https://doi.org/10.17576/apjitm-2020-0902-01
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 9 No. 2 December 2020: 1 - 10
e-ISSN: 2289-2192

Received: 24 April 2020
Accepted: 15 September2020
Published: 1 December 2020

QUANTIFYING SEMANTIC SHIFT VISUALLY ON A MALAY DOMAIN-SPECIFIC CORPUS USING TEMPORAL WORD EMBEDDING APPROACH

SABRINA TIUN
SAIDAH SAAD
NOR FARIZA MOHD NOOR
AZHAR JALALUDIN
ANIS NADIAH CHE ABDUL RAHMAN

ABSTRACT

In this study, we propose an alternative approach to analyzing a domain-specific time series corpus for detecting word evolution. The method trains a target corpus in time series into a temporal word embedding (TWE) model. The advantage of TWE is that one can see how the meaning of a word changes over time. We have chosen the TWEC approach to model a Malay domain-specific time-series corpus, the Malaysian Hansard Corpus (MHC), to a TWE model and called the model as MHC-TWEC. Two primary analyses, i.e., self-similarity analysis and user-defined method analysis, were performed to validate the effectiveness of the MHC-TWEC model in quantifying semantic shift on MHC visually. From those analyses, we visually find out that the TWE model can capture the semantic shift in the temporal corpus (the MHC).

*Keywords**:** temporal word embedding, temporal corpus, Malaysian Hansard Corpus*

INTRODUCTION

One of the latest popular approaches in the world of natural language processing is the so-called word embedding (WE). The word embedding (WE) model is a neural network based on the distributional semantic model. The distributional hypothesis states that semantically similar words tend to have similar contextual distributions (Desagulier, 2019). In the WE context, if two words have similar vectors, then they have the same distribution. An application of WE base on periodical is called temporal word embedding (TWE) or also known as dynamic word embedding (DWE). Different from WE, TWE is to associate different vectors to a word at different times (Di Carlo, 2019). Since the TWE model is built over time, an analysis by comparing the same words or related words across time is possible, which is not possible in the WE model. In addition, we can also see the semantic shift of a word based on the close neighbouring words around the target word over time (Di Carlo, 2019). The close neighbouring words define the meaning of a word, and the changing surrounding word of the target word over time will give the idea of how the meaning of the target word has changed.

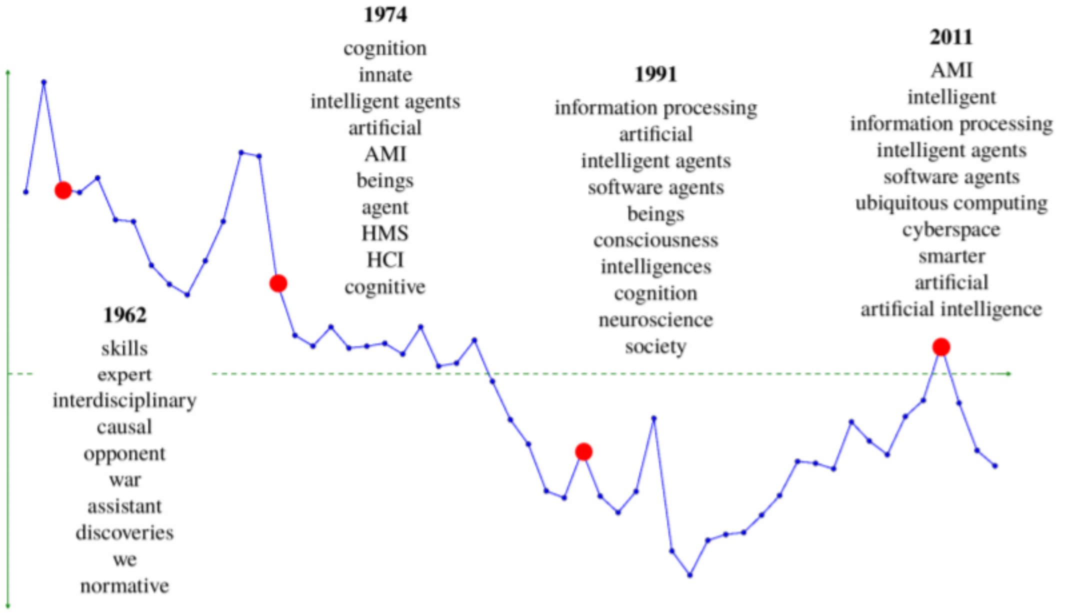
The applications of the TWE model fall under two circumstances: For a TWE with a time-series corpus of a more extended period (or also known as a diachronic corpus), usually the related research concerns on language topics, i.e., the study on the meaning of words over time. For a time-series corpus with a shorter period (called temporal corpus), it is more on mining texts for culture semantic shift, i.e., detecting event for an actionable purpose (Kutuzou et al., 2018).

