similarity," in *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, ser. SemEval '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 441–448.

[26] S. A. Noah, N. Omar, and A. Y. Amruddin, "Evaluation of lexical-based approaches to the semantic similarity of malay sentences," *Journal of Quantitative Linguistics*, vol. 22, no. 2, pp. 135–156, April 2015 2015.

[27] Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *IEEE Trans. on Knowl. and Data Eng.*, vol. 18, no. 8, pp. 1138–1150, Aug. 2006.

[28] M. A. Sultan, S. Bethard, and T. Sumner, "Dlscu at semeval-2016 task 1: Supervised models of sentence similarity," in *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEvalNAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, 2016, pp. 650–655.

[29] A. A. Hasan, S. Tiun, M. M. Yusof, U. A. Mokhtar, and D. I. Jambari, "Enhanced feature for short document classification," *Journal of Engineering and Applied Sciences*, vol. 12, no. 13, pp. 3534–3540, 2017.

[30] C. D. Boom, S. V. Canneyt, S. Bohez, T. Demeester, and B. Dhoedt, "Learning semantic similarity for very short texts," *CoRR*, vol. abs/1512.00765, 2015.

[31] R. Ferreira, F. Cavalcanti, George D.C. Freitas, R. D. Lins, S. J. Simske, , and M. Riss, "Combining sentence similarities measures to identify paraphrases," *Comput. Speech Lang.*, vol. 47, pp. 59–73, Jan. 2018.

[32] P. H. Duong and N.-T. Nguyen, Hien T.and Huynh, "Measuring similarity for short texts on social media," in *Proceeding of the 5th International Conference on Computational Social Networks, CSoNet 2016, Ho Chi Minh City, Vietnam, August 2-4, 2016*. Springer International Publishing, 2016, pp. 249–259.

[33] J. Oliva, J. I. Serrano, M. D. del Castillo, and A. Iglesias, "Symss: A syntax-based measure for short-text semantic similarity," *Data Knowl. Eng.*, vol. 70, no. 4, pp. 390–405, Apr. 2011.

[34] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, no. 13, pp. 13–18, April 2013.

[35] S. Hassan and R. Mihalcea, "Semantic relatedness using salient semantic analysis," in *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, ser. AAAI'11. AAAI Press, 2011, pp. 884–889. [Online]. Available: http://dl.acm.org/citation.cfm?id=2900423.2900564

[36] G. Tsatsaronis, I. Varlamis, and M. Vazirgiannis, "Text relatedness based on a word thesaurus," *J. Artif. Int. Res.*, vol. 37, no. 1, pp. 1–40, Jan. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1861751.1861752

[37] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Commun. ACM*, vol. 7, no. 3, pp. 171–176, Mar. 1964.

[38] A. Husin and K. Ruhana Ku-Mahamud, "Ant system and weighted voting method for multiple classifier systems," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, p. 4705, 12 2018.

[39] M. Whitehead and L. Yaeger, *Sentiment Mining Using Ensemble Classification Models*. Dordrecht: Springer Netherlands, 2010, pp. 509–514.


## BIOGRAPHY OF AUTHORS


**Arifah Che Alhadi** is currently a senior lecturer at School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu (UMT). She received her Master degree in Information Technology from National University of Malaysia in 2005. Her recent research work focuses on Information Retrieval especially in the short text analysis and retrieval. Besides that, she also involved in the research on ontology development and sentiment analysis. She is actively conducting her research in her field of interests through the supervision of undergraduate students. She is also actively involved as paper reviewers for national and international journals, conferences, seminars and symposiums in her field.

**Aziz Deraman** received his Bachelor from UKM in 1982, Master from Glasgow University in 1984 and PhD from UMIST in 1992 and has served Universiti Kebangsaan Malaysia from 1984 to 2007 and Universiti Malaysia Terengganu since 2007. He is presently a senior professor of Software Engineering specialising in software process, software management and certification. He maintains a diverse research interest including IT strategic planning, Software process and certification, reusability in multimedia object, medical computing, smart education management and community computing.

**Masita Masila Abdul Jalil** received her B.Eng. (Hons) degree in Computer System Engineering from University of Warwick, UK in 1997. Upon graduation, she joined Celcom (M) Sdn Bhd, a mobile service provider, as System Engineer. She later pursued her study in engineering business management at University of Warwick and obtained her Master's degree in 2000. In 2013, she received her Ph.D degree in Information System from Universiti Kebangsaan Malaysia (UKM). She is currently a senior lecturer at the School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu. Her research interests include Information System, software reuse and computer science education.

**Wan Nural Jawahir Wan Yussof** received her B.IT in Software Engineering and M.Sc. in Artificial Intelligence from Kolej Universiti Sains dan Teknologi Malaysia. In 2014, she obtained her Ph.D. from Universiti Malaysia Terengganu. She is currently a senior lecturer at School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu. Her research interests are in 2D/3D image analysis and underwater video processing.

**Rosmayati Mohemad** is currently a senior lecturer at School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu (UMT). She received her PhD degree in Computer Science from National University of Malaysia in 2013. Her research interests are in knowledge engineering and decision support system and currently focusing on modelling ontology for supporting decision-making process in the domains of special education and forensics science. She is an editor of books, conference proceedings and also reviewer of international journals and conference papers. She is a profesional technologist of Malaysian Board of Technologies and also a member of IEEE society.