MODEL PEMBENAMAN PERKATAAN DINAMIK UNTUK BERITA TWITTER BAHASA MELAYU MENGGUNAKAN PENDEKATAN PEMBENAMAN PERKATAAN TEMPORAL

Jacqueline Hii Sing Hee[1]*

Sabrina Tiun[2]

[1,2]*Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor Darul Ehsan, Malaysia*

Abstrak

Bahasa manusia berkembang dari semasa ke semasa, begitu juga dengan makna perkataan. Pembenaman perkataan dinamik (PPD) boleh mengesan evolusi perkataan dari semasa ke semasa dengan membahagikan korpus kepada tempoh masa yang berbeza dan memperolehi representasi vektor perkataan yang mempunyai konteks yang serupa dengan lebih dekat untuk setiap tempoh masa. Masalah yang perlu diatasi dalam projek ini adalah penyediaan data berita teks dari Twitter, pemilihan akaun Twitter yang sesuai dan cara yang paling sesuai untuk mengekstrak data berita teks Twitter bahasa Melayu. Selain itu, pemilihan metode yang sesuai untuk membina model PPD bagi data teks Twitter merupakan salah satu masalah yang perlu diselesaikan dalam projek ini. Data *tweet* dari akaun Agensi Berita Nasional Malaysia atau BERNAMA, iaitu akaun berita bahasa Melayu rasmi akan diekstrak untuk mengatasi masalah yang disebutkan. Di samping itu, pendekatan pembenaman perkataan temporal (PPT) telah dipilih sebagai metode yang sesuai untuk membina model PPD bagi data teks Twitter. Skop projek ini memberi fokus kepada domain COVID-19 di mana COVID-19 adalah isu popular yang masih diberi perhatian oleh warganegara Malaysia. Projek ini memperkenalkan model PPD untuk meneroka evolusi perkataan dari semasa ke semasa. Tambahan pula, model PPD domain COVID-19 digunakan untuk meneroka trend perkataan dalam berita bahasa Melayu terhadap domain COVID-19 pada platform media sosial dari semasa ke semasa. Model ini

menggunakan pendekatan *Temporal Word Embeddings with a Compass* (TWEC) untuk melatih data berita Twitter yang disasarkan berdasarkan tempoh masa. Dengan pendekatan TWEC dalam model, evolusi sesuatu perkataan dalam data berita Twitter bahasa Melayu mengikut masa yang berbeza dapat dianalisis. Hasilan projek ini dapat ditunjukkan melalui antaramuka di laman sesawang.

Kata kunci: Pembenaman Perkataan Dinamik, Pembenaman Perkataan Temporal, Korpus Twitter

Pengenalan

Bahasa merupakan adaptasi utama bagi manusia, terutamanya untuk berkomunikasi antara satu sama lain. Bahasa berkembang daripada keperluan seharian manusia. Bahasa lisan dan tulisan merupakan salah satu hasilan perkembangan bahasa. Pada era globalisasi ini, kemunculan perkataan baharu dianggap biasa malahan makna perkataan berubah dari semasa ke semasa. Sebagai contoh, perkataan "Python" hanya dikaitkan dengan ular pada masa dahulu, namun kini turut dikaitkan dengan bahasa pengaturcaraan. Untuk mengetahui perhubungan perkataan antara satu sama lain, pembenaman perkataan telah diperkenalkan.

Pembenaman perkataan atau *word embedding* (PP) digunakan untuk merepresentasi perkataan-perkataan atau dokumen dengan melibatkan bentuk vektor dan menunjukkan perkataan-perkataan yang mempunyai konteks serupa dalam kelompok masing-masing. PP adalah berdasarkan konsep hipotesis pengagihan atau *distributional hypothesis*. Hipotesis pengagihan dinyatakan sebagai persamaan pengedaran dan persamaan makna adalah berkorelasi (Sahlgren 2008). Teknik PP bermula daripada pendekatan statistik dan berkembang kepada pendekatan rangkaian neural (Yao et al. 2018). Walau bagaimanapun, pendekatan statistik dan pendekatan rangkaian neural tidak melibatkan aspek temporal. Oleh itu, pembenaman perkataan dinamik atau *dynamic word embedding* (PPD) semakin diperkenalkan kerana ia mengambil kira aspek temporal. Model PPD meningkatkan pemahaman

tentang bagaimana perkataan berkaitan antara satu sama lain dan juga evolusi sesuatu perkataan dari semasa ke semasa (Yao et al. 2018).

Salah satu pendekatan bagi PPD adalah pendekatan pembenaman perkataan temporal atau *temporal word embedding* (PPT). Pendekatan PPT mempelajari vektor yang menangkap makna perkataan dalam tempoh masa tertentu (Di Carlo, Bianchi & Palmonari 2019). Di Carlo, Bianchi dan Palmonari (2019) juga menyatakan bahawa pendekatan tersebut dapat menjejaki sebarang perubahan semantik dalam perkataan dan membolehkan untuk mencari istilah berbeza yang berkongsi makna serupa dalam tempoh masa yang berbeza.

Pada era teknologi ini, penggunaan media sosial berkembang dengan pesat terutamanya semasa dan selepas mengalami pandemik COVID-19. Semua orang dapat memuat naik apa sahaja informasi atau perasaan sendiri secara awam ke mana-mana platform media sosial. Dengan hal demikian, representasi teks menjadi semakin penting supaya dapat menganalisis perkembangan sesuatu perkataan sepanjang tempoh masa yang berbeza. Salah satu platform media sosial yang popular ialah Twitter. Pemprosesan teks pada mesej-mesej disiarkan dalam Twitter atau *tweet* membolehkan untuk mendapatkan trend perkataan terkini. Selain itu, pengekstrakan data daripada Twitter dikatakan lebih mudah kerana data tersebut adalah sumber terbuka.

Objektif projek ini adalah untuk menyediakan data berita teks Twitter bahasa Melayu, membina model PPD untuk berita Twitter bahasa Melayu menggunakan pendekatan PPT, dan menghasilkan antaramuka untuk memudahkan pengguna mengguna model PPD berita Twitter.

Skop projek ini memberi fokus kepada domain COVID-19. Hal ini demikian kerana COVID-19 adalah isu popular yang masih diberi perhatian oleh warganegara Malaysia.