In this paper, we want to see the word used in the Malaysian parliament over time using the TWE approach. By proving TWE can capture the semantic shift of words used in Malaysian parliament over time, we offer alternative approaches for researchers from other related fields such as digital humanities or political science to analyze the parliamentary debates. Thus, following the same equation of TWE by Del Coco (2018), our TWE model is formulated as, that given $w = TWE_t$ (i), where the vector of word w, is in the period of t, t = 1, …, n. In a sense that, given a word "ancaman" (threat) that is referring to a different of perspective across time. It is interpreted that the "ancaman" (threat) faced by a government over time is different. For example, the "ancaman" during the early years is more on physical war, such as fighting over communism. However, as the years go on, the threat to the government is more on psychological war, such as drugs, fake news, and misinformation. Thus, by comparing the result given by the TWE model across the different time periods, we can see what and how the meaning of "ancaman" (threat*)* has changed/shifted over time. As we mentioned previously, the modeling of TWE is based on time-series text corpora (Di Carlo, 2019), which is a corpus that has been temporally partitioned. Therefore, in this study, we choose the Malaysian Hansard Corpus (MHC) (Nor Fariza et al., 2019; Imran et al. 2017). MHC is a temporal and domain-specific corpus. By applying TWE on MHC, we then name our TWE model as MHC-TWEC. The corpus is partitioned temporally based on sessions of the parliamentary debates that have been arranged into 13 sessions. In detail, as mentioned in Nor Fariza et al. (2019), the MHC contains parliamentary proceedings from P1(1959) until P13 (2018) (see Nor Fariza et al. (2019) for more detail). In this study, by modeling the MHC onto TWE, we want to find out the answers to a question of whether TWEC model can capture the semantic changes in a domain-specific temporal corpus (e.g., MHC corpus)? As to answer for the question, we will perform two types of TWE analysis; the self-similarity analysis (Del Coco, 2018) and user-defined method analysis (Boudih, 2018).

For the rest of this paper, we will present some related works in the RELATED WORKS section, followed by the research method and dataset used in RESEARCH METHOD section. In the section RESULTS AND DISCUSSION, we will then discuss the analysis result and provide some discussions on the obtained results. Finally, we conclude our study in the CONCLUSION section.

RELATED WORKS

As mentioned previously, an increasingly popular of natural language processing approach on mining information based on time series events is called Temporal Word Embedding (TWE). TWE captures the evolution of words throughout periods. In other words, by applying TWE, we can analyze how the trends of words evolve. Pointed out by Almasian (2019), the shifted word semantic due to the changes in the attitudes of speakers or the general environment of the speakers causes word evolution. Such as in the work of Ruldoph and Blei (2018), they use TWE to capture what are words highly related to the target word 'intelligence' on a temporal basis on the ACM journal abstract from 1951 to 2014.



