Volume 20(3), August 2020 http://doi.org/10.17576/gema-2020-2003-05

Model Peringkasan Teks Ekstraktif Dwibahasa menggunakan Fitur Kekangan Corak Tekstual

Suraya Alias[a]
suealias@ums.edu.my
Universiti Malaysia Sabah

Mohd Shamrie Sainin[b]
shamrie@ums.edu.my
Universiti Malaysia Sabah

Siti Khaotijah Mohammad
sitijah@usm.my
Universiti Sains Malaysia

ABSTRAK

Di dalam era pencarian maklumat digital, sebuah ringkasan yang dijana secara automatik dapat membantu pembaca mendapatkan maklumat penting dan relevan dengan lebih mudah. Sebahagian besar kajian dan set data penanda aras dalam bidang peringkasan teks secara automatik adalah dalam bahasa Inggeris. Justeru itu, terdapat keperluan kajian dalam bahasa Melayu agar potensi dalam bidang ini lebih kompetitif. Kajian ini juga menyoroti masalah dalam mengenal pasti dan menjana maklumat penting dalam penyediaan ringkasan ekstraktif. Ini kerana model perwakilan teks yang sedia ada seperti BOW mempunyai kelemahan dalam perwakilan semantik yang kurang tepat dan model N-gram pula mempunyai isu penghasilan dimensi vektor kata yang sangat tinggi. Dalam kajian ini, sebuah model peringkasan teks dwibahasa dinamakan MYTextSumBASIC telah dibangunkan untuk menghasilan ringkasan ekstraktif secara automatik dalam versi bahasa Melayu dan bahasa Inggeris. Model MYTextSumBASIC ini menggunakan model perwakilan teks dikenali sebagai FASP yang telah diimprovisasi dengan menggunakan tiga Fitur Kekangan Corak Tekstual iaitu kekangan item kata, kekangan kata urutan bersebelahan dan kekangan saiz urutan. Terdapat tiga fasa utama dalam rangka kerja model MYTextSumBASIC iaitu pembangunan korpus ringkasan bahasa Melayu, pembangunan model MYTextSumBASIC menggunakan perwakilan FASP dan penilaian ringkasan. Dalam fasa penilaian, dengan menggunakan 100 wacana berita bahasa Melayu, prestasi ringkasan yang dihasilkan secara automatik oleh MYTextSumBASIC telah mengatasi ringkasan dari model Baseline (Lead) dan OTS dengan nilai purata tertinggi bagi dapatan semula (R) ialah 0.5849, kejituan (P) ialah 0.5736 dan skor-F (Fm) ialah 0.5772. Bagi penilaian secara manual oleh pakar bahasa, kaedah MYTextSumBASIC telah menghasilkan skor kebolehbacaan sebanyak 4.1 dan 3.87 untuk skor isi kandungan ringkasan yang dihasilkan menggunakan set data rawak. Eksperimen selanjutnya menggunakan set data tanda aras bahasa Inggeris DUC 2002 sebanyak 102 wacana berita juga telah menunjukkan model MYTextSumBASIC telah mengatasi sistem terbaik dan tercorot dalam perbandingan tersebut dengan nilai purata dapatan semula ROUGE-1 (0.43896) dan ROUGE-2 (0.19918). Kesimpulan dari penilaian ringkasan dapat merumuskan bahawa kaedah perwakilan teks FASP yang digunakan sebagai fitur oleh MYTextSumBASIC boleh diaplikasi untuk teks dwibahasa dengan prestasi kompetitif melalui perbandingan dengan model peringkasan teks bahasa Inggeris yang sedia ada.

---

[a] *Pengarang utama*
[b] *Pengarang koresponden*

eISSN: 2550-2131
ISSN: 1675-8021

Kata Kunci: Fitur Kekangan Corak Tekstual; Peringkasan Teks; Pertumbuhan Corak-Tersusun; Bahasa Melayu

Bilingual Extractive Text Summarization Model using Textual Pattern Constraints

ABSTRACT

In the era of digital information, an auto-generated summary can help readers to easily find important and relevant information. Most of the studies and benchmark data sets in the field of text summarization are in English. Hence, there is a need to study the potential of Malay language in this field. This study also highlights the problems in identifying and generating important information in extractive summaries. This is because existing text representation models such as BOW has weaknesses in inaccurate semantic representation, while the N-gram model has the issue of producing very high word vector dimensions. In this study, a bilingual text summarization model named MYTextSumBASIC has been developed to generate an extractive summary automatically in Malay and English. The MYTextSumBASIC summarizer model applies a text representation model known as FASP using three Textual Pattern Constraints, namely word item constraints, adjacent word constraints and sequence size constraints. There are three main phases in the framework of MYTextSumBASIC model, which are the development of the Malay language corpus, the development of MYTextSumBASIC model using FASP and the summary evaluation phase. In the summary evaluation phase, using the Malay language data sets of 100 news articles, the summaries produced by MYTextSumBASIC outperformed the summary generated by Baseline (Lead) and OTS summarizer with the highest average for retrieval (R) is 0.5849, precision (P) is 0.5736 and the F-score (Fm) is 0.5772. For manual evaluation by linguists, the MYTextSumBASIC method yielded a reading score of 4.1 and 3.87 for summary content generated using a random data set. Further experiments using the 2002 DUC English benchmark data set of 102 news articles have also shown that the MYTextSumBASIC model outperformed the best and lowest systems in the comparison with the mean retrieval values of ROUGE-1 (0.43896) and ROUGE-2 (0.19918). These findings conclude that the FASP text representation feature along with the textual pattern constraints used by our model can be used for bilingual text with competitive performance compared to other text summarization models.

Keywords: Textual Pattern Constraint; Text Summarization; Sequential Pattern-Growth; Malay language

PENGENALAN

Dalam era kebanjiran data di atas talian, pengguna sering menghadapi masalah untuk mencari dan menyisih lambakan maklumat yang sedia ada. Justeru itu, sebuah ringkasan yang dijana secara automatik dapat memberikan persepsi tentang maklumat utama yang berkaitan di dalam teks dengan lebih mudah dan cepat. Bidang Peringkasan Teks Automatik atau ATS adalah satu proses automatik bagi menghasilkan ringkasan daripada satu atau lebih sumber input dokumen. Menurut (Nenkova & McKeown, 2011), terdapat beberapa jenis ringkasan iaitu ringkasan umum seperti ringkasan berita atau artikel, ringkasan bersifat fokus-pertanyaan berdasarkan topik pengguna yang khusus/spesifik, ataupun ringkasan berdasarkan sentimen yang berupaya meringkaskan pendapat pengguna. Sementara itu, terdapat dua kaedah untuk menghasilkan ringkasan secara automatik iaitu melalui kaedah ekstraktif atau abstraktif (Mahajani, Pandya, Marijuso & Sharma, 2019).

Bagi kaedah ekstraktif, sebuah model peringkasan teks akan memilih dan merangkaikan ayat-ayat yang paling penting bagi menghasilkan versi ringkasan dokumen yang lebih pendek tanpa mengubah sumber dan struktur asal ayat tersebut. Ayat-ayat akan diberi skor mengikut kepentingannya berdasarkan fitur tertentu seperti fitur luaran dan fitur isi kandungan. Antara fitur luaran yang lazim digunakan adalah tajuk, kedudukan ayat dan frekuensi perkataan. Sementara itu, fitur bagi isi kandungan pula merujuk pada ayat-ayat yang membawa maklumat paling signifikan dan mengandungi perkataan-perkataan yang sama dengan tajuk dokumen tersebut (Ferreira et al., 2014; Litvak & Last, 2013). Bagi menghasilkan ringkasan, model peringkas ekstraktif akan memilih ayat dengan skor tertinggi dan isi terpenting.

Sebaliknya, kaedah abstraktif pula akan menghasilkan ringkasan dengan mengolah, memparafrasa dan menggabungkan maklumat dari ayat-ayat yang berkaitan untuk membentuk sebuah ayat yang baru. Kaedah ini adalah lebih mirip kepada ringkasan yang dihasilkan oleh manusia di mana Pemprosesan Bahasa Tabii (NLP) secara ekstensif dan maklumat terdahulu diperlukan. Oleh kerana proses penjanaan ringkasan berdasarkan kaedah abstraktif agak rumit dan kompleks, kaedah ekstraktif lebih mendominasi bidang kajian ATS sehingga kini. Namun demikian, masih banyak ruang untuk diperbaiki dan keperluan untuk menghasilkan ringkasan berkualiti secara automatik telah dibincangkan dalam (Gambhir & Gupta, 2017).

Sebahagian besar kajian terdahulu dan terbaharu dalam bidang ringkasan teks secara ekstraktif kebanyakannya di dalam bahasa Inggeris. Set data tanda aras yang sedia ada, khususnya daripada Document Understanding Conference (DUC) dan Text Analysis Conference (TAC) yang diurus oleh National Institute of Standards and Technology (NIST) juga kebanyakannya dalam bahasa Inggeris. Masalah utama bidang kajian ini adalah kekurangan set data tanda aras dalam bahasa Melayu yang boleh digunakan untuk menilai ringkasan-ringkasan yang dihasilkan seperti dinyatakan dalam (Jusoh, Masoud, & Alfawareh, 2011; Zamin & Ghani, 2010). Isu ini telah mendorong pembangunan korpus ringkasan Bahasa Melayu oleh (Alias, Mohammad, Hoon, & Ping, 2016) untuk mengekstrak corak ringkasan yang dihasilkan oleh manusia yang mana lebih fleksibel dan padat agar dapat memanfaatkan bidang peringkasan teks dalam bahasa Melayu. Terkini dapat dilihat kajian dalam bidang peringkasan teks bahasa Melayu semakin mendapat perhatian seperti dalam menghasilkan ringkasan isi utama menggunakan fitur kata tanpa seliaan oleh (Noah, Ali, & Hasan, 2018). Namun demikian, isu kekurangan set data aras piawai untuk penilaian masih perlu ditangani dan memerlukan perhatian agar bidang ini lebih kompetitif.

Isu seterusnya merupakan masalah dalam mengenal pasti dan mengekstrak maklumat yang penting untuk menjana sesebuah ringkasan secara automatik. Ini kerana model perwakilan teks yang sedia ada seperti BOW mempunyai kelemahan dalam perwakilan semantik yang kurang tepat disebabkan urutan susunan perkataan yang tidak dikekalkan (Ning, Yuefeng, & Sheng-Tang, 2012). Sementara itu, model N-gram pula mempunyai isu penghasilan dimensi vektor kata yang sangat tinggi, dimana isu kekosongan nilai data membawa kepada masalah dalam pencarian kombinasi perkataan yang penting dan signifikan (Kim, Park, Lu, & Zhai, 2012; Le & Mikolov, 2014).