Contoh nilai keusahawanan dan komersial projek ini adalah mengaplikasi model PPD dalam sistem pencadang dan serangan siber. Model PPD boleh digunakan untuk membina sistem pencadang yang berasaskan kandungan untuk padanan perniagaan antarabangsa. Gohourou et al. (2017) melatih model PP dengan artikel berita yang berkaitan dengan domain perniagaan dan menggunakan model

tersebut untuk mendapati informasi yang mungkin diperlukan oleh pengguna tanpa memperolehi pengetahuan terdahulu yang khusus daripada pengguna. Tambahan pula, model tersebut menawarkan kemungkinan baharu untuk mengekstrak maklumat daripada data teks yang besar (Gohourou et al. 2017). Model PPD dapat digunakan untuk memantau dan memahami langkah serangan siber dieksploitasi, dan membenderakan perubahan dalam cara serangan siber berlaku (Shen & Stringhini 2019). Dengan lebih memahami serangan siber dan evolusinya, pengetahuan ini dapat digunakan untuk meningkatkan kesedaran situasi siber dan membangunkan pertahanan proaktif (Shen & Stringhini 2019).

Penyediaan set data yang sesuai dalam sesuatu projek adalah proses penting untuk menjayakan pelaksanaan projek tertentu. Oleh sebab projek ini berkaitan dengan bahasa Melayu, fokus penyediaan set data adalah set data *tweet* bahasa Melayu. Beberapa kajian lepas berkaitan penyediaan set data *tweet* bahasa Melayu telah dipilih untuk kajian. Dalam kajian '*English-Malay Word Embeddings Alignment for Cross-lingual Emotion Classification with Hierarchical Attention Network*' oleh Hao dan Yan (2022), pengkaji mendapatkan subset data *tweet* daripada Malaya Documentation yang merangkumi Melayu Malaysia dan Melayu Indonesia secara rawak. Pendekatan hibrid, iaitu secara algoritma mesin dan manusia secara manual, telah digunakan untuk mengekstrak hanya *tweet* Melayu Malaysia sebagai korpus kajian tersebut. Proses pra-pemprosesan data seperti menormalkan kontraksi bahasa Melayu dan menyemakkan ejaan perkataan mengikut konteks telah dijalankan pada korpus kajian tersebut. Dalam kajian '*An Enhancement of Malay Social Media Text Normalization for Lexicon-Based Sentiment Analysis*' oleh Bakar, Idris dan Shuib (2019), pengkaji mengumpulkan data *tweet* daripada pengguna Twitter Malaysia dengan web scripting dan membahagikan kepada kategori masing-masing untuk membangunkan model penormalan teks Melayu. Model penormalan tersebut merangkumi kamus-kamus yang membolehkan penukaran perkataan hingar kepada perkataan bermakna, tokenisasi lanjutan, pengesanan token Melayu atau Inggeris, peraturan leksikal, penggantian token hingar, n-gram, dan detokenisasi. Model penormalan

tersebut telah mencapai keputusan dengan 83.55% dalam ketepatan dan 84.61% dalam *recall*. Proses pra-pemprosesan data dalam kajian-kajian lepas akan menjadi sebagai panduan untuk projek ini. Metode yang disebutkan dalam kajian-kajian lepas akan diubahsuaikan, dikurangkan atau ditambahkan metode pembersihan lain untuk menyesuaikan teks data projek ini.

Pembinaan model untuk menganalisis perkataan dalam berita Twitter bahasa Melayu adalah proses yang paling penting dalam projek ini. Projek ini melibatkan pemprosesan bahasa tabii dan PP adalah elemen penting dalam pemprosesan bahasa tabii. Beberapa kajian lepas berkaitan PP telah dipilih untuk kajian. Dalam kajian '*Word Embedding for Analogy Making*' oleh Si (2022), pengkaji membina graf pengetahuan songsang menggunakan informasi daripada Wikidata kerana kebanyakan konsep dalam Wikidata adalah unik. Kemudian, pengkaji menggunakan graf tersebut untuk mengira benam. Hasil dapatan telah menunjukkan bahawa PP dapat diguna untuk meneroka hubungan analogi dengan bantuan graf pengetahuan. Dalam kajian '*Word Embedding Evaluation for Sinhala*' oleh Lakmal et al. (2020), pengkaji telah memperkenalkan penilaian komprehensif untuk bahasa Sinhala dengan tiga jenis pendekatan PP, iaitu *word2vec*, *FastText*, dan *GloVe* pada set data *Common Crawl*. Hasil dapatan telah menunjukkan bahawa pendekatan *FastText* mendapatkan keputusan yang paling tinggi antara tiga pendekatan. pengkaji juga menerangkan bahawa tiada pendekatan PP universal yang akan sentiasa memberikan prestasi terbaik untuk setiap tugasan pemprosesan bahasa tabii. Melalui kajian-kajian lepas, model PP tidak sesuai digunakan untuk projek ini kerana korpus yang digunakan dalam kajian-kajian lepas tidak melibatkan aspek temporal. Oleh sebab data teks Twitter bahasa Melayu melibatkan aspek temporal, model PPD lebih sesuai untuk projek ini.

Pemilihan model yang sesuai untuk dibangunkan amat penting dalam pelaksanaan sesuatu projek. Yao et al. (2018) telah menyatakan bahawa model PPD dapat meneroka evolusi sesuatu perkataan mengikut perubahan masa. Oleh itu, model PPD amat sesuai untuk digunakan dalam projek ini di mana terlibat aspek temporal. Dalam kajian '*Enriching Word Embeddings with Temporal and Spatial Information*' oleh Gong, Bhat dan Viswanath (2020), pengkaji telah memperkenalkan satu

model untuk belajar representasi perkataan yang dapat menjejaki semantik mengikut perubahan masa dan lokasi. Melalui penggunaan model PPD, pengkaji menunjukkan bagaimana model PPD menangkap evolusi bahasa dari semasa ke semasa dan perubahan lokasi. Dalam kajian '*Tracking Short-Term Temporal Linguistic Dynamics to Characterize Candidate Therapeutics for COVID-19 in the CORD-19 Corpus*' oleh Powell dan Sentz (2021), pengkaji mengaplikasi PPD dalam kajian tersebut untuk menentukan kemungkinan mencari dan mengukur perubahan satu set terapeutik calon yang dikenal pasti dalam kajian guna semula dadah dengan contoh temporal korpus CORD-19 dari semasa ke semasa. Namun, hasil dapatan mendapati ada yang mempamerkan kelemahan atau perubahan hubungan semantik di mana kemungkinan dipengaruhi oleh penerbitan kajian baru yang memberi kesan positif atau negatif kepada pertimbangan terapeutik calon sebagai rawatan untuk COVID-19. Melalui kajian-kajian lepas, model PPD biasanya digunakan apabila melibatkan aspek temporal. Oleh itu, model PPD sesuai untuk digunakan dalam pembinaan model projek ini.