(a) INTELLIGENCE in ACM abstracts (1951–2014)

FIGURE 1. Words that are highly related to the word 'intelligence' extracted from TWE of a corpus on ACM abstract from 1951 to 2014 (Ruldoph and Blei, 2018)

Most of the early works of TWE (Hamilton et al., 2016; Kulkarni et al., 2015; Zhang et al., 2016), they train time-sliced embedding separately and independently. Due to that, the dimensions of the time-sliced embeddings are not comparable across time, and a method such as ad-hoc alignment techniques (Hamilton et al., 2016; Kulkarni et al., 2015; Zhang et al., 2016) are required to glue the embeddings together. A few better approaches to overcome the ad hoc alignment is by employing the matrix factorization as the perspective of embeddings (Levy et al., 2014) or exponential family embeddings as the general embedding (Ruldoph and Blei, 2018). The recent and more straightforward approach in TWE is called the Temporal Word Embedding Compass (TWEC) proposed by Di Carlo et al., (2019). The following describes the TWEC model in detail.

TWEC MODEL

The TWE or the Temporal Word Embedding Compass has been proposed by Di Carlo et al., (2019). Di Carlo et al. (2019) uses the atemporal compass to ensure the time-sliced embeddings are comparable across time. In TWEC (Di Carlo et al., 2019), two kinds of WE trainings are required. First, the training of the whole corpus, D, to build an atemporal embedding compass (see Figure 2). The atemporal embedding compass freezes all the vectors in D. Second, training the time-sliced corpus independently with the frozen pre-trained atemporal embedding. The vectors in the frozen pre-training embedding will align all the vectors in C^t across time. TWEC trains corpora based on the Word2vec model and use CBOW as the hyperparameter setting.

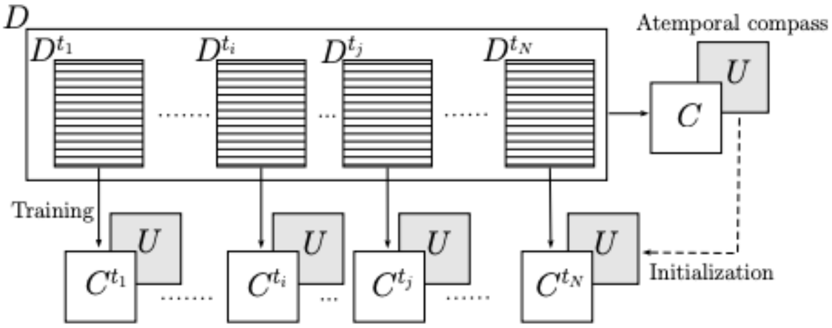


FIGURE 2. The TWEC model where the temporal context embeddings C^t are independently trained with pre-trained atemporal target embeddings U (Di Carlo, 2019)

TWEC has been evaluated based on many case studies and has outperformed comparatively for both model accuracy and analogies (Di Carlo et al., 2019). TWEC model has been proven relatively easy to use, and its source code can downloadable at https://github.com/valedica/twec). Therefore, in this study, we will use the TWEC model on MHC.

RESEARCH METHOD

Based on the two main factors; easy to be used and outperform than any previous TWE models, this study uses the TWEC model on MHC. Following the work of Sabrina et al. (2020) (since we will use their WE model to answer our RQ-1), we use the same corpus, the MHC, to be trained as the TWE model. MHC consists of 13 small corpora based on the allocated parliamentary sessions (Nor Fariza et al., 2019). However, training the MHC for the TWE model, we only consider monolingual texts (Malay), which is from P3 (1971-1973) until P13 (2013-2018). We ignore P1(1959-1964) and P2(1965-1970) since both debates are written in a mixture of English and Malay languages.