Kajian ini menerangkan secara terperinci proses pembangunan sebuah model peringkasan teks dwibahasa (MYTextSumBASIC) untuk menghasilan ringkasan ekstraktif secara automatik dalam versi bahasa melayu dan bahasa Inggeris. Model MYTextSumBASIC ini menggunakan model perwakilan teks berdasarkan teknik Pertumbuhan-Corak Tersusun yang dinamakan *Frequent Adjacent Sequential Pattern* (FASP) sebagai fitur utama untuk mengekstrak maklumat penting dan relevan di dalam sesebuah dokumen. Bagi tujuan ini, kaedah FASP telah diimprovisasi dengan menggunakan tiga Fitur Kekangan Corak Tekstual iaitu: 1) kekangan item kata, 2) kekangan kata urutan bersebelahan dan 3) kekangan saiz urutan.

Hasil daripada kajian ini telah mendapati bahawa perwakilan teks FASP boleh digunakan sebagai fitur yang mewakili maklumat signifikan dan relevan bagi sesebuah ayat dalam membentuk sesebuah ringkasan. Eksperimen menggunakan set data bahasa Inggeris dan bahasa Melayu juga telah menunjukkan bahawa perwakilan fitur FASP tidak dipengaruhi oleh jenis bahasa. Ini disebabkan kerana penjanaan fitur FASP adalah berdasarkan corak data tanpa seliaan dan ia boleh diaplikasi pada domain-domain set data yang berlainan.

SOROTAN KAJIAN

Di dalam bidang ATS, bagi mengenalpasti topik utama di dalam sesebuah teks atau dokumen, kebanyakan model peringkasan ekstraktif menjalankan tiga tugas utama seperti yang diterangkan dalam (Nenkova & McKeown, 2012). Pertama, penjanaan perwakilan teks yang mengandungi sumber penting atau topik utama yang mewakili kandungan input diperlukan. Input daripada teks tersebut boleh diwakili oleh fitur yang mengandungi senarai vektor perkataan, N-gram dan model graf dengan tahap granulariti yang berbeza. Seterusnya, proses kedua ialah mendapatkan skor ayat berdasarkan perwakilan teks input tadi. Di sini, setiap model peringkasan akan menggunakan pelbagai kaedah seperti kaedah statistikal, pembelajaran mesin dan model graf untuk menganggar hubungan dan kepentingan setiap ayat yang akan diekstrak. Akhir sekali, proses pemilihan ayat akan dilakukan berdasarkan saiz keperluan ringkasan tersebut. Beberapa teknik pemilihan ayat boleh digunakan seperti pendekatan rakus, algoritma pengoptimuman global dan pendekatan secara penggugusan untuk menghasilkan ringkasan akhir.

KAEDAH PERWAKILAN TEKS

Bagi menghasilkan sebuah ringkasan, kebanyakan model peringkasan yang sedia ada sering menggunakan kaedah perwakilan teks yang asas, iaitu menggunakan model bahasa tradisional seperti Bag-of-Words (BOW) bagi mewakili setiap istilah dalam teks yang dijadikan sebagai vektor kata kunci *n*-dimensi seperti yang terdapat dalam (Conroy, Schlesinger, O'leary, & Goldstein, 2006). Sementara itu, terdapat model-model peringkasan yang lain telah mengeksploitasi model kebarangkalian model N-gram sebagai fitur ayat (Clarke & Lapata, 2008) dan ada juga yang mewakilkan model N-gram tersebut dalam bentuk graf (Ganesan, Zhai, & Han, 2010; Khan, Salim, Reafee, Sukprasert, & Kumar, 2015; Van Lierde & Chow, 2019). Walau bagaimanapun, terdapat beberapa isu ketara di dalam model-model bahasa ini seperti perwakilan semantik yang kurang tepat dan maksud perkataan yang terpesong. Sebagai contoh seperti yang terdapat dalam perwakilan BOW yang mengalami masalah dalam perbandingan persamaan dalam ayat kerana urutan susunan perkataan yang tidak dikekalkan (Ning, Yuefeng, & Sheng-Tang, 2012). Sebaliknya, isu ketara bagi model N-gram adalah kombinasi saiz perkataan yang berdimensi tinggi akan terhasil di mana tidak semua kombinasi perkataan boleh didapati iaitu isu kekosongan nilai data (Kim, Park, Lu, & Zhai, 2012; Le & Mikolov, 2014).

Oleh sebab ini, sebilangan penyelidik telah beralih arah dengan memanipulasikan teknik perwakilan teks dengan menggunakan Corak Kerap atau Frequent Patterns (FP) yang didapati di dalam teks atau dikenali juga sebagai "*textual patterns*". Di sini, FP yang mengikut urutan kedudukan item dinamakan Frequent Sequential Pattern (FSP) atau Corak Tersusun Kerap. Perwakilan teks berdasarkan corak mempunyai kelebihan untuk mengkorelasi perkataan-perkataan secara tabii dan pada masa yang sama juga masih dapat mengekalkan hubungan semantik perkataan tersebut. Melalui perwakilan FSP juga corak unit teks yang bermakna dari dokumen dapat ditemui tanpa perlu bergantung pada pengetahuan linguistik atau pengetahuan yang terdahulu.

Contohnya, terdahulu (Ledeneva, Gelbukh, & García-Hernández, 2008) telah mengkaji kegunaan Maximal Frequent Patterns (MFS) untuk mewakili isi kandungan yang signifikan daripada sebuah dokumen. Model peringkasan berdasarkan corak yang telah dibangunkan dapat memperbaiki nilai panggil balik ringkasan ekstraktif tunggal jika dibandingkan dengan model peringkas yang menggunakan kaedah perwakilan BOW dan N-gram. Sementara itu, percubaan oleh Qiang, Chen, Ding, Xie dan Wu (2016) dengan menggunakan perwakilan Closed Patterns (CP) adalah untuk menyingkirkan ayat-ayat lewah dengan mengekalkan ayat-ayat penting telah menunjukkan kegunaan model peringkasan berdasarkan corak dalam bidang peringkasan dokumen yang pelbagai. Kajian terbaharu oleh (Xie, Wu, & Zhu, 2017) pula telah mencadangkan penggunaan pengehadan kad bebas untuk mengekstrak FSP sebagai frasa utama telah berjaya mengenal pasti topik penting sebuah dokumen. Kajian mereka telah menunjukkan keputusan yang memberangsangkan dengan menggunakan ujian set data penanda aras frasa utama di mana telah menyokong kemampuan perwakilan teks berdasarkan corak dalam bidang ATS.

TEKNIK-TEKNIK PERINGKASAN TEKS SECARA EKSTRAKTIF

Bahagian sorotan kajian ini memperkenalkan secara ringkas teknik-teknik sedia ada yang telah diaplikasi dalam proses peringkasan ekstraktif yang secara umumnya boleh diklasifikasikan kepada kaedah statistikal yang berdasarkan fitur, pembelajaran mesin dan pembelajaran mendalam, model graf dan pendekatan berdasarkan corak.

KAEDAH STATISTIK

Model peringkasan berdasarkan kaedah statistik mengeksploitasi fitur daripada dokumen bagi mengekstrak ayat-ayat penting untuk dimasukkan ke dalam sebuah ringkasan. Iaitu, apabila lebih tinggi skor ayat yang dihitung berdasarkan fitur-fitur tertentu, lebih besar peluang untuk ayat tersebut dipilih untuk dimasukkan ke dalam ringkasan. (Luhn, 1958) yang dikenali sebagai perintis dalam bidang ATS yang telah memperkenalkan pendekatan kaedah statistikal ini. Kajian terdahulu beliau adalah berasaskan frekuensi perkataan dan frasa dengan memberi tumpuan kepada dokumen-dokumen teknikal. Sedekad kemudian, beberapa fitur aras luaran telah digunakan oleh (Edmundson, 1969) untuk mengenalpasti kepentingan sesebuah petikan atau ayat. Fitur-fitur tersebut adalah seperti tajuk, kedudukan ayat dan perkataan petunjuk sebagai contoh frasa "*in summary*" dan "*in conclusion*" di dalam sebuah dokumen. Sehingga kini, fitur-fitur ini kekal sebagai heuristik dalam fasa penetapan skor ayat bagi kebanyakan sistem ATS dan ia juga telah dijadikan rujukan dalam kajian ini. Fitur-fitur penting lain seperti frekuensi perkataan, TF-IDF, saiz ayat, kedudukan ayat, persamaan ayat dengan tajuk dan persamaan leksikal juga telah dikaji secara ekstensif oleh model-model peringkasan terdahulu seperti SUMBASIC oleh (Nenkova & Vanderwende, 2005).

Kaedah Statistik ini mempunyai kelebihan bagi set data yang mempunyai fitur yang teratur seperti set data berita di mana ia telah disokong dalam kajian terbaharu dalam (Ferreira et al., 2013). Namun demikian, untuk set data yang bersifat lebih umum dan tidak teratur seperti data dari media sosial, tidak banyak fitur luaran dokumen dapat dieksploitasi secara statistik dan ia memerlukan gabungan dari kaedah-kaedah peringkasan teks yang lain.

PEMBELAJARAN MESIN DAN PEMBELAJARAN MENDALAM

Teknik pembelajaran mesin boleh dibahagikan kepada tiga jenis pendekatan iaitu: diselia, tidak diselia dan separa diselia. Dalam pendekatan diselia, model peringkasan yang dilatih akan belajar memilih ayat-ayat yang penting daripada ringkasan-ringkasan yang telah dihasilkan oleh manusia. Terkini, (Verma, Yadav & Jain, 2019) telah menggabungkan fitur ayat seperti

gaya teks, panjang saiz ayat dan sintaks ayat dengan pendekatan pembelajaran mesin bagi memastikan bahawa fitur teks generik ini berupaya menghasilkan ekstrak ringkasan umum walaupun bagi domain-domain yang berlainan. Isu-isu berbangkit dalam melatih model peringkasan seperti kesulitan dalam melatih ayat-ayat panjang dan mengelaskan ayat-ayat yang serupa selalunya berlaku semasa memetakan teks kepada model urutan-demi urutan (Seq2Seq). Dalam kes ini, (M. Denil, 2014) telah mengeksploitasi Dynamic Convolution Neural Network (DCNN) yang berupaya belajar untuk menapis perlingkaran secara hierarki pada peringkat dokumen dan ayat. Pendekatan baru teknik visual mereka telah berupaya memberi pemahaman terhadap proses pembelajaran tersebut dan berupaya menghasilkan ringkasan ekstraktif dengan mengekalkan nuansa semantik ayat tersebut. Sementara itu, model peringkas (Narayan, Cohen, & Lapata, 2018) telah dioptimumkan dengan menggunakan pengukuhan objektif pembelajaran melalui tugas menetapkan ranking ayat di mana ayat dengan ranking lebih tinggi selalunya akan dipilih berdasarkan kekerapan ayat tersebut muncul dan skor tinggi oleh ayat tersebut.