Pemilihan metode yang sesuai untuk pembangunan model adalah penting untuk melaksanakan projek dengan lancar. Dalam projek ini, pendekatan PPT telah dipilih untuk membina model PPD. Beberapa kajian lepas berkaitan pendekatan ini telah dipilih untuk kajian. Dalam kajian '*Temporal Word Embeddings for Narrative Understanding*' oleh Volpetti, Vani dan Antonucci (2020), pengkaji menggunakan pendekatan PPT dalam pengenalpastian peranan watak daripada buku *Harry Potter* oleh *J.K. Rowling*, dan evolusinya mengikut masa yang berubah-ubah secara automatik. Hasil dapatan telah menunjukan bahawa PPT dapat mencapai objektif kajian pengkaji dengan memerhati pergerakan dan kedudukan dalam ruang vektor. Dalam kajian '*ATTACK2VEC: Leveraging Temporal Word Embeddings to Understand the Evolution of Cyberattacks*' oleh Shen dan Stringhini (2019), pengkaji telah menunjukkan penggunaan PPT dalam memodelkan konteks langkah serangan siber dan menjejaki evolusinya. Oleh sebab serangan siber selalu berubah berdasarkan masa, PPT yang melibatkan aspek temporal adalah metode sesuai untuk kajian tersebut. Melalui penggunaan PPT, ATTACK2VEC juga menunjukkan berkesannya dalam membenderakan perubahan dalam cara

serangan siber berlaku. Melalui kajian-kajian lepas, pendekatan PPT sering digunakan terutamanya semasa melibatkan aspek temporal. Oleh itu, pendekatan PPT akan digunakan dalam projek ini untuk membina model PPD pada data teks Twitter bahasa Melayu.

Tiga kajian lepas telah dipilih untuk bincangkan secara kritis. Dalam kajian '*Quantifying semantic shift visually on a Malay domain-specific corpus using temporal word embedding approach*' oleh Tiun et al. (2020), pengkaji telah menggunakan pendekatan PPT untuk melihat perkataan yang digunakan dalam parlimen Malaysia mengikut perubahan masa. Korpus kajian tersebut adalah daripada *Malaysian Hansard Corpus* (MHC) di mana korpus kajian tersebut adalah temporal dan domain spesifik. Analisis *self-similarity* dan analisis metode *user-defined* digunakan untuk mengukur keberkesanan model kajian tersebut yang dipanggil sebagai model MHC-TWEC. Analisis *self-similarity* adalah untuk mengukur perubahan maksud perkataan dari semasa ke semasa secara visual di mana melibatkan pengiraan vektor daripada perkataan yang sama bagi ruang vektor yang berbeza dan alat analisis tersebut hanya sesuai untuk dilakukan pada data siri masa secara logik. Analisis metode *user-defined* digunakan untuk bayangkan persamaan semantik antara dua perkataan mengikut perubahan masa yang berdasarkan persamaan kosinus dalam graf 2-dimensi. Dapatan kajian tersebut menunjukkan keberkesanan model MHC-TWEC dalam mengukur anjakan semantik secara visual dan model MHC-TWEC mampu mengukur anjakan semantik pada domain korpus yang melibatkan siri masa. Kajian tersebut menjadi panduan untuk projek ini dan boleh diubahsuai untuk mengaplikasi dan menganalisis pada korpus domain media sosial bahasa Melayu. Metode analisis model yang digunakan dalam kajian tersebut juga sesuai untuk mengukur keberkesanan model projek ini kerana korpus projek ini juga melibatkan data siri masa. Dalam kajian '*Visualizing Trends of Key Roles in News Articles*' oleh Xia et al. (2019), pengkaji telah menggunakan peranan semantik dan pembenaman perkataan untuk menganalisis hubungan peranan utama dalam berita merentasi tempoh masa yang berbeza dan membayangkan trend peranan utama dan perubahan topik berita mengikut peredaran masa. Korpus kajian tersebut adalah melibatkan *Trump dataset* yang merangkumi tajuk

berita dalam bahasa Inggeris dan *Newsroom dataset* yang merangkumi artikel berita bahasa Inggeris. Pengkaji membuat visualisasi hutan menggunakan peranan semantik di mana setiap aktiviti peranan utama diwakili oleh satu pokok. Pengkaji menggunakan *word2vec* dan penjajaran *Procrustes* ortogonal untuk menjejaki perubahan trend berita setiap bulan. Dapatan kajian tersebut menunjukan sistem yang dibangunkan dapat menjejaki tindakan dan berita tergempar, dan dapat perkataan bermakna yang melibatkan perubahan paling banyak dapat dikesan. Sistem pengkaji melibatkan pendekatan *word2vec* dan penjajaran *Procrustes* ortogonal dalam model PPD. Penjajaran *Procrustes* ortogonal melibatkan prosedur penjajaran yang kompleks di mana mungkin akan menyebabkan kesilapan dalam proses (Jun 2021). Satu keterbatasan penjajaran *Procrustes* ortogonal adalah metode ini memerlukan penjajaran ruang benam dengan persimpangan kosa kota di antara semua ruang benam. Keterbatasan ini menyebabkan perkataan baru yang muncul pada masa kemudian tidak dapat dibandingkan (Jun 2021). Satu cadangan untuk menyelesaikan masalah tersebut adalah menggunakan pendekatan *Temporal Word Embeddings with a Compass* (TWEC) dalam sistem tersebut untuk menjejaki trend peranan utama dalam artikel berita dengan lebih berkesan. Dalam kajian '*A Novel Method of Extracting Topological Features from Word Embedding*' oleh Gholizadeh, Seyeditabari & Zadrozny (2020), pengkaji telah mencadangkan kaedah baru untuk menggunakan homologi berterusan bagi mengekstrak ciri topologi daripada representasi PP dokumen teks dan mengaplikasi ciri topologi tersebut dalam klasifikasi teks. Korpus kajian tersebut adalah kajian daripada arXiv dan ulasan filem daripada IMDB. Homologi berterusan biasanya digunakan dalam analisis data topologi untuk mentafsir ruang benam untuk setiap dokumen teks. Dalam kajian tersebut, tiga pendekatan PP telah digunakan dan pendekatan *ConceptNet Numberbatch* telah menunjukkan keputusan yang terbaik di antara ketiga-tiga pendekatan. Dapatan kajian ini menunjukkan penggunaan ciri topologi terutamanya semasa dokumen teks yang panjang dapat meningkatkan keputusan kajian tersebut dengan penggunaan homologi berterusan dan pembenaman perkataan. Dalam kajian tersebut, pengkaji menganalisis perbezaan dimensi pembenaman dalam siri masa. Namun, pendekatan yang

digunakan dalam kajian tersebut tidak mengambil kira aspek siri masa atau temporal. Oleh itu, satu cadangan untuk kajian tersebut adalah menggunakan model PPD seperti pendekatan PPT untuk memaksimumkan keberkesanan pembenaman perkataan dalam kajian tersebut.