TABLE 1. The statistic description of Malaysian Hansard Corpus (MHC) (Nor Fariza et al., 2019)

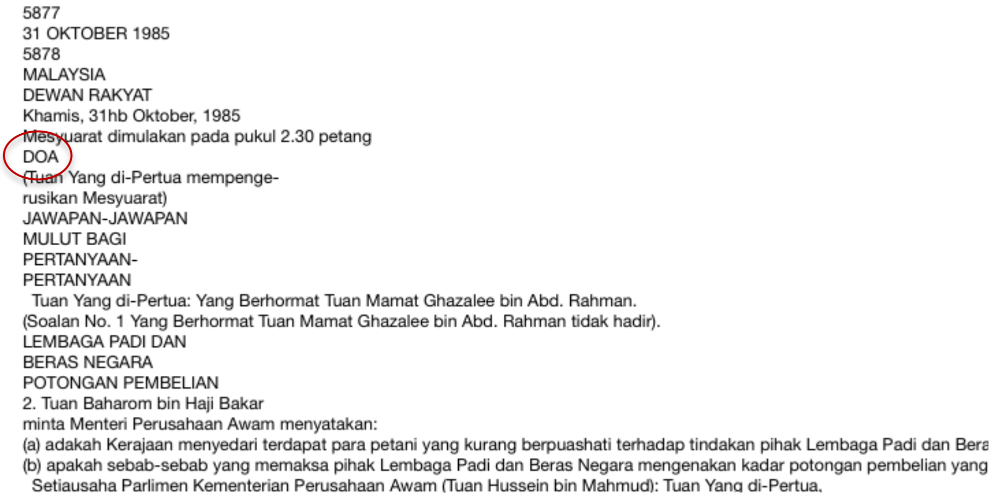
Parliment	Duration (years)	Size (word token)
1	1959 – 1964	6060551
2	1964 - 1969	9893721
3	1971 – 1973	6264859
4	1974 - 1978	8040934
5	1978 - 1981	8691728
6	1978 - 1981	9485250
7	1986 -1990	9106187
8	1990 - 1994	15171864
9	1990 - 1994	12919341
10	1999 - 2003	14123916
11	2004 - 2007	17047556
12	2004 - 2007	22188820
13	2013 - 2018	18517944
	TOTAL size (word token)	157512671

To perform semantic quantifying analysis visually on MHC, it requires three stages of processing:

1. Cleaning and pre-processing the MHC and;
2. building the MHC-TWEC model and;
3. building the semantic evaluation analysis tool (see Figure 4).

The detail on involved processings for each stage will be explained in the following:

Stage 1: Clean and pre-process MHC: The MHC pre-processing has gone through three main phases for this study: (1) Text cleaning and (2) text normalization and (3) stop word removal. In the text cleaning, symbols, characters and numbers were removed since they were considered as insignificant features. As mentioned by Sabrina et al. (2020), the content of the corpus was only considered after the keyword “DOA” or “DO’A” (see Figure 3) due to a parliament debate that will only begin after a prayer has been recited. For text normalization, we only change any capital letters into lowercase letters. In stop word removal, we use the Malay same stop word list used by Sabrina et al. (2020).



```
5877
31 OKTOBER 1985
5878
MALAYSIA
DEWAN RAKYAT
Khamis, 31hb Oktober, 1985
Mesyuarat dimulakan pada pukul 2.30 petang
DOA
(Tuan Yang di-Pertua mempenge-
rusikan Mesyuarat)
JAWAPAN-JAWAPAN
MULUT BAGI
PERTANYAAN-
PERTANYAAN
  Tuan Yang di-Pertua: Yang Berhormat Tuan Mamat Ghazalee bin Abd. Rahman.
(Soalan No. 1 Yang Berhormat Tuan Mamat Ghazalee bin Abd. Rahman tidak hadir).
LEMBAGA PADI DAN
BERAS NEGARA
POTONGAN PEMBELIAN
2. Tuan Baharom bin Haji Bakar
minta Menteri Perusahaan Awam menyatakan:
(a) adakah Kerajaan menyedari terdapat para petani yang kurang berpuashati terhadap tindakan pihak Lembaga Padi dan Bera
(b) apakah sebab-sebab yang memaksa pihak Lembaga Padi dan Beras Negara mengenakan kadar potongan pembelian yang
  Setiausaha Parlimen Kementerian Perusahaan Awam (Tuan Hussein bin Mahmud): Tuan Yang di-Pertua,
```

FIGURE 3: An Excerpt of MHC file Containing the Word ‘DOA.’