Walau bagaimanapun, untuk teknik pembelajaran mesin yang diselia, ia memerlukan banyak proses latihan dan data yang berlabel. Ini berbeza dengan teknik yang tidak diselia seperti pengelompokan yang menghasilkan ringkasan berdasarkan pola atau struktur kelompok ayat yang ditemui dari dokumen yang diberikan.

MODEL GRAF

Kaedah perwakilan teks dengan menggunakan pendekatan graf merupakan salah satu kaedah yang mendapat perhatian penyelidik di dalam bidang Carian Maklumat. Menggunakan pendekatan graf, setiap nod verteks boleh digunakan untuk mewakili unit teks seperti perkataan, frasa atau ayat dan sisi graf akan menghubungkan verteks-verteks yang berkaitan. Algoritma ranking berdasarkan graf TextRank berfungsi dengan menggunakan "undian" untuk mempertimbangkan setiap verteks di mana undian yang tinggi menunjukkan kepentingan verteks tersebut. Baru-baru ini (Xie et al., 2017) telah cuba memperbaiki teknik penetapan skor ayat dengan menggabungkan model berdasarkan graf dengan model baru co-ranking bagi hubungan antara perkataan dan ayat yang dinamakan CoRank. Anggapan mereka bahawa setiap perkataan sepatutnya mempunyai berat pincang telah menunjukkan keputusan lebih baik berbanding teknik dasar TextRank dengan menggunakan sumber berita bahasa Cina dan set data DUC 2002.

Penyelidikan terkini oleh (Van Lierde & Chow, 2019) telah menekankan isu-isu dalam proses mengekstrak ayat ringkasan yang berulang dimana nod tersebut diberikan skor yang tinggi dengan anggapan ayat yang berulang mempunyai signifikan yang tinggi. Model peringkasan berdasarkan hipergraf yang mereka hasilkan merujuk kepada hipergraf berdasarkan tema, di mana setiap tema dipertimbangkan berdasarkan kepentingannya dalam korpus tersebut dan berdasarkan pertanyaan tentuan pengguna untuk mengurangkan masalah pemilihan ayat.

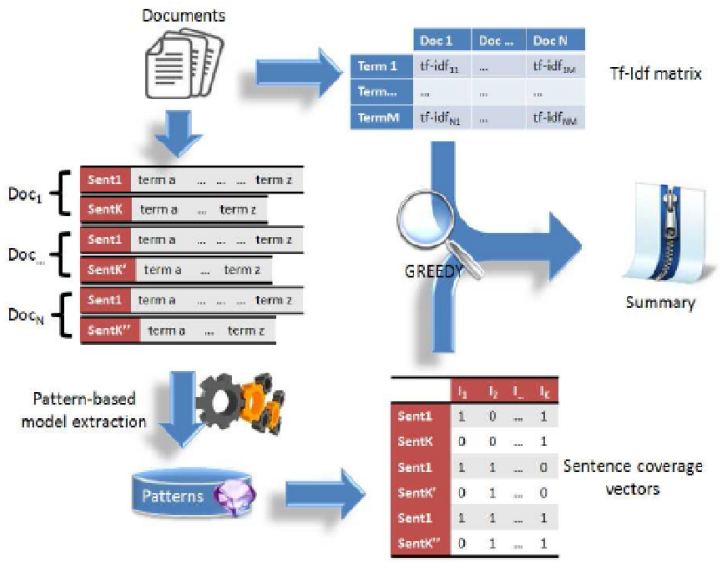
Secara keseluruhannya, model graf mampu menghasilkan ringkasan yang kompetif dan keputusan yang baik. Namun begitu, masalah kehilangan maklumat penting (kira-kira 48% hingga 60%) seperti dinayatakan di kajian (Boudin & Morin, 2013) adalah diwarisi dari masalah perwakilan teks menggunakan N-gram memerlukan perhatian dan penambahanbaikan dalam bidang ini.

TEKNIK BERDASARKAN CORAK

Berbeza dengan pendekatan graf, model peringkasan berdasarkan corak cuba untuk menangani isu perwakilan unit teks dengan menggunakan corak-corak yang terdapat dalam teks tanpa bergantung kepada maklumat terdahulu mahupun ilmu linguistik. Sebagai contoh dengan menggunakan perwakilan FP, MFS dan CP. MFS boleh didefinisikan sebagai FP yang tidak terkandung dalam mana-mana FP yang bersaiz lebih besar. Ia boleh digunakan sebagai

perwakilan yang kompak memandangkan saiz setiap MFS ditentukan sendiri mengikut teks. Para penyelidik telah menguji perwakilan MFS berbanding model BOW dan N-gram dalam model peringkasan masing-masing menggunakan set data DUC 2002. Mereka telah mendapati ia telah mengurangkan dimensi ruang vektor dan berjaya menyatakan isi kandungan utama di dalam sesebuah dokumen. Keputusan ringkasan ekstraktif tunggal dengan kepanjangan 100 perkataan telah melaporkan nilai panggil balik purata sebanyak 0.44085 dalam (Ledeneva et al., 2008) dan nilai skor-F sebanyak 0.44698 dengan menggunakan pemberat Boolean dalam (García-Hernández & Ledeneva, 2009) menggunakan penilaian ROUGE.

(Baralis, Cagliero, Jabeen, & Fiori, 2012) pula telah mengeksploitasi penggunaan Perlombongan Set Item Kerap dalam model mereka yang dinamakan PatTextSum seperti yang diilustrasi dalam Rajah 1. Model peringkas berdasarkan set item ini menggabungkan model perwakilan FP dan BOW untuk memilih ayat-ayat yang paling signifikan dan tidak berlebihan secara automatik untuk disertakan dalam ringkasan menggunakan pemberat statistik TF-IDF. Mereka menjalankan kajian berbanding dengan model peringkas OTS dan TexLexAn menggunakan penilaian ROUGE. OTS oleh (Rotem, 2019) merupakan model peringkasan sumer terbuka berbilang bahasa yang mempunyai bahasa Melayu sebagai salah satu tetapan pengaturan bahasa mereka. OTS menggunakan fail XML untuk menentukan peraturan tatabahasa POS dan kamus kata umum yang dikecualikan seperti "the", "a" dan beberapa kata hubung. Hasil eksperimen PatTextSum melaporkan keputusan yang signifikan dalam skor ROUGE-2 dengan nilai skor-F sebanyak 0.210 dalam set data domain teknologi dan 0.141 dalam set data domain bencana alam.



RAJAH 1. Model PatTextSum (Baralis, Cagliero, Jabeen, & Fiori, 2012)

Kajian semasa yang menggunakan perwakilan CP oleh Qiang et al. (2016) telah menunjukkan kebolehan perwakilan berdasarkan corak jika dibandingkan dengan kaedah berasaskan istilah yang lain dalam proses peringkasan dokumen yang pelbagai dan ia juga mengatasi (Baralis et al., 2012) dalam penilaian ROUGE-2 skor-F. Dengan menggunakan Corak Tertutup CP, ayat-ayat paling penting akan diekstrak dan ayat-ayat lewah akan dikurangkan. CP juga adalah perwakilan kompak bagi FSP yang tidak termasuk dalam FSP lain yang mempunyai nilai sokongan yang sama dan inilah yang membezakannya daripada MFS. Kajian penyelidik ini juga membuat perbandingan terhadap perwakilan berdasarkan corak yang lain seperti FP dan set item tertutup. Penyelidik telah menyimpulkan bahawa semua

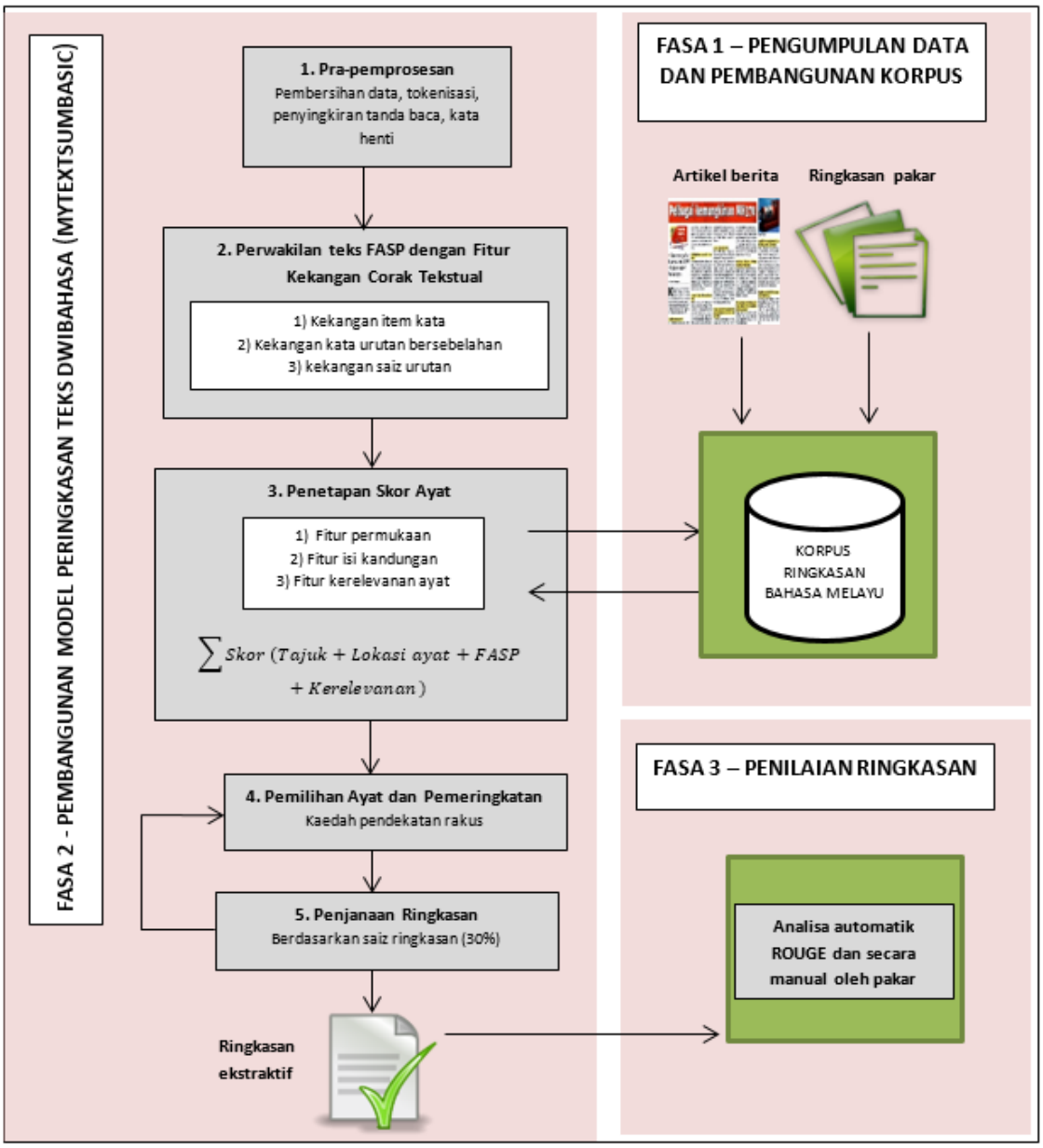
strategi adalah efektif dengan cara tersendiri, aturan atau jujukan perkataan dianggap penting bagi keperluan penghasilan sesebuah ringkasan.

Bagi merumuskan bahagian ini, pendekatan perwakilan kompak MFS dan CP boleh dikatakan menarik dan telah pun digunakan dalam kajian-kajian terdahulu. Lantaran itu, kajian ini mencadangkan agar kekuatan hubung kait setiap jujukan istilah juga patut dikaji dengan mengeksploitasi penggunaan Corak Tersusun Kerap (FSP) sebagai perwakilan teks. Selain daripada itu, bahawa model peringkasan berasaskan corak mempunyai faedah untuk mengaitkan hubungan antara perkataan dengan memelihara kalimat semantik. Perwakilan teks berasaskan corak juga telah menghasilkan keputusan yang lebih baik berbanding model bahasa yang ada seperti yang telah dibincangkan dalam bahagian sorotan kajian untuk menyokong penyelidikan kajian ini.

KERANGKA KERJA MODEL PERINGKASAN TEKS DWIBAHASA

Kajian ini memperkenalkan model peringkasan teks dwibahasa MYTextSumBASIC menggunakan perwakilan teks pertumbuhan-corak tersusun FASP sebagai fitur utama untuk mengekstrak maklumat penting dan relevan bagi menghasilkan ringkasan secara automatik. Perwakilan FASP ini dijana tanpa penyeliaan, iaitu berdasarkan corak data daripada teks dan ia tidak memerlukan rujukan dari sumber maklumat luaran mahupun interpretasi linguistik. Rangka kerja model digambarkan dalam Rajah 2 di mana ianya melibatkan tiga fasa utama iaitu:

1. Fasa 1: Pengumpulan data dan pembangunan korpus ringkasan bahasa Melayu,
2. Fasa 2: Pembangunan model Peringkasan Teks Dwibahasa (MYTextSumBASIC) menggunakan FASP dengan Fitur Kekangan Corak Tekstual,
3. Fasa 3: Penilaian Ringkasan.



RAJAH 2. Kerangka kerja model MYTextSumBasic menggunakan FASP dengan Fitur Kekangan Corak Tekstual

PENGUMPULAN DATA DAN PEMBANGUNAN KORPUS

Fasa pertama adalah prosedur mengumpul data ringkasan untuk membangunkan korpus artikel berita bahasa Melayu mengikut langkah penyediaan data DUC 2002 bagi artikel berita bahasa Inggeris. Artikel-artikel tersebut telah dimuat turun daripada arkib laman sesawang berita Bernama Library and Infolink Service (BLIS) di mana sumber berita adalah dari Berita Harian, Utusan Malaysia dan Bernama. Bagi mengelak daripada meringkaskan artikel yang terlalu pendek, panjang artikel-artikel berita tersebut hendaklah melebihi 200 patah perkataan. Kesemua 100 artikel tersebut diberi kepada tiga orang ahli panel pakar linguistik bagi menghasilkan ringkasan ekstraktif secara manual sebagai rujukan dalam proses penilaian peringkasan nanti. Sebanyak 300 artikel ringkasan rujukan telah dikutip dalam fasa ini dimana ia telah digunakan untuk membangunkan korpus bahasa Melayu untuk kajian ini (rujuk Fasa 1 di dalam Rajah 2).

SET DATA BAHASA MELAYU

Set data bahasa Melayu bagi kajian ini mengandungi 100 buah artikel berita dari dua domain berlainan iaitu Bencana Alam dan Peristiwa. Kata kunci spesifik tajuk berita digunakan seperti "*gempa bumi Kinabalu*" dan "*pesawat MH370 hilang*" untuk mencari artikel-artikel yang berkaitan. Bagi mengelakkan pensampelan data yang tidak seimbang, 10 artikel berita tertinggi telah dipilih secara rawak untuk set data bagi setiap tajuk untuk kajian ini. Setiap domain mengandungi 50 buah artikel, dimana jumlah perkataan unik yang ditemui adalah 11,989 dari dengan jumlah saiz 40,917 patah perkataan. Setiap artikel mempunyai purata 407 patah perkataan dengan jumlah 1,883 potong ayat. Perincian bagi set data bahasa Melayu ini ditunjukkan dalam Jadual 1. Sebagai contoh, dalam domain Bencana Alam, lima set data daripada tiga tajuk utama Banjir Kuala Krai (BKK), Tanah Runtuh Cameron (TRC) dan Gempa Bumi Kinabalu (GBK) dimuat turun dengan label BKK_1, BKK_2, TRC, GBK_1, GBK_2. Bagi set data Peristiwa, tajuk-tajuknya adalah berkenaan kehilangan pesawat MH370 sebanyak 3 set data iaitu dilabel sebagai MH370_1, MH370_2, MH370_3, penembakan pesawat MH17 dilabel MH17 dan kehilangan kapal tangki MT Orkim Harmony di Malaysia dilabel KMH.

JADUAL 1. Perincian set data Bahasa Melayu

Domain	**Bencana Alam**	**Peristiwa**	**Jumlah**
Bilangan artikel	50	50	100
Jumlah perkataan unik	5826	6163	11,989
Set data artikel	BKK_1, BKK_2, TRC, GBK_1, GBK_2	MH370_1, MH370_2, MH370_3, MH17, KMH	10 set
Kata Kunci Tajuk Berita	banjir kuala krai, gempa bumi Kinabalu, tanah runtuh Cameron	mangsa pesawat MH17, pesawat MH370 hilang, pampasan waris MH370, kapal Malaysia hilang	6 tajuk berita
Jumlah perkataan			40,917
Jumlah ayat			1,883
Purata ayat per-artikel			407

SET DATA BAHASA INGGERIS

Kajian ini juga menggunakan set data tanda aras bahasa Inggeris untuk menguji prestasi ringkasan yang dihasilkan menggunakan perwakilan teks FASP. Artikel-artikel yang dipilih adalah daripada set pertama dokumen data tanda aras English DUC 2002 yang boleh dirujuk perinciannya dalam Jadual 2. Set data bahasa Inggeris mempunyai 55,475 patah perkataan, 2,611 potong ayat dan diambil daripada 102 buah artikel berita. Terdapat lima set data topik bagi domain Bencana Alam dan lima lagi set data topik berita untuk satu Peristiwa. Secara purata, setiap set mengandungi 10 artikel berita yang berkaitan dengan topik-topik tertentu seperti "*Hurricane Gilbert*" dan "*The Eruption of Mt. Pinatubo in the Philippines*".

JADUAL 2. Perincian set data Bahasa Inggeris

Domain	**Bencana Alam**	**Peristiwa**	**Jumlah**
Bilangan artikel	50	52	102
Jumlah perkataan unik	6631	5826	12,457
Set data artikel	d077, d089, d097, d103, d109	d080, d084, d095, d101, d116	10 set
Jumlah perkataan			55,475
Jumlah ayat			2,611

Seterusnya, fasa kedua kajian ini adalah kerja-kerja merekabentuk dan membangunkan model peringkasan teks dwibahasa (MYTextSumBASIC). Aliran proses model ini digambarkan dalam Fasa 2 Rajah 2 di mana ia melibatkan proses-proses berikut: 1) Pra-pemprosesan, 2) Perwakilan teks menggunakan FASP dengan Fitur Kekangan Corak Tekstual, 3) Penetapan Skor Ayat, 4) Pemilihan Ayat dan Pemeringkatan dan 5) Penjanaan ringkasan secara automatik.

1) Pra-pemprosesan

Dalam langkah pra-pemprosesan ini, artikel-artikel berita bahasa Melayu yang telah dimuat turun diformat menjadi bentuk XML berdasarkan set data bahasa Inggeris DUC 2001-2002. Bagi setiap fail XML, proses-proses berikut akan dilaksanakan iaitu: 1) Pembahagian setiap terma unik dari koleksi ayat-ayat, 2) Menyingkirkan tanda baca dan tanda koma di antara nombor-nombor, 3) Memisahkan ayat-ayat menggunakan noktah (.) dan 4) Menyingkirkan kata henti perkataan bahasa Melayu. Selepas proses ini, semua terma-terma unik dalam setiap ayat di dalam artikel akan menjalani proses perwakilan teks.

2) Perwakilan teks FASP

Sebuah fitur perwakilan teks bernama *Frequent Adjacent Sequential Pattern* (FASP) telah dibangunkan dengan menggunakan kaedah Pertumbuhan-Corak Tersusun berdasarkan strategi pecah dan perintah. Asas bagi kaedah ini ialah dengan memecahkan set data teks kepada set yang lebih kecil berdasarkan corak kekerapan yang ditemui secara global. Kemudian, corak berikutnya akan di takluk secara berlelaran berdasarkan corak kekerapan setempat. Perbezaan utama antara teknik FASP dan algoritma asal *PrefixSpan* (Pei et al., 2004) adalah pada fasa Penjanaan Corak.

Idea dalam kajian ini adalah dengan memperkenalkan tiga kekangan corak tekstual iaitu: 1) kekangan item kata, 2) kekangan kata urutan bersebelahan dan 3) kekangan saiz urutan, sebelum proses penggabungan corak terkerap dijalankan. Berikut adalah Langkah-langkah yang terlibat dalam menjana perwakilan teks FASP.

Langkah 1: Penetapan dan Penjanaan senarai *prefixTerms*

Dalam langkah penetapan, satu set istilah kerap dimana kandungannya adalah terma-terma unik bersaiz 1-urutan ditetapkan dalam senarai *prefixTerms* dan diwakilkan sebagai α. Untuk proses ini, setiap ayat per-dokumen akan disimpan mengikut senarai urutannya untuk mengekalkan aturan urutan perkataan iaitu $s = \{t_1, t_2, \ldots t_k\}$. Seterusnya, nilai sokongan global yang diwakilkan sebagai *gSupp(P)* akan dikira. Di sini, setiap nilai kekerapan terma akan dikira sebagai sebagai satu tanpa menghiraukan berapa kali ia muncul dalam dokumen yang sama mengikut Persamaan 1. Nilai *gSupp(P)* yang dikira akan digunakan sebagai pemberat (w) sesebuah istilah dalam pelaksanaan perwakilan fitur-vektor.

$$w(t_m) = \{nilai\ 1\ jika\ terma\ index\ t_m\ berada\ di\ dokumen\ n, 0\ untuk\ kes\ lain\ \} \quad (1)$$

Sebagai contoh, dengan merujuk kepada Jadual 3, jika ambang sokongan minimum (*min_sup*) diberi nilai dua, ini bermakna setiap istilah unik sepatutnya dijumpai ≥ 2 kali dalam set dokumen secara keseluruhan. Istilah "pesawat" dan "mh370" dalam Jadual 3 kedua-duanya mempunyai nilai kekerapan tiga dan membawa maksud yang istilah-istilah itu terdapat dalam ketiga-tiga dokumen $\{d1, d2, d3\}$ tanpa mengira berapa kali ia muncul dalam setiap dokumen.

Senarai *prefixTerms* boleh ditulis dalam bentuk *"<term>:gSupp(P)"* seperti *{<pesawat>:3; <mh370>:3; <penumpang>:2; <terputus>:2; <hubungan>:2*".

JADUAL 3. Sampel set *prefixTerms* dalam set data MH370

Terms	$gSupp(P)$	***Document Occurrences***
<pesawat>	3	*{d1,d2,d3}*
<mh370>	3	*{d1,d2,d3}*
<kehilangan>	3	*{d1,d2,d3}*
<penerbangan>	3	*{d1,d2,d3}*
<penumpang>	*2*	*{d2,d3}*
<terputus>	2	*{d1,d3}*
<hubungan>	2	*{d1,d3}*

Langkah 2: Penjanaan FASP menggunakan Fitur Kekangan Corak Tekstual

Selepas proses penetapan, langkah seterusnya adalah untuk membahagi ruang carian setiap dokumen berdasarkan senarai *prefixTerms* dalam Langkah 1. Setiap istilah dalam senarai *prefixTerms* digunakan sebagai *prefix* untuk menakluk senarai aturan perkataan seterusnya secara rekursif. Ini dilakukan dengan menggabungkan α yang didapati dari Langkah 1 dengan istilah yang diurutan bersebelahannya daripada senarai ayat-ayat dalam setiap dokumen $t_{(k+1)}$ yang diwakilkan sebagai β dalam kajian ini. Penjanaan FASP untuk setiap dokumen sehingga istilah saiz vektor *m* adalah berpandukan Persamaan 2:

$$FASP_{\mathrm{m}} = \alpha \cup \beta \tag{2}$$

Walau bagaimanapun, sebelum proses penggabungan ini dilakukan, tiga kekangan corak tekstual telah diperkenalkan iaitu kekangan item kata, kekangan kata urutan bersebelahan dan kekangan saiz urutan kata.

1) Kekangan Item Kata

Kekangan item kata merujuk kepada istilah yang digunakan bagi menjana setiap kombinasi FASP. Di mana semua istilah hanyalah terdiri daripada terma-terma dalam set *prefixTerms* α yang ditetapkan tadi. Kekangan item kata ini adalah untuk mengurangkan masalah penjanaan kombinasi susunan corak yang tidak kerap dan untuk mengekalkan satu unjuran kecil set data vektor. Ini bermaksud, untuk menjana FASP, terma yang berada di susunan bersebelahan $t_{(k+1)}$ ataupun β dalam ayat *s* mestilah termasuk sebagai salah satu item dalam senarai *prefixTerms* atau ($\beta \in \alpha$) tadi. $FASP_m$ hanya akan terjana jika nilai β semasa adalah kerap dan tidak lewah dalam dokumen itu. Kekangan item boleh ditulis sebagai $K_{item}(FASP)$. Katakan α adalah *prefixTerms* yang telah ditetapkan, FASP sebagai set corak yang akan dilombong dan *s* sebagai ayat dengan susunan istilah *t*. Persamaan bagi kekangan item kata boleh ditulis sebagai:

$$K_{item}(FASP) \equiv (\forall_t : 1 \leq t \leq len(s); s[t] \in \alpha)$$

di mana bagi setiap istilah *t*, panjangnya sepatutnya sama atau lebih besar daripada 1 dan kurang daripada jumlah panjang saiz ayat susunan *len(s)*. Item *t* dalam aturan *s* juga mestilah salah satu elemen daripada set α. Contohnya, bagi memaparkan FASP 2-urutan yang mengandungi "pesawat mh370", istilah "mh370" tersebut mestilah terlebih dahulu memenuhi kekangan item kata iaitu nilai sokongannya *sup* mestilah melebihi nilai ambang sokongan minimum (min_sup) σ dan memenuhi kekangan item kata.

$$if\ \sup(X) \geq \sigma \wedge K_{item}(X) = true$$

2) Kekangan Kata Urutan Bersebelahan

Kekangan Kata Urutan Bersebelahan merupakan kekangan ke atas aturan susunan kata $(k + 1)$ yang seterusnya, iaitu ia dilakukan berdasarkan aturan sebenar susunan ayat-ayat di dalam setiap dokumen yang dianalisa. Pendekatan asal dalam algorithma *PrefixSpan* masih mengekalkan aturan item yang kerap, tetapi kedudukan item kerap yang seterusnya tidak semestinya di urutan yang bersebelahan. Namun demikian, dalam kajian ini kekangan urutan susunan ini adalah sangat penting apabila melibatkan bahasa tabii. Ini kerana makna bagi setiap frasa bergantung kepada aturan atau urutan susunan perkataan. Kekangan ini dilakukan untuk mengekalkan semantik ayat agar FASP yang dijana boleh dizahirkan sebagai satu fitur set peraturan yang menggariskan isi kandungan sebuah dokumen. Kekangan aturan susunan bersebelahan ditulis sebagai $K_{adj}(FASP)$. Katakan t_{k+1} dalam ayat *s* diwakilikan sebagai β di mana ia akan mengikut aturan perkataan dalam *s*; β perlulah salah satu elemen dalam set p*refixTerms* seperti yang diterangkan dalam kekangan item dan telah disemak untuk setiap susunan penjanaan FASP dengan mengambil kira kekangan kekangan saiz ayat. Persamaan bagi Kata Urutan Bersebelahan boleh ditulis sebagai:

$$K_{adj}(FASP) \equiv (\forall_s : \beta = t_{k+1}; 1 \leq \beta \leq len(s); \beta \in \alpha)$$

Sebagai contoh, FASP 2-urutan bagi istilah "pesawat mh370" di mana awalan α ialah "pesawat" dan β ialah "mh370" dapat dijana berdasarkan susunan ayat berita asal "pesawat mh370 yang hilang…" di dalam dokumen berita tadi. Istilah yang berada di urutan bersebelahan iaitu "mh370" dikatakan berpadanan dengan aturan susunan FASP apabila ia memenuhi syarat kekangan $K_{adj}(FASP) = true$ tadi.

3) Kekangan Saiz Urutan.

Kekangan saiz urutan merujuk kepada had kepanjangan bagi sebuah corak seperti jumlah maksimum atau minimum istilah dalam setiap corak. Panjang FASP telah ditetapkan dalam kajian ini berdasarkan saiz corak signifikan bi-gram atau trigram dalam bidang Dapatan Kembali Maklumat di mana kekangan saiz adalah bersamaan 2-urutan dan 3-urutan FASP. Tujuan mengehadkan saiz FASP adalah untuk mempercepatkan proses perlombongan dan untuk menampung penemuan signifikan corak tekstual. Kekangan saiz urutan diwakili sebagai $K_{len}(FASP)$. Sebagai contoh, untuk mencari sehingga 3-urutan FASP atau trigram daripada dokumen-dokumen teks tersebut, ianya boleh ditulis sebagai:

$$K_{len}(FASP) \equiv (len(FASP) \geq 3)$$

Langkah 3: Penemuan fitur FASP

Dalam langkah ini, $FASP_m$ yang tidak lewah daripada setiap vektor dokumen akan dimasukkan ke dalam keseluruhan vektor istilah FASP secara tersusun untuk mengekalkan aturannya. Seluruh koleksi dokumen diwakili oleh sebuah matriks *NxM* di mana *N* ialah bilangan dokumen dan *M* adalah susunan vektor FASP yang dijana daripada kesemua dokumen. Dalam langkah ini, FASP yang ditemui digunakan sebagai fitur dalam VSM untuk mewakili setiap dokumen di mana FASP diberi nilai pemberat tertentu seperti Boolean, TF-IDF dan sokongan. Jadual 4 menunjukkan contoh perwakilan dokumen menggunakan FASP dengan pemberat Boolean.

JADUAL 4. Sampel set data MH370 menggunakan perwakilan FASP dengan pemberat Boolean

FASP	pesawat	mh370	penumpang	pesawat mh370	terputus hubungan	penumpang pesawat	…	*M*
d_1	1	1	0	1	1	0	…	....
d_2	1	1	1	1	0	1	…	…
d_3	1	1	1	1	1	1	…	…
N	…	…	…	…	…	…	…	…

Perbandingan prestasi FASP diantara model BOW dan N-gram telah dinilai menggunakan set data bahasa Melayu dan bahasa Inggeris dalam (Alias, Mohammad, Hoon, & Ping, 2018) untuk menjalankan tugas mencari persamaan ayat dan dokumen. Dapatan tersebut telah menyokong penggunaan model perwakilan teks FASP sebagai fitur utama untuk mengekstrak maklumat penting dan relevan bagi menghasilkan ringkasan bagi kajian ini.

Sebagai contoh, jika pengguna menetapkan nilai ambang sokongan minimum *min_sup* σ dan ambang keyakinan minimum *min_conf* δ, sebuah Peraturan Susunan (*Sequential Rule*) $X \rightarrow Y$ yang memenuhi syarat σ dan δ boleh dijanakan. Peraturan susunan $X \rightarrow Y$ adalah hubungan susunan di antara dua set susunan atau istilah. Dalam kajian ini, susunan $X \rightarrow Y$ adalah dalam urutan susunan bersebelahan mengikut penjanaan FASP. Sokongan *support* peraturan $X \rightarrow Y$ adalah bilangan kekerapan urutan susunan itu didapati dalam set data dokumen ataupun *gSupp(X)*. Nilai keyakinan *Confidence* atau *Conf* dalam sesebuah peraturan adalah bilangan urutan susunan yang mempunyai $X \cup Y$ dibahagikan dengan bilangan sokongan urutan susunan yang mempunyai X. Peraturan Susunan ditakrifkan dalam Persamaan 3 seperti berikut:

$$\begin{gathered} support\ (X \rightarrow Y) = support\ (X \cup Y) \\ and\ conf(X \rightarrow Y) = support\ (X \cup Y)/support(X) \\ if\ conf\ (X \rightarrow Y\ ) \geq min_{conf} \geq, then\ X \rightarrow Y\ is\ a\ Sequential\ Rule \end{gathered} \tag{3}$$

Jadual 5 menunjukkan beberapa contoh FASP berserta nilai sokongan dan keyakinan yang ditemui dalam set data MH370_1 dengan nilai *min_sup* ≥ 2 dan nilai *min_conf* ≥ 0.5. Jumlah artikel dalam set data MH370_1 ialah 10.

JADUAL 5. Penemuan FASP daripada set data bahasa Melayu MH370_1 berserta nilai sokongan dan keyakinan

FASP	**Support**	**Confidence *Conf***
kehilangan → pesawat	10	10/10 = 1.0
kehilangan pesawat → mh370	7	7/10 = 0.7
penumpang	8	8/8 = 1.0
penumpang → pesawat	4	4/8 = 0.5
penumpang pesawat → mh370	2	2/4 = 0.5
anak	5	5/10 = 0.5
anak → kapal	3	3/5 = 0.6
anak kapal → mh370	2	2/3 = 0.67

Sebagai contoh, *X* mewakili istilah "*anak*" yang mempunyai nilai sokongan *Support* sebagai 5, yang bermaksud istilah itu didapati dalam 5 dari 10 artikel dalam set MH370_1. Sokongan untuk FASP 2-urutan "*anak → kapal*" yang diwakili oleh *Y* ialah 3. Ini bermakna istilah tersebut didapati dalam 3 artikel daripada set tersebut (tanpa mengambil kira berapa kali ia muncul di dalam ayat-ayat artikel). Ini bermakna, jika *prefixTerms* ialah istilah "*anak*", ia boleh disimpulkan dengan nilai keyakinan 0.6 (*Conf*) bahawa istilah "*kapal*" boleh hadir secara urutan bersebelahan di dalam sesebuah ayat. Di mana FASP tersebut dapat mewakili maklumat yang signifikan dan menarik dalam set data masing-masing.

Satu lagi contoh untuk FASP 3-urutan ialah "*anak kapal mh370*" dengan sokongan 2. Dengan menggunakan peraturan susunan, *X* sekarang mewakili FASP bagi "*anak kapal*" dan Y mewakili "*anak kapal mh370*". Nilai keyakinan bagi peraturan "*anak kapal*" → "*anak kapal mh370*" ialah 0.67. Ini bermaksud, jika istilah "*anak*" dan "*kapal*" muncul bersama, terdapat kemungkinan keyakinan sebanyak 0.67 *Conf* yang istilah "*mh370*" juga akan turut muncul (secara urutan bersebelahan) di dalam sesebuah ayat. Berdasarkan penemuan ini, ia boleh disimpulkan bahawa corak tekstual menarik yang ditemui menggunakan kaedah FASP boleh digunakan sebagai fitur untuk mewakili dokumen atau ayat bagi menyokong penggunaannya dalam model MYTextSumBasic. Tambahan pula, eksperimen terdahulu menggunakan set data bahasa Inggeris telah menunjukkan bahawa perwakilan FASP juga tidak bergantung pada jenis bahasa.

3) Penetapan Skor Ayat

Kajian meluas dalam bidang ATS telah dijalankan bagi mencari kombinasi fitur ayat terbaik dan optimum dalam menetapkan skor ayat. Walau bagaimanapun, kajian ini tertumpu pada empat fitur asas ayat yang secara konsistennya dianggap penting dan signifikan berdasarkan kajian terdahulu dan terkini. Fitur-fitur tersebut adalah seperti kedudukan ayat, ayat yang paling banyak persamaan dengan tajuk dokumen, ayat yang mempunyai frasa kata kunci yang kerap, dan ayat-ayat yang hampir serupa. Dalam model MYTextSumBasic ini, proses penetapan skor telah menggunakan kombinasi linear berikut: 1) fitur permukaan, 2) fitur isi kandungan dan 3) fitur kerelevanan ayat. Sumbangan dalam kajian ini ialah dalam pengiraan skor bagi fitur isi kandungan dan fitur kerelevanan ayat yang improvasi dengan menggunakan perwakilan FASP.

a) Fitur Permukaan

Fitur permukaan adalah sifat dokumen atau ayat. Dalam contoh ini, fitur permukaan sepotong ayat akan dinilai berdasarkan ayat-ayat yang banyak persamaan dengan tajuk, kedudukan ayat dalam dokumen dan panjang ayat.

$$f_1 : Pertindihan\ Ayat\ Dengan\ Tajuk$$

Secara umum, tajuk sesebuah dokumen menggambarkan tema dan isi kandungan dokumen itu. Ayat yang mempunyai istilah-istilah yang sama atau bertindih dengan tajuk dianggap penting dan berkaitan dengan tema utama. Bagi ayat *s*, skor untuk fitur f_1 telah dinormalisasi menggunakan saiz panjang tajuk seperti yang ditunjukkan dalam Persamaan 4:

$$Skor_{f_1}(s) = \frac{\sum_{i=1}^{n} t_i \cap Ayat\ Tajuk}{Panjang\ Ayat\ Tajuk} \tag{4}$$

$f_2 : Kedudukan\ Ayat$

Bagi menetapkan skor untuk fitur kedudukan ayat, kajian ini telah menilai kekerapan setiap ayat berdasarkan kedudukannya dalam artikel dengan merujuk kepada ringkasan yang dihasilkan oleh panel pakar bahasa secara manual. Skor bagi fitur kedudukan ayat telah dibulatkan kepada nilai lebih rendah mengikut kedudukannya dan telah dinormalisasi berdasarkan jumlah artikel dalam korpus ringkasan bahasa Melayu, iaitu 100. Keputusannya ditunjukkan dalam Jadual 6. Kedudukan ayat pertama didapati bahawa ayat tersebut telah termasuk dalam ayat ringkasan sebanyak 92 kali dan telah memberi skor 0.9 kepada fitur f_2 dalam Persamaan 5. Dapat juga diperhatikan bahawa hanya kedudukan tiga teratas ayat dalam kajian ini yang muncul secara signifikan manakala kedudukan ayat terakhir didapati tidak kerap disertakan oleh ahli panel (hanya 37 kali) didalam penyediaan ringksan mereka. Dapatan

ini adalah bercanggah dengan penemuan dalam kajian terdahulu yang menggunakan peringkas bahasa Inggeris oleh (Ferreira et al., 2013).

Melalui pemerhatian kajian ini, telah didapati bahawa ahli panel telah memastikan ringkasan mereka mengandungi topik dokumen dengan menyertakan maklumat signifikan yang ditemui dalam bahagian awal dokumen tersebut. Maklumat signifikan atau relevan yang lain kemudiannya ditambah bagi memenuhi syarat untuk menghasilkan ringkasan ekstraktif dengan saiz kepanjangan 30% daripada artikel sebenar. Selain proses peringkasan secara manual, ahli panel juga melakukan pemampatan ayat ke atas ayat-ayat ringkasan yang dipilih untuk mengeluarkan unsur yang tidak perlu bagi mengekalkan kesinambungan dan kebolehbacaan ringkasan itu. Oleh demikian, dapat diperhatikan dalam Jadual 6 bahawa nilai skor kedudukan ayat lain (termasuk ayat terakhir) adalah rendah kerana kekerapan yang rendah dan kedudukan ayat yang berlainan dipilih oleh ahli panel.

JADUAL 6. Nilai skor ayat yang telah dinormalisasi mengikut kedudukan ayat dalam korpus ringkasan bahasa Melayu

Kedudukan ayat	**Kekerapan**	**Nilai Normalisasi** $Score_{f_2}$
Ayat pertama	92	0.9
Ayat kedua	85	0.8
Ayat ketiga	83	0.7
Ayat terakhir	37	0.2
Ayat lain-lain	20	0.2

$$Skor_{f_2}(s) = Skor\ Normalisasi\ Kedudukan\ Ayat \tag{5}$$

b) Fitur Isi Kandungan

Fitur isi kandungan merangkumi perkataan-perkataan utama yang menyamai kandungan dalam dokumen. Contoh perkataan yang mengandungi isi kandungan yang paling kerap digunakan adalah perkataan-perkataan yang mengandungi topik/tema, istilah pengenalan khas, kata kunci perkataan dengan kekerapan tinggi dan perkataan sentroid.

$$f_3 : Pertindihan\ FASP\ dengan\ ayat$$

Intuisi kajian ini merujuk kepada kajian oleh (Nenkova & McKeown, 2012), di mana ayat-ayat yang mengandungi perkataan yang kerap atau topik pengenalan khas boleh mempengaruhi faktor pemilihan ayat tersebut kerana ia memberi gambaran subjek utama dalam teks tersebut. Dalam kes ini, FASP merupakan salah satu fitur yang mewakili maklumat signifikan atau penting dalam ayat tersebut. Untuk tujuan itu, apabila lebih banyak FASP yang terkandung dalam sesebuah ayat maka lebih tinggi nilai skor ayat itu, seterusnya menunjukkan kepentingan ayat tersebut. Skor itu kemudiannya dinormalisasikan menggunakan saiz kepanjangan ayat. Di sini, FASP dijana sehingga kekerapan 3-urutan. Skor untuk fitur ayat f_3 ditunjukkan dalam Persamaan 6 sebagai:

$$Skor_{f_3}(s) = \frac{\sum_{i=1}^{n} t_i \cap FASP}{Panjang\ ayat\ N} \tag{6}$$

c) Fitur Kerelevanan ayat

Fitur kerelevanan ayat mengeksploitasi hubungan antara ayat-ayat dalam sesebuah dokumen. Ayat-ayat yang mengandungi persamaan dengan ayat yang mempunyai skor tertinggi dianggap sedang membincangkan topik yang sama.

$$f_4: Persamaan\ Ayat$$

Di sini, ayat dengan skor tertinggi dijadikan ayat pendahuluan. Setiap ayat diubah menjadi vektor FASP. Kemudian, nilai kaitan atau kerelevanan antara ayat pendahuluan dan ayat-ayat lain dikira dengan menggunakan formula persamaan kosinus.
Skor bagi fitur ayat f_4 dinyatakan dalam Persamaan 7, di mana $\vec{v}$ mewakili ayat tersebut dan $\vec{w}$ ialah ayat pendahuluan.

$$Skor_{f_4}(s) = sim_{cosine}\ (\vec{v}, \vec{w})$$
$$\frac{\vec{v} \cdot \vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}} \quad (7)$$

Akhirnya, semua skor bagi fitur ayat f_1 hingga f_4 tadi akan dikira untuk setiap ayat di dalam artikel. Jumlah skor untuk setiap ayat ditunjukkan dalam Persamaan 8.

$$Jumlah\ Skor\ Ayat(s) = \sum_{i=1}^{4} Skor\ f_i(s) \quad (8)$$

Kombinasi linear fungsi penetapan skor ayat ditulis sebagai:

$$\therefore Jumlah\ Skor\ Ayat(s) = \sum Skor\ (Tajuk + Lokasi\ Ayat + FASP + Kerelevanan)$$

4) Pemilihan Ayat dan Pemeringkatan

Proses pemilihan ayat-ayat untuk disertakan dalam sesebuah ringkasan adalah bergantung kepada nilai jumlah skor ayat. MYTextSumBasic menggunakan pendekatan rakus dalam memilih ayat berdasarkan jumlah skor setiap ayat daripada jumlah kombinasi linear fitur-fitur ayat dalam Persamaan 8. Skor ayat akan disusun berdasarkan nilai tertinggi untuk memilih ayat-ayat relevan yang mempunyai isi kandungan yang signifikan untuk dimasukkan sebagai ayat dalam ringkasan. Berdasarkan pemerhatian kajian ini terhadap penjanaan ringkasan yang dihasilkan oleh panel pakar bahasa, aliran dan kesinambungan isi kandungan biasanya dikekalkan mengikut kedudukan ayat. Jadi, ayat yang dipilih berdasarkan skor tadi akan dinilai semula pemeringkatannya berdasarkan kedudukan ayat bersebelahan untuk mengekalkan kesinambungan isi kandungan artikel itu.

5) Penjanaan Ringkasan

Selepas proses pemilihan ayat, proses penjanaan ringkasan akan dilakukan berdasarkan saiz ringkasan di mana saiznya mewakili 30% pengekstrakan daripada dokumen asal. Semasa proses penjanaan ringkasan, tiada tatacara pemangkasan dilakukan berdasarkan garis panduan DUC 2002 yang membenarkan had + <=15 perkataan melebihi saiz ringkasan. Ayat ringkasan yang terpilih secara rakus berdasarkan skor tertinggi telah disusun berdasarkan nombor baris ayat untuk mengekalkan kesinambungan berita itu.

Algoritma: MYTextSumBasic(*D, min_sup, min_conf*)

Input: Dokumen D yang mengandungi set ayat-ayat $S=\{s_1,s_2,...s_n\}$, min_sup, min_conf
Output: Ringkasan ss

1: Pra-pemprosesan
2: Penjanaan perwakilan teks FASP
    2.1: Penetapan dan Penjanaan senarai prefixTerm
    2.2: Penjanaan FASP menggunakan fitur kekangan corak tekstual
    2.3: Penemuan fitur FASP
3: Penetapan Skor Ayat
    3.1: Untuk setiap ayat $s_i \in S$ lakukan
    3.2: Tokenisasi s_i menggunakan pembatas ruang
    3.3: $Skor(s_i) = \sum_{i=1}^{4} Skor\ f_i(s_i), \forall_i \in \{Tajuk + Lokasi + FASP + Kerelevanan\}$
4: Pemilihan Ayat dan Pemeringkatan
    4.1: Untuk setiap $Skor(s_i) \in S$ lakukan
    4.2: Pemilihan ayat (s_i) berdasarkan kaedah Rakus
    4.3: Sisih $Skor(s_i)$ secara urutan menurun
5: Penjanaan Ringkasan
    5.1: MaxSS = D saiz * 0.3
    5.2: Selagi $(s_i.saiz) <$ MaxSS
    5.3: Tambah s_i $kepada$ ss
    5.4: Sisih s_i secara ln_i menurun
6: Kembalikan ouput ringkasan ss

RAJAH 3. Algoritma MYTextSumBasic menggunakan perwakilan FASP

Aliran kerja penuh model algoritma MYTextSumBasic menggunakan perwakilan FASP untuk menghasilkan ringkasan ekstraktif secara automatik ditunjukkan dalam Rajah 3. Proses-proses berikut telah dijalankan seperti yang diterangkan secara terperinci di bahagian atas iaitu:

1) Pra-pemprosesan, 2) Perwakilan teks menggunakan FASP dengan Fitur Kekangan Corak Tekstual, 3) Penetapan Skor Ayat Skor dengan menggunakan fitur Tajuk, Lokasi Ayat, FASP dan Kerelevanan, 4) Pemilihan Ayat secara rakus dan Pemeringkatan mengikut urutan ayat dan 5) Penjanaan saiz maksimum ringkasan (patah perkataan) berdasarkan saiz yang ditentukan pengguna iaitu sebanyak 30% daripada saiz asal. Ayat s_i yang dipilih akan ditambah selagi tidak melebihi saiz maksimum MaxSS. Output dari algoritma ini merupakan sebuah ringkasan ekstraktif yang memaparkan maklumat penting dan signifikan dari sesebuah artikel.

PENILAIAN RINGKASAN DAN PERBINCANGAN

Fasa yang terakhir adalah untuk membandingkan ringkasan automatik yang dihasilkan oleh model MYTextSumBasic dengan ringkasan-ringkasan yang dihasilkan secara manual oleh manusia untuk dataset bahasa Melayu dan Inggeris. Kaedah penilaian untuk ringkasan yang dihasilkan oleh MYTextSumBasic dijalankan secara automatik dan manual seperti berikut:

a) **Penilaian Automatik:** ROUGE (Recall-Oriented Understudy for Gisting Evaluation) versi 2.0 digunakan untuk penilaian ringkasan secara automatik dalam kajian ini. ROUGE merupakan tanda aras yang digunakan dalam persidangan-persidangan DUC untuk menilai kualiti sesebuah model dan juga boleh digunakan untuk menilai kualiti model peringkasan dalam pelbagai bahasa. ROUGE akan mengira persamaan di antara ringkasan automatik yang dihasilkan oleh model dengan ringkasan yang dihasilkan oleh pakar bahasa yang

dijadikan sebagai rujukan berasaskan model N-gram dimana nilai ROUGE-1 adalah 1-gram dan ROUGE-2 adalah 2-gram. Penilaian menggunakan nilai ROUGE-1 dilaporkan mempunyai hubung kait tertinggi dengan rumusan yang dihasilkan oleh manusia, dengan nilai keyakinan 95%.

b) **Penilaian Manual:** Kajian ini mengikut garis panduan DUC 2005 di mana penilaian manual kebiasaannya dilakukan oleh ahli panel yang sama. Ringkasan yang dihasilkan akan dinilai berdasarkan tahap kebolehbacaannya dan isi kandungannya ringkasan tanpa perbandingan dengan ringkasan manual yang sedia ada. Terdapat lima kualiti linguistik bagi penilaian kebolehbacaan ringkasan iaitu Q1: Kebolehan mengikut tatabahasa/nahu, Q2: Tidak lewah (duplikasi ayat), Q3: Kejelasan rujukan, Q4: Fokus, Q5: Struktur dan kesinambungan ayat. Penilaian oleh ahli panel adalah berdasarkan skala lima mata Likert iaitu: 1 = Sangat Lemah, 2 = Lemah, 3 = Sederhana, 4 = Bagus dan 5 = Sangat Bagus.

EKSPERIMEN SET DATA BAHASA MELAYU

Berdasarkan garis panduan DUC, untuk tujuan eksperimen ini kami telah menjana sebuah ringkasan dasar sebagai penanda-aras dengan kaedah mengekstrak bilangan ke-*N* ayat terawal daripada sebuah dokumen. Model peringkas ini dinamakan Baseline (Lead) yang melakukan pemilihan ayat melalui kaedah rakus menggunakan ayat-ayat bilangan ke-*N* yang terawal sehingga ke jumlah perkataan yang diperlukan (Jones, 2007) . Kami juga membandingkan pendekatan peringkas MYTextSumBasic dengan peringkas pelbagai bahasa dinamakan OTS yang turut mempunyai bahasa Melayu sebagai salah satu penetapan bahasanya. Peringkas OTS adalah peringkas sumber terbuka yang dibangunkan oleh (Rotem, 2019) yang boleh diakses secara atas talian. Menggunakan penetapan bahasa, kami mengubah pengekod bahasa kepada MS iaitu bahasa Melayu. Model OTS ini menggunakan kaedah pemilihan ayat TF-IDF berdasarkan model N-gram.

Jadual 7 adalah keputusan keputusan ROUGE-1 dan ROUGE-2 di antara model MYTextSumBasic, Baseline (Lead) dan OTS menggunakan set data Bahasa Melayu iaitu sebanyak 100 wacana berita. Ia dapat dilihat bahawa prestasi ringkasan yang dihasilkan oleh MYTextSumBASIC telah mengatasi ringkasan dari model Baseline (Lead) dan OTS dengan nilai purata tertinggi bagi dapatan semula (R) ialah 0.5849, kejituan (P) ialah 0.5736 dan skor-F (Fm) ialah 0.5772 bagi penilaian ROUGE-1. Keputusan bagi ROUGE-2 juga adalah konsisten, ini juga menunjukkan bahawa dengan menggunakan perwakilan teks FASP sebagai fitur dalam pemilihan ayat dalam menghasilkan ringkasan automatik dapat mengekstrak maklumat penting berbanding model peringkas yang lain.

JADUAL 7. Keputusan ROUGE-1 dan ROUGE-2 di antara model MYTextSumBasic, Baseline (Lead) dan OTS menggunakan set data Bahasa Melayu (100 wacana berita)

	ROUGE-1			ROUGE-2		
Model	**R**	**P**	**Fm**	**R**	**P**	**Fm**
Baseline (Lead)	0.5718	0.5653	0.5666	0.4602	0.4508	0.4536
***MYTextSumBasic**	**0.5849**	**0.5736**	**0.5772**	**0.4751**	**0.4593**	**0.4654**
OTS	0.5593	0.5812	0.5356	0.4203	0.3956	0.4052

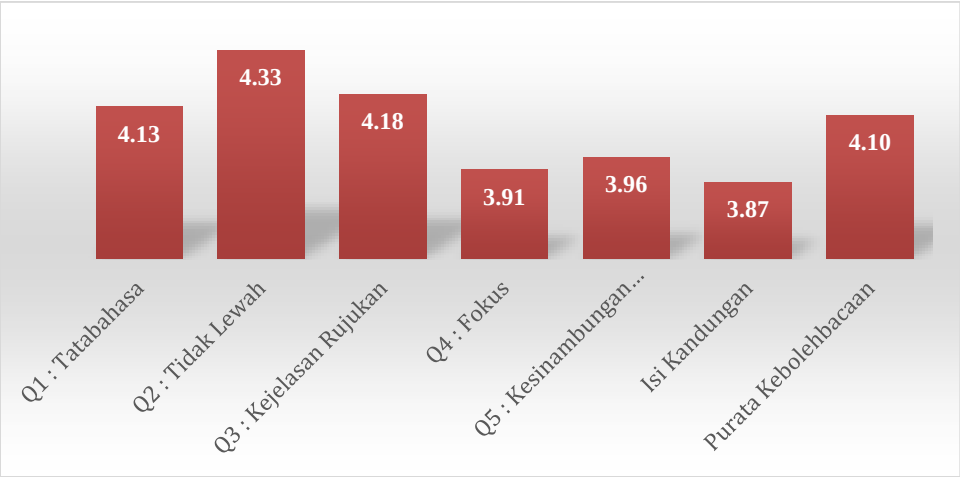
Seterusnya, Jadual 8 menunjukkan keputusan terperinci eksperimen menggunakan set data Bahasa Melayu dengan mengasingkan set data 50 buah artikel dari domain Bencana Alam dan Peristiwa. Manakala, set data Rawak adalah artikel-artikel yang dipilih secara rawak iaitu 30 buah artikel daripada set data Bencana Alam dan Peristiwa. Dapatan dari Jadual 8 menunjukkan keputusan positif MYTextSumBASIC dalam persamaan dengan ringkasan-ringkasan yang dihasilkan oleh pakar bahasa sebagai rujukan. Nilai purata tertinggi bagi

dapatan semula (R) ialah 0.5868, kejituan (P) ialah 0.58 dan skor-F (Fm) ialah 0.5822 bagi set data Peristiwa mengatasi keputusan dari model-model peringkasan yang lain. Ia juga dapat dilihat dalam keputusan penilaian, walaupun menggunakan pendekatan konvensional iaitu hanya mengambilkira ayat-ayat terawal sahaja, ringkasan yang dihasilkan oleh Baseline (Lead) telah mengatasi peringkas dari sumber terbuka iaitu OTS yang menggunakan kaedah TF-IDF.

JADUAL 8. Keputusan ROUGE-1 di antara model MYTextSumBasic, Baseline (Lead) dan OTS menggunakan domain set data Bencana Alam, Peristiwa dan Rawak

Domain Set Data	**Bencana Alam**			**Peristiwa**			**Rawak**		
Model	**R**	**P**	**Fm**	**R**	**P**	**Fm**	**R**	**P**	**Fm**
Baseline (Lead)	0.5756	0.5539	0.5616	0.5679	0.5754	0.5701	0.5637	0.5386	0.5472
***MYTextSumBasic**	**0.5829**	**0.5672**	**0.5722**	**0.5868**	**0.58**	**0.5822**	**0.5813**	**0.5496**	**0.5621**
OTS	0.5614	0.4965	0.5254	0.5571	0.5399	0.5459	0.5551	0.4961	0.5216

Dalam kajian ini, penilaian secara manual ke atas ringkasan-ringkasan yang dijana oleh model MYTextSumBASIC telah dilakukan mengikut garis panduan DUC 2005 di mana ringkasan-ringkasan tersebut akan dinilai secara manual untuk kebolehbacaannya dan isi kandungannya berdasarkan skala lima mata Likert dari 1 hingga 5 oleh pakar bahasa. Merujuk kepada Rajah 4, kaedah MYTextSumBASIC telah menghasilkan skor kebolehbacaan sebanyak 4.1 dan 3.87 untuk isi kandungan ringkasan yang dihasilkan. Ringkasan yang dihasilkan juga mendapat nilai tidak lewah setinggi 4.33 yang menunjukkan tiada pengulangan kepada isi kandungan ringkasan, nilai tatabahasa yang bagus sebanyak 4.13 dan kualiti kesinambungan stuktur ayat setinggi 3.96. Secara rumusan, keputusan ini menunjukkan bahawa kualiti dan tahap kebolehbacaan ringkasan yang dihasilkan secara automatik oleh model MYTextSumBASIC boleh dikategorikan sebagai bagus dan boleh ditambahbaik.



RAJAH 4. Keputusan penilaian ringkasan secara manual model MYTextSumBasic

Seterusnya, sampel artikel berita bahasa Melayu yang asal (Artikel 21) dan ringkasan automatik yang dihasilkan oleh MYTextSumBASIC boleh dirujuk dalam Jadual 9. Ringkasan yang dihasilkan adalah 30% daripada jumlah perkataan asal di dalam artikel tersebut. Antara contoh FASP yang dapat ditemukan dalam artikel tersebut adalah istilah *maklumat radar, pesawat mh370* dan *kehilangan pesawat* yang merupakan antara kandungan utama yang relevan bagi artikel tersebut dan merupakan fitur utama dalam pemilihan ayat ringkasan.

JADUAL 9. Sampel artikel ringkasan yang dihasilkan secara automatik oleh model MYTextSumBasic

Artikel 21	**MYTextSumBASIC**
Kongsi data radar demi jejaki MH370 Tarikh: 14-03-2014 Sepang: Demi menjejaki 227 penumpang dan 12 anak kapal pesawat MH370 Penerbangan Malaysia (MAS) yang masih hilang, Malaysia berkompromi apabila berkongsi maklumat radar milik Angkatan Tentera Malaysia (ATM). Menyedari risiko pendedahan rahsia ketenteraan boleh mengancam keselamatan, Malaysia tetap berkongsi data dengan Federal Aviation Administration (FAA) dan National Transport Security Board (NTSB) demi siasatan menjejaki kehilangan pesawat terbabit. Menteri Pertahanan selaku Pemangku Menteri Pengangkutan, Datuk Seri Hishammuddin Hussein, menegaskan perkongsian data radar terbabit hanya untuk siasatan dan bukan kepada umum. "Namun, kita tidak akan dedahkan data radar kepada umum, kerana ia akan menjejaskan operasi mencari dan menyelamat (SAR)."Pendirian ini bukan bermakna kami cuba menyembunyikan sesuatu, sebaliknya siasatan sedang dijalankan dengan bantuan dan kerjasama pelbagai pihak dalam dan luar negara. "Kita hanya berkongsi maklumat data radar tentera dengan Amerika Syarikat dan China bagi tujuan operasi SAR demi menjejaki MH370," katanya dalam sidang media di Lapangan Terbang Antarabangsa Kuala Lumpur (KLIA). Tidak perlu diragui. Hishammuddin memberi jaminan, keupayaan radar tentera dalam mengawasi pergerakan ruang udara negara ini tidak perlu diragui, kerana keupayaan kelengkapan ATM itu diperakui negara luar dan badan antarabangsa. "Namun, semua radar milik tentera dan penerbangan awam negara ini, tidak menerima sebarang isyarat kecemasan daripada juruterbang MH370, sebelum ia dilaporkan hilang awal pagi Sabtu lalu," tegasnya. Pesawat Boeing 777-200ER itu hilang daripada radar pada jam 1.23 pagi sejurus memasuki ruang udara Vietnam, selepas dibenarkan berlepas dari Lapangan Terbang Antarabangsa Kuala Lumpur (KLIA) menuju Beijing, China mulai 12.41 pagi. Sebelum dilaporkan hilang, MH370 dikesan menghantar isyarat data terakhir menerusi `Aircraft Communications Addressing and Reporting System' (ACARS) kepada pengeluar enjinnya, Roll Royces dan pembuat pesawat, Boeing, pada jam 1.07 pagi. (290 perkataan)	Sepang: Demi menjejaki 227 penumpang dan 12 anak kapal **pesawat MH370** Penerbangan Malaysia (MAS) yang masih hilang, Malaysia berkompromi apabila berkongsi **maklumat radar** milik Angkatan Tentera Malaysia (ATM). Menyedari risiko pendedahan rahsia ketenteraan boleh mengancam keselamatan, Malaysia tetap berkongsi data dengan Federal Aviation Administration (FAA) dan National Transport Security Board (NTSB) demi siasatan menjejaki **kehilangan pesawat** terbabit. "Kita hanya berkongsi **maklumat data radar** tentera dengan Amerika Syarikat dan China bagi tujuan operasi SAR demi menjejaki MH370," katanya dalam sidang media di Lapangan Terbang Antarabangsa Kuala Lumpur (KLIA). (87 perkataan)

EKSPERIMEN SET DATA BAHASA INGGERIS

Eksperimen seterusnya adalah untuk menunjukkan bahawa model MYTextSumBASIC menggunakan perwakilan teks FASP adalah dwibahasa dan mempunyai kelebihan dalam mengenalpasti maklumat terpenting daripada sesebuah teks. Penilaian telah dilakukan dengan menggunakan 102 buah dokumen yang dipilih daripada set pertama dokumen bahasa Inggeris bagi data DUC 2002 iaitu D