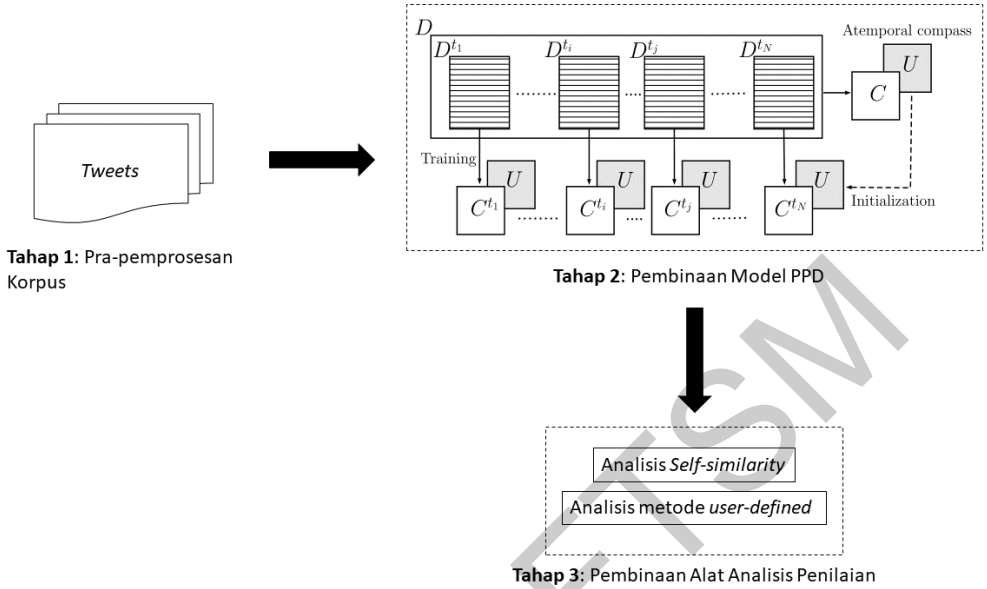
Dalam projek ini, model PPD digunakan untuk meneroka evolusi perkataan sepanjang tempoh masa yang berbeza menggunakan data Twitter. Selain itu, model PPD domain COVID-19 digunakan untuk meneroka trend perkataan dalam berita bahasa Melayu terhadap domain COVID-19 pada platform media sosial mengikut perubahan masa. Model ini menggunakan pendekatan PPT untuk melatih data Twitter. Seperti yang dinyatakan sebelum ini, pendekatan tersebut membolehkan untuk menjejaki istilah berbeza yang mempunyai maksud yang seerti sepanjang tempoh masa yang berbeza. Selain itu, data yang digunakan hanya terhad kepada berita Twitter dalam bahasa Melayu. Akhir sekali, hasilan projek ini dapat ditunjukkan melalui antaramuka di laman sesawang.

Laporan teknik ini merangkumi seperti berikut: metodologi kajian menerangkan tentang pembinaan model PPD bagi berita Twitter bahasa Melayu. Keputusan dan perbincangan adalah menjelaskan hasilan projek ini. Akhir sekali, kesimpulan telah dibuat.

Metodologi Kajian

Korpus projek ini adalah set data yang dikumpulkan daripada akaun berita Twitter bahasa Melayu rasmi Malaysia dari Mac 2020 hingga September 2021. Korpus tersebut dikumpulkan mengikut empat kategori, iaitu Ekonomi (*Economy*), Kesihatan (*Health*), Keselamatan (*Safety*) dan Pendidikan (*Education*). Korpus tersebut akan menjalani pra-pemprosesan untuk membersihkan data teks.

Projek ini menggunakan pendekatan *Temporal Word Embeddings with a Compass* (TWEC) pada korpus berita Twitter bahasa Melayu yang merangkumi data teks Twitter sepanjang 19 bulan di mana dibahagikan kepada 19 korpus kecil. Pembinaan model dalam projek ini melibatkan tiga tahap, iaitu pra-pemprosesan korpus, pembinaan model PPD dan pembinaan alat analisis penilaian. Tahap-tahap bagi pembinaan model dalam projek ini adalah seperti yang ditunjukkan dalam Rajah 1.



Rajah 1 Rajah pembinaan model bagi model PPD

Pra-pemprosesan korpus akan menjalani tiga fasa, iaitu pembersihan teks, penormalan teks, dan penyingkiran kata henti. Dalam fasa pembersihan teks, simbol, nombor dan emoji akan disingkirkan kerana ia tidak menunjukkan ciri penting. Dalam fasa penormalan teks, penukaran huruf besar kepada huruf kecil akan dilakukan. Dalam fasa penyingkiran kata henti, senarai kata henti bahasa Melayu akan diperoleh daripada sumber dalam talian dan digunakan pada korpus Twitter.

Terdapat dua jenis korpus terlibat dalam latihan pendekatan TWEC, iaitu korpus diakronik secara keseluruhan dan korpus siri masa. Pada tahap ini, model PPD dilatih untuk kedua-dua korpus menggunakan pendekatan TWEC (Di Carlo, Bianchi & Palmonari 2019). Korpus diakronik secara keseluruhan atau dipanggil sebagai *D*, adalah gabungan 19 korpus kecil dan dilatih sebagai kompas pembenaman atemporal atau dikenali sebagai *U* (rujuk Rajah 1). Korpus siri masa merupakan 19 korpus kecil secara berdikari di mana akan menjadi C^{t1} hingga C^{tN} dalam model TWEC. Memandangkan model TWEC dibina berasaskan tetapan *Word2Vec* pada *Continuous bag-of-words* (CBOW), tetapan yang sama (*Word2Vec* dan CBOW) digunakan untuk melatih kedua-dua korpus.

Pada tahap pembinaan alat analisis penilaian, dua jenis analisis persamaan dilakukan untuk menjalankan penilaian secara visual. Model TWEC digunakan untuk melakukan analisis *self-similarity* (Del Coco 2018) dan analisis metode *user-defined* (Boudih 2018) selepas korpus dilatih ke dalam model TWEC.

Analisis *self-similarity* dicadangkan oleh Del Coco (2018). Objektif analisis ini adalah untuk membandingkan vektor perkataan yang sama dalam ruang vektor yang berbeza supaya perubahan dalam semantik mengikut tempoh masa yang berbeza dapat diukur. Secara khusus, *self-similarity* adalah persamaan kosinus antara dua vektor bagi perkataan yang sama *i,* pada masa *t*, berkenaan dengan tempoh masa sebelumnya, *t-1*, dan persamaan adalah ditakrifkan seperti yang berikut (Del Coco 2018):

$$T_t(i) = \frac{TWE_t(i) \cdot TWE_{t-1}(i)}{\|TWE_t(i)\|_2 \|TWE_{t-1}(i)\|_2}$$

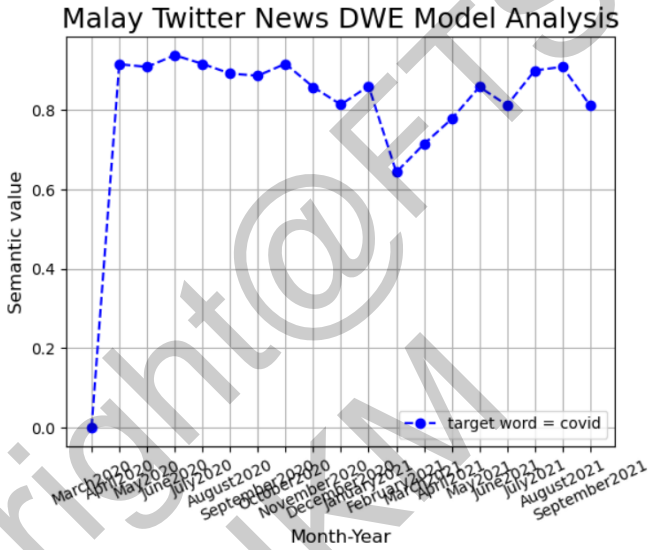
Oleh sebab analisis *self-similarity* hanya boleh digunakan pada data siri masa secara logik, ia sesuai untuk terpakai bagi model PPD projek ini. Melalui analisis ini, trend perkataan dalam korpus Twitter dapat dilihat melalui model PPD projek ini secara visual.

Analisis metode *user-defined* dicadangkan oleh Boudih (2018). Tujuan analisis ini adalah untuk membayangkan persamaan semantik antara dua perkataan mengikut perubahan masa dalam graf 2-dimensi. Pengkaji mencadangkan bahawa paksi-x sebagai tempoh masa yang berubah manakala paksi-y sebagai persamaan kosinus antara vektor bagi dua perkataan. Analisis ini menunjukkan kaitan antara dua perkataan, iaitu perkataan rujukan dan perkataan sasaran, mengikut perubahan masa dengan ditunjukkan dalam graf garis. Secara khusus, analisis metode *user-defined* adalah berdasarkan persamaan kosinus di mana digunakan untuk kiraan persamaan semantik pasangan perkataan (perkataan rujukan *x* dan perkataan sasaran *y*) pada masa tertentu *t* dan persamaan adalah ditakrifkan seperti yang berikut (Boudih 2018):

$$x_t, y_t = \frac{x_t \cdot y_t}{|x_t| \cdot |y_t|}$$

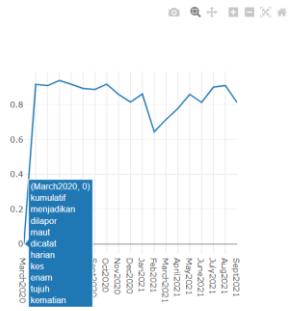
Keputusan dan Perbincangan

Dalam projek ini, dua jenis analisis persamaan telah dilakukan pada model PPD untuk menjalankan penilaian secara visual. Bahagian ini akan menunjukkan hasilan kedua-dua jenis analisis, iaitu analisis *self-similarity* dan analisis metode *user-defined*, dan menerangkan tentang hasilan yang diperoleh daripada dua jenis analisis. Selain itu, bahagian ini juga akan menunjukkan antaramuka yang direka bentuk.

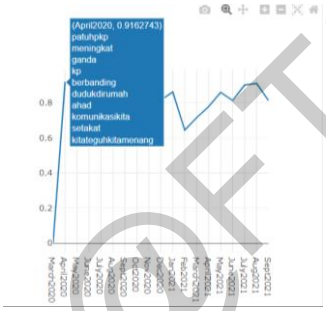


Rajah 2 Graf analisis *self-similarity* model PPD bagi perkataan ‘covid’

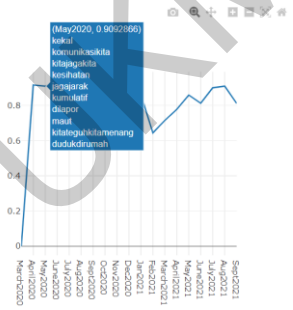
Untuk memperoleh nilai *self-similarity* model PPD, persamaan kosinus antara vektor bagi Mac 2020, dan vektor-vektor bagi bulan lain, iaitu bermula sejak April 2020 hingga September 2021, telah dikira. Melalui analisis ini, betapa kuatnya konsep bagi sesuatu perkataan diunjurkan mengikut perubahan masa dalam korpus berita Twitter bahasa Melayu dapat ditunjukkan. Dalam Rajah 2, graf tersebut menunjukkan hampir semua bulan memperoleh nilai *self-similarity* yang hampir maksimum. Oleh sebab demikian, konsep perkataan ‘covid’ boleh diandaikan sebagai konsep yang kuat dalam korpus berita Twitter bahasa Melayu.



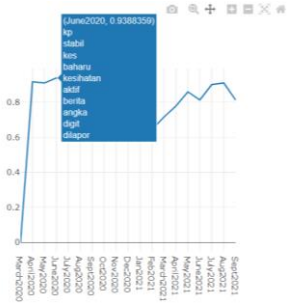
Rajah 3 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi Mac 2020



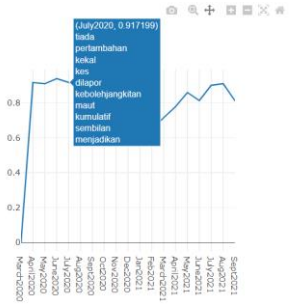
Rajah 4 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi April 2020



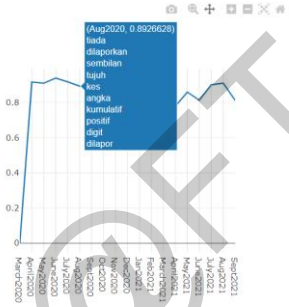
Rajah 5 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi Mei 2020



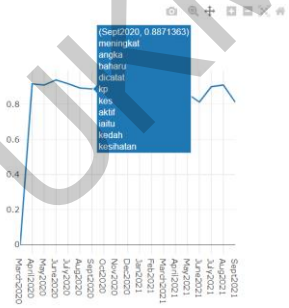
Rajah 6 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi Jun 2020



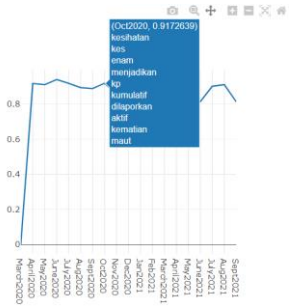
Rajah 7 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi Julai 2020



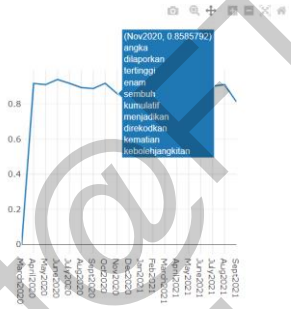
Rajah 8 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi Ogos 2020



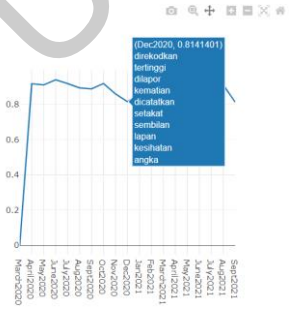
Rajah 9 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi September 2020



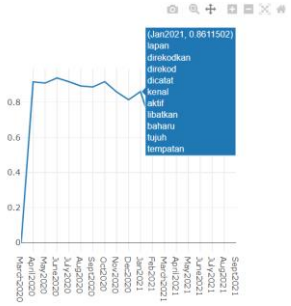
Rajah 10 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi Oktober 2020



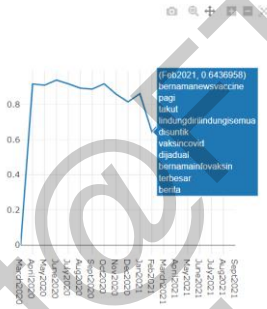
Rajah 11 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi November 2020



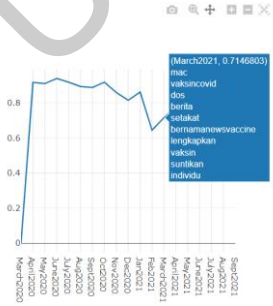
Rajah 12 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi Disember 2020



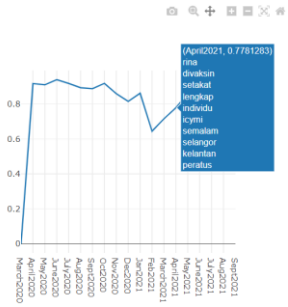
Rajah 13 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi Januari 2021



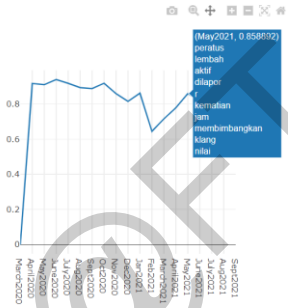
Rajah 14 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi Februari 2021



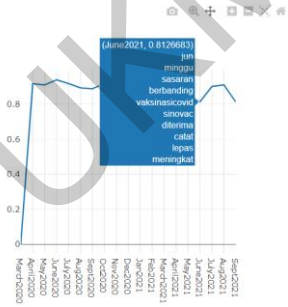
Rajah 15 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi Mac 2021



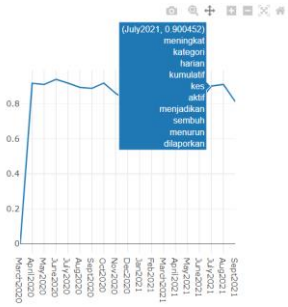
Rajah 16 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi April 2021



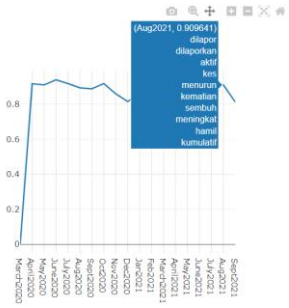
Rajah 17 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi Mei 2021



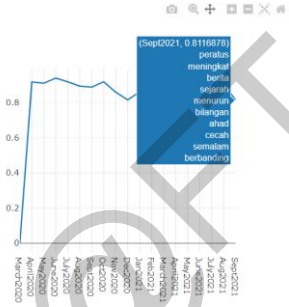
Rajah 18 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi Jun 2021



Rajah 19 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi Julai 2021



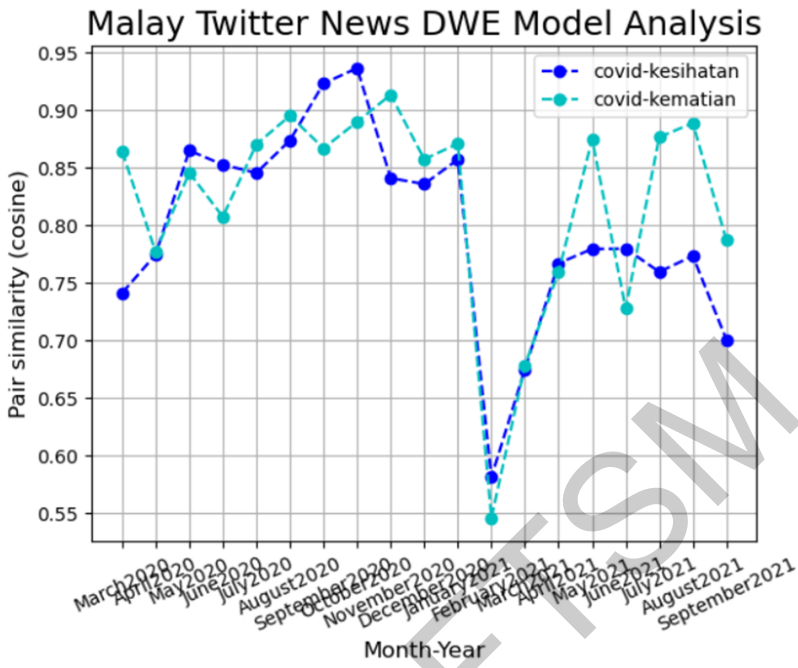
Rajah 20 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi Ogos 2021



Rajah 21 Graf menunjukkan perkataan paling serupa bagi perkataan ‘covid’ bagi September 2021

Dari Rajah 3 hingga Rajah 21, trend bagi perkataan ‘covid’ berubah dari ‘patuhpkp’, ‘dudukdirumah’ ke ‘vaksincovid’, ‘sembuh’. Ini jelas menunjukkan bahawa pada peringkat awal penularan COVID-19, kebanyakan berita Twitter bahasa Melayu mencatatkan berkaitan dengan Perintah Kawalan Pergerakan (PKP) dan menggalakkan rakyat duduk di rumah. Hal ini adalah untuk mengawal penularan COVID-19 di Malaysia pada tempoh masa tersebut. Sejak Februari 2021, kemunculan vaksin COVID-19 menyebabkan perkataan paling serupa bagi perkataan ‘covid’ telah berubah kepada yang berkaitan dengan vaksin. Walaupun maksud perkataan ‘covid’ tidak bertukar, perkataan paling serupa bagi perkataan ‘covid’ menunjukkan perubahan.

Secara ringkas, trend bagi perkataan ‘covid’ menunjukkan perubahan seperti yang ditunjukkan pada bentuk garisan dalam Rajah 2 dan perkataan berkelompok mengikut bulanan.



Rajah 22 Trend konsep perkataan ‘covid’-‘kesihatan’ dan konsep perkataan ‘covid’-‘kematian’ dalam korpus berita Twitter bahasa Melayu pada Mac 2020 hingga September 2021

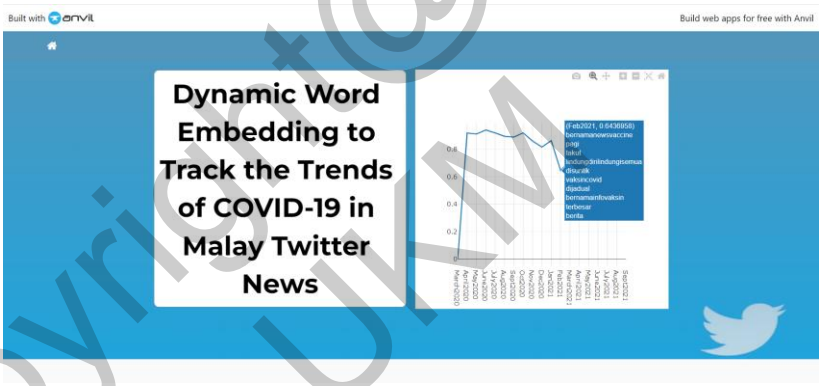
Rajah 22 menunjukkan perbandingan antara dua perkataan mengikut bulanan. Untuk analisis metode *user-defined*, perkataan ‘covid’ dipilih sebagai perkataan rujukan, dan perkataan ‘kesihatan’ dan perkataan ‘kematian’ dipilih sebagai perkataan sasaran.

Dalam Rajah 22, kedua-dua pasangan perkataan ‘covid’-‘kesihatan’ dan ‘covid’-‘kematian’ menunjukkan perkaitan semantik yang sederhana dalam model PPD. Pada Februari 2021, graf menunjukkan penurunan dalam kedua-dua pasangan perkataan. Penurunan tersebut menunjukkan konsep ‘covid’ berkaitan dengan ‘kesihatan’ dan ‘kematian’ kurang disebutkan dalam berita Twitter bahasa Melayu pada bulan tersebut berbanding bulan-bulan lain. Hal ini diwujudkan mungkin kerana kemunculan vaksin COVID-19 banyak mendapat perhatian berita Twitter bahasa Melayu seperti yang ditunjukkan dalam Rajah 14.

Secara ringkas, Model PPD dibuktikan bahawa boleh mengukur anjakan semantik secara visual pada korpus berita Twitter bahasa Melayu dengan menunjukkan perkembangan trend perkataan pasangan mengikut bulanan melalui analisis metode *user-defined*.

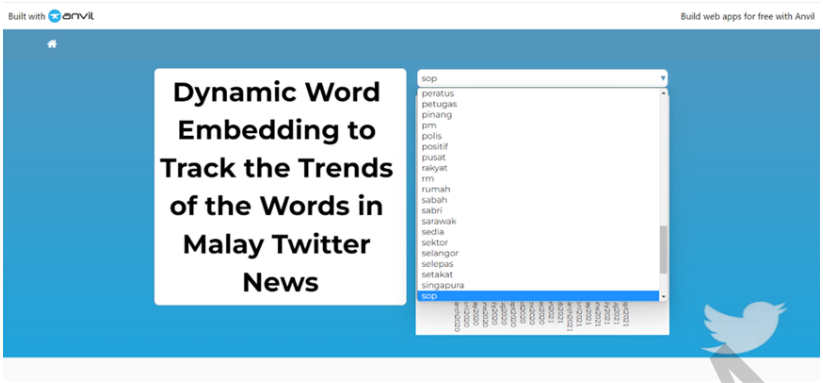
Rajah 23 Antaramuka utama bagi model PPD

Rajah 23 menunjukkan antaramuka utama bagi projek ini dengan dua butang yang akan membawa pengguna ke antaramuka masing-masing.

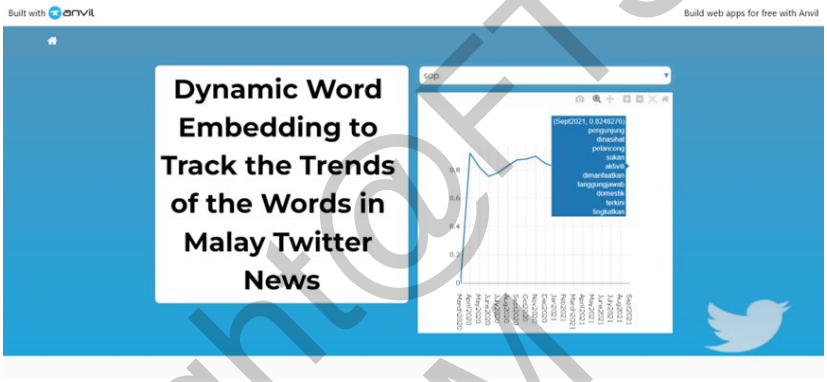


Rajah 24 Antaramuka menunjukkan analisis pada domain COVID-19

Rajah 24 menunjukkan antaramuka yang ditunjukkan kepada pengguna selepas klik butang pertama, ‘COVID-19 Trends’ dalam antaramuka utama. Satu graf yang mengandungi analisis pada domain COVID-19 ditunjukkan dan perkataan paling serupa bagi perkataan ‘covid’ ditunjukkan apabila pengguna mengarahkan tetikusnya ke titik-titik tertentu.



Rajah 25 Antaramuka menunjukkan senarai perkataan bagi analisis perkataan



Rajah 26 Antaramuka menunjukkan analisis perkataan

Rajah 25 menunjukkan antaramuka yang ditunjukkan kepada pengguna selepas klik butang kedua, 'List of Words Trends' dalam antaramuka utama. Pengguna dapat memilih perkataan-perkataan daripada senarai perkataan dan melihat trend bagi perkataan tertentu seperti yang ditunjukkan dalam Rajah 26. Perkataan paling serupa bagi perkataan terpilih oleh pengguna ditunjukkan apabila pengguna mengarahkan tetikusnya ke titik-titik tertentu.

Kesimpulan

Projek ini menggunakan pendekatan *Temporal Word Embeddings with a Compass* (TWEC) bagi membina model pembenaman perkataan dinamik (PPD) untuk berita Twitter bahasa Melayu. Korpus projek ini merupakan korpus Twitter yang diekstrak daripada akaun Twitter Agensi Berita Nasional

Malaysia (BERNAMA) bermula sejak Mac 2020 hingga September 2021 mengikut kata kunci dari empat kategori. Model ini menggunakan pendekatan TWEC untuk melatih korpus Twitter dan dua jenis analisis persamaan, iaitu analisis *self-similarity* dan analisis metode *user-defined*, telah dijalankan pada model PPD untuk menjalankan penilaian secara visual. Daripada dua jenis analisis persamaan, model PPD yang menggunakan pendekatan TWEC dapat menjejaki trend perkataan dalam berita bahasa Melayu pada platform media sosial dari semasa ke semasa. Akhir sekali, antaramuka telah direka bentuk untuk memudahkan pengguna menggunakan model PPD.

Kekuatan model PPD adalah dapat menganalisis model PPD secara visual. Trend perkataan telah ditunjukkan dalam antaramuka yang direka bentuk. Graf garis adalah digunakan untuk menunjukkan *self-similarity* dan perkataan paling serupa bagi perkataan tertentu pada setiap bulan bermula sejak Mac 2020 hingga September 2021. Ini dapat memudahkan membuat perbandingan untuk sesuatu perkataan pada masa yang berbeza.

Kekangan model PPD adalah menjumpai singkatan atau abjad dalam senarai perkataan dan juga perkataan paling serupa bagi sebahagian perkataan dalam antaramuka yang direka bentuk. Sebagai contoh, ‘pm’ yang ditunjukkan dalam senarai perkataan merupakan singkatan bagi Perdana Menteri manakala ‘sop’ ialah singkatan bagi prosedur operasi standard. Selain itu, korpus Twitter hanya melibatkan tempoh masa dari Mac 2020 hingga September 2021. Ini akan memberi impak kepada trend perkataan pada masa yang akan datang kerana akan menunjukkan perbezaan dengan trend perkataan tahun 2020 dan tahun 2021.

Cadangan untuk menambahbaikkan projek di masa hadapan adalah menambah lebih banyak tinjauan dalam platform media sosial yang berlainan menggunakan pendekatan PPT. Terdapat banyak platform media sosial yang digunakan oleh seluruh dunia malahan munculnya platform media sosial baharu seperti Threads. Ini dapat membantu dalam memahami trend masa kini secara mendalam. Selain itu, model PPD boleh ditambahbaik dengan meneroka cara yang lebih sesuai untuk melakukan pra-pemprosesan korpus supaya masalah kewujudan singkatan dan abjad dapat dielakkan.

Penghargaan

Pertama sekali saya ingin mengucapkan setinggi-tinggi penghargaan dan terima kasih kepada penyelia Dr. Sabrina binti Tiun. Beliau telah banyak meluangkan masa dalam memberi tunjuk ajar, bantuan, dan nasihat yang begitu berguna dan bernilai sepanjang projek ini. Panduan beliau yang berharga dan teliti telah menginspirasi saya dalam cara yang tidak terhitung jumlahnya.

Selain itu, saya juga ingin mengucapkan terima kasih kepada pensyarah-pensyarah Fakulti Teknologi dan Sains Maklumat yang telah banyak mendidik saya sepanjang tempoh masa pembelajaran di Universiti Kebangsaan Malaysia. Keterlibatan mereka telah meningkatkan kematangan intelektual saya yang akan membatu saya pada masa akan datang.

Seterusnya, ucapan terima kasih tidak terhingga untuk ahli keluarga yang tercinta dan rakan-rakan yang telah banyak memberi sokongan sepanjang projek ini.

Akhir sekali, seikhlas tulus kata terima kasih kepada semua pihak yang telah memberi bantuan, nasihat, dan bimbingan secara langsung dan tidak langsung dalam menjayakan projek ini.

RUJUKAN

Bakar, M.F.R.A., Idris, N. & Shuib, L. 2019. An Enhancement of Malay Social Media Text Normalization for Lexicon-Based Sentiment Analysis. *Proceedings of the 2019 International Conference on Asian Language Processing, IALP 2019*: 211–215.

Boudih, A.M. 2018. A New Way of Visualizing Semantic Similarity over Time: 1–67.

Di Carlo, V., Bianchi, F. & Palmonari, M. 2019. Training temporal word embeddings with a compass. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*: 6326–6334.

Del Coco, P.E.J. 2018. Temporal Text Mining: From Frequencies to Word Embeddings.

Gholizadeh, S., Seyeditabari, A. & Zadrozny, W. 2020. A Novel Method of Extracting

Topological Features from Word Embeddings.

Gohourou, D., Kurita, D., Kuwabara, K. & Huang, H.H. 2017. International business matching using word embedding. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10191 LNAI: 181–190.

Gong, H., Bhat, S. & Viswanath, P. 2020. Enriching Word Embeddings with Temporal and Spatial Information: 1–11.

Hao, L.Y. & Yan, J.L.S. 2022. English-Malay Word Embeddings Alignment for Cross-lingual Emotion Classification with Hierarchical Attention Network. *WASSA 2022 - 12th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Proceedings of the Workshop*: 113–124.

Jun, Y. 2021. Measuring Semantic Changes Using Temporal Word Embeddings. https://towardsdatascience.com/measuring-semantic-changes-using-temporal-word-embedding-6fc3f16cfdb4 [30 November 2022].

Lakmal, D., Ranathunga, S., Peramuna, S. & Herath, I. 2020. Word Embedding Evaluation for Sinhala. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*(May): 1874–1881.

Powell, J. & Sentz, K. 2021. Tracking Short-Term Temporal Linguistic Dynamics to Characterize Candidate Therapeutics for COVID-19 in the CORD-19 Corpus. *Proceedings of SenSys 2020: 18th ACM Conference on Embedded Networked Sensor Systems (SenSys 2020)*, hlm. Association for Computing Machinery.:

Sahlgren, M. 2008. The distributional hypothesis. *Italian Journal of Linguistics* 20(1): 33–53.

Shen, Y. & Stringhini, G. 2019. ATTACK2VEC: Leveraging temporal word embeddings to understand the evolution of cyberattacks. *Proceedings of the 28th USENIX Security Symposium*: 905–921.

Si, M. 2022. Word Embedding for Analogy Making. *International Conference on Innovative Computing and Cloud Computing*.

Tiun, S., Saad, S., Mohd Noor, N.F., Jalaludin, A. & Che Abdul Rahman, A.N. 2020. Quantifying Semantic Shift Visually on a Malay Domain Specific Corpus Using Temporal Word Embedding Approach. *Asia-Pacific Journal of Information Technology and Multimedia* 09(02): 1–10.

Volpetti, C., Vani, K. & Antonucci, A. 2020. Temporal Word Embeddings for Narrative Understanding. *ACM International Conference Proceeding Series*: 68–72.

Xia, C., Zhang, H., Moghtader, J., Wu, A. & Chang, K.W. 2019. Visualizing trends of key roles in news articles. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Proceedings of System Demonstrations*(2008): 247–252.

Yao, Z., Sun, Y., Ding, W., Rao, N. & Xiong, H. 2018. Dynamic word embeddings for evolving semantic discovery. *WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining* 2018-Febua: 673–681.

Jacqueline Hii Sing Hee (A179361)
Dr. Sabrina Tiun
Fakulti Teknologi & Sains Maklumat,
Universiti Kebangsaan Malaysia