The output of Stage 1 is a clean dataset that will be used as an input to perform the TWEC training (Di Carlo, 2019), Stage 2 in Figure 4.

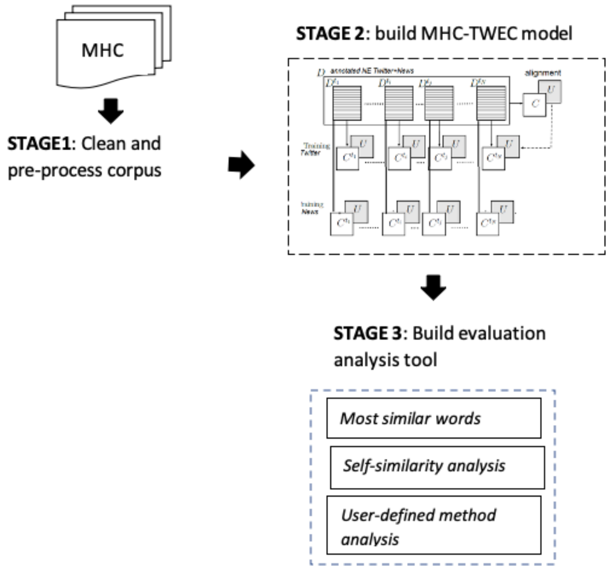

MHC
STAGE1: Clean and pre-process corpus
STAGE 2: build MHC-TWEC model
alignment
STAGE 3: Build evaluation analysis tool
Most similar words
Self-similarity analysis
User-defined method analysis

FIGURE 4. The processing stages of MHC-TWEC model for semantic analysis

Stage 2: Build the MHC-TWEC model. The TWEC approach requires two types of corpora, the whole corpus and the time-series corpus for training. At this stage, we train the TWE model for the entire corpus and time-sliced corpus using the TWEC approach (Di Carlo, 2019). To build MHC-TWEC model, the whole corpus, which is D, will be the combination of P3 until P13 corpus and trained as U, the atemporal embedding compass (see Figure 2). The time-sliced corpus is an individual corpus of P3 until P13, which in TWEC model, will be C^t … C^N (see Figure 2). Since TWEC model is built based on Word2vec on the CBOW setting, both our whole and time-sliced corpora are trained in the same setting (Word2Vec, CBOW).

Stage 3: Build evaluation analysis: At the stage, the visual evaluation will be carried out by performing two kinds of similarity analysis. Once the corpus has been trained into the TWEC model, the TWEC model will be used to perform visual analyses of Del Coco (2018) and Boudiah (2018).

1. *Self-similarity:* Following the work of Del Coco (2018), we use the same self-similarity equation as in Equation 1. The reason for performing this analysis is to quantify the change in word meaning visually across different periods where the vectors from the same word of different vector space are computed. To be specific, the self-similarity, T, of a given word vector in TWE time *i*, TWE_t (*i*), to the word vector in TWE time *i*-1, TWE_t (*i*-1), is computed as (Del Coco, 2018):

$$\text{self-similarity } (T_t(i)) = \frac{TWE_i(t).TWE_{i-1}(t)}{|TWE_i(t)|.|TWE_{i-1}(t)|} \qquad (1)$$

Since self-similarity analysis is to quantify semantic change from one period of time to another period, thus logically, it can only be performed on time-series data. Therefore, it is only applicable to our MHC-TWEC model. Using this analysis, we can see how word evolves in MHC using the MHC-TWEC model. In other words, we can find out if there is a semantic shift captured using the MHC-TWEC model, visually.

2. *User-defined method analysis.* In the study of Boudih (2018), the user-defined method analysis is aimed to visual a semantic similarity over time between two words in a two-dimensional graph. They suggested placing the variable period (temporal) on the *x*-axis, while the cosine similarity of a pair of two words vectors is placed on the *y*-axis. This visualization method shows the semantic relation strength of word pairs over time. The equation that is used to calculate the semantic similarity of the word pair (word reference and word target) is based on a cosine similarity (see Equation 2). In Equation 2 or we named it as 'pair-similarity', given a word reference *x*, and a target word, *y*, the similarity strength of the pair words on a specific time, *t*, is:

$$\text{pair-similarity } (x_t, y_t) = \frac{x_t \cdot y_t}{|x_t| . |y_t|} \qquad (2)$$

In the following, we perform and discuss the two visual analyses: Self-similarities analysis, and user-defined method analysis.

RESULTS AND DISCUSSION

In this section, we will show the results of the two analyses: Self-similarity analysis and user-defined method analysis in visually quantifying the semantic of MHC. Besides, we also provide the readers with some discussions on what can be interpreted with the results.

ANALYSIS I: SELF-SIMILARITY ANALYSIS

Self-similarity analysis is a method that can quantify semantics visually across time by comparing vectors of the same word in different vector spaces (Del Coco, 2018). In this analysis, we can see how strong the concept is projected over time in a particular corpus. Figure 5 shows the time-series for the word "keselamatan" (security) from the year 1971 (P3) until the year 2018 (P13) from MHC. With near-maximum self-similarity on almost all the years, we are suggesting that the concept of the word "keselamatan" (security) is strong in MHC. We can also see the selected most related word for each period moving from the word "antarabangsa", "Indonesia", "singapura" towards the word "awam", "kakitangan", "polis". This trend of the selected most related words of target word 'keselamatan' (*security*) is suggesting the semantic shift from security on external (international) affairs to security on local (domestic) affairs.

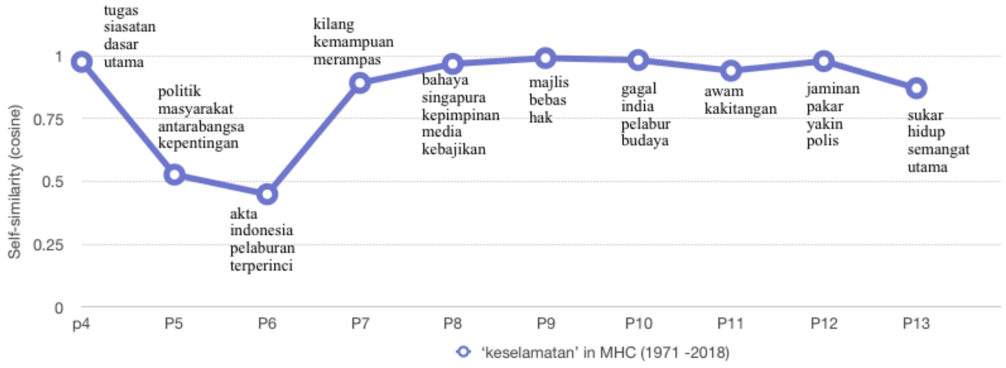


FIGURE 5. The self-similarity of the word 'keselamatan' (*security*) generated from the MHC-TWEC model.

Although the semantic shift is not apparent in the sense that the word "keselamatan" (security) does not have a substantial difference in meaning. In this analysis, the concern probably on what kind of issues are related to the word concept "keselamatan" (security). As mentioned by Sabrina et al. (2020), MHC is a domain-specific corpus. Thus, the sense or the meaning of a word throughout the corpus will be the same. The only difference will be regarded as to what words are related to the meaning of the word "keselamatan" (security). In another perspective, although this analysis does not show a huge leap or shift of semantic words, a semantic shift is evident. The model captures a shift (based on the clustered words) on how the concept "keselamatan"(security) moves from concerning international affairs to a domestic affair. This capture shows that the TWE approach is capable of quantifying the semantic shift on a domain-specific corpus. To conclude, looking at the trend, which is the shape of line in Figure 5, and the word clustered at each period, MHC-TWEC can visually quantify semantic in MHC.

ANALYSIS II: USER-DEFINED METHOD ANALYSIS

In Figure 6, the graph represents a comparison over time between two words. As defined by Boudih (2018), a high cosine similarity (value of each period is near to 1) indicates a strong semantic association between two words in a specific period. In this analysis, based on the MHC-TWEC model, we choose the word "keselamatan" (security) as a reference vector on "maklumat" (information) and "media"(media) vectors.

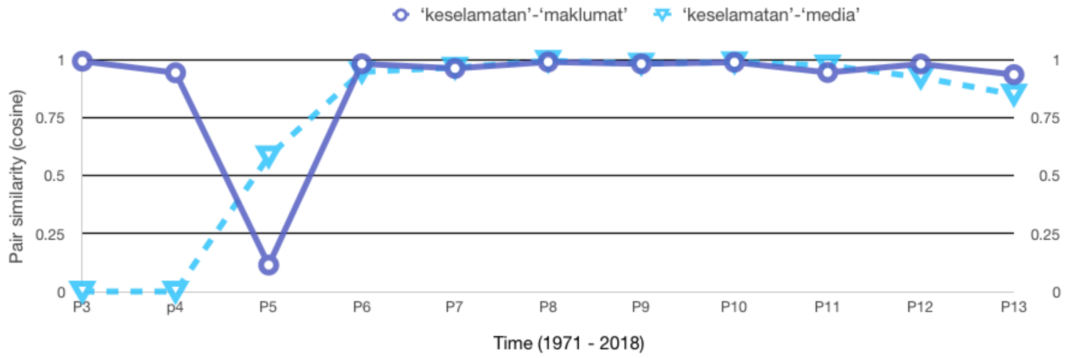


FIGURE 6. The trend of a word concept 'keselamatan' (security) –'maklumat'(*information*) and word concept 'keselamatan' (*security*) –'media'(*media*) in MHC in the year of 1971 to 2018.

From Figure 6, MHC-TWEC captures a relatively strong semantic relatedness both pairs of words of "keselamatan"(security) –"maklumat" (information) and "keselamatan"(security) – "media"(media). Figure 4 can be interpreted that both pairs of words co-occurred for a significant amount of time; therefore, both pairs of concepts are highly semantically similar. However, for the word pair "keselamatan"(security) – "media"(media) since the pair similarity value is at 0; this indicates that there is no co-occurrence at P3 (1971-1973) and P4 (1974-1978). Our guess is probably; the word "media" is not widely used before P5 (1978-1981). Another noticeable trend is, in Figure 6, there is a deep dive of word pair "keselamatan"(security) – "maklumat"(information). The deep dive could indicate the concept "keselamatan"(security) related to "maklumat"(information) was less debated in the Malaysian parliament during that specific period compared to other periods. The trend on how a pair word evolves over the years via this analysis proves that the TWE model can visually quantify semantics shifts on a domain-specific corpus such as MHC.

To conclude, the visual quantification of semantic shift based on TWE model presented in this paper can be used as a clue or support on detecting any trend or change information (events). Clue or support on detecting any trend or change information is a very useful ‘tool’ for a corpus-based research study, i.e. culturomics study or digital humanities, in general.

CONCLUSION

In this study, we proposed an alternative approach to analyzing a corpus for detecting word evolution. The approach trains a target corpus in time series into a temporal word embedding (TWE) model instead of word embedding (WE). The advantage of TWE is that one can see how the meaning of a word changes over time. We have chosen the TWEC approach to model the temporal corpus of MHC to the TWE model and called the model as MHC-TWEC. Two visual analyses: Self-similarity analysis and user-defined method analysis, were performed to validate the effectiveness of our MHC-TWE model in quantifying the semantic shift of a domain-specific time series corpus. From this analysis, we visually can see how effective is MHC-TWEC in quantifying semantic shift. Both the self-similarity analysis and user-defined method analysis show visually how MHC-TWEC is capable of capturing the trend of a word or a pair of words over time. To conclude, based on all of the analyses, it shows that the MHC-TWEC model is capable of quantifying a semantic shift on a specific-domain of a time-series corpus.

ACKNOWLEDGMENT

This research work is funded by Ministry of Higher Education Malaysia under research code: FRGS/1/2020/ICT02/UKM/02/1.

REFERENCES

Almasian S., Spitz A. & Gertz M. 2019. Word Embeddings for Entity-Annotated Texts. In: Azzopardi L., Stein B., Fuhr N., Mayr P., Hauff C., Hiemstra D. (eds) Advances in Information Retrieval. ECIR 2019. Lecture Notes in Computer Science, 11437(1): 307-322.

Boudih A. M. 2018. A New Way of Visualizing Semantic Similarity over Time. Master Thesis, Tilburg University: Netherlands.

Del Coco, P. 2018. Temporal Text Mining: From Frequencies to Word Embeddings. Thesis paper: Università di Bologna: Italy.

Desagulier, G. 2019 . Can word vectors help corpus linguists?, Studia Neophilologica, 91(2): 219-240.

Di Carlo, V., Bianchi, F. & Palmonari, M. 2019. Training Temporal Word Embeddings with a Compass. The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), pp. 6326-6334.

Hamilton, W. L., Leskovec, J. & Jurafsky, D. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1489–1501.

Imran Ho Abdullah, Anis Nadiah Che Abdul Rahman. & Azhar Jaludin (2017). The Malaysian Hansard Corpus. Culturomics Workshop UKM. 27 April 2017.

Levy, O. & Goldberg, Y. 2014. Neural word embedding as implicit matrix factorization. In Neural Information Processing System, pp. 2177–2185.

Nor Fariza Mohd Nor, Azhar Jalaludin, Anis Nadiah Che Abdul Rahman, Imran Ho Abdullah & Sabrina Tiun. 2019. A Corpus Driven Analysis of Representations Around the Word ‘ekonomi’ in Malaysian Hansard Corpus. GEMA Online® Journal of Language Studies, 19(4):66-95.

Kulkarni, V., Al-Rfou, R., Perozzi, B. & Skiena, S. 2015. Statistically Significant Detection of Linguistic Change. In Proceedings of the 24th International Conference on World Wide Web. ACM, pp. 625–635.

Kutuzov, A. Øvrelid, L., Szymanski, T. & Velldal, E. 2018. Diachronic Word Embeddings and Semantic Shifts: A Survey. Proceedings of the 27th International Conference on Computational Linguistics, pp. 1384-1397.

Ruldoph, M. & Blei, D. 2018. Dynamic Embeddings for Language Evolution. WWW '18: Proceedings of the 2018 World Wide Web Conference, pp. 1001-1011.

Sabrina Tiun, Nor Fariza Mohd Nor, Azhar Jalaludin & Anis Nadiah Che Abdul Rahman. 2020. Word Embedding for Small and Domain-specific Malay Corpus. In: Alfred R., Lim Y., Haviluddin H., On C. (eds) Computational Science and Technology. Lecture Notes in Electrical Engineering, 603:435-443.

Zhang, Y., Jatowt, A., Bhowmick S. S. & Tanaka, K. 2016. The Past is Not a Foreign Country: Detecting Semantically Similar Terms across Time. IEEE Transactions on Knowledge and Data Engineering, pp. 2793–2807.

Sabrina Tiun
Saidah Saad
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
sabrinatiun@ukm.edu.my, saidah@ukm.edu.my

Nor Fariza Mohd Noor
Azhar Jalaludin
Anis Nadiah Che Abdul Rahman
Faculty of Social Sciences & Humanities
Universiti Kebangsaan Malaysia
fariza@ukm.edu.my, azharj@ukm.edu.my, p87706@siswa.ukm.edu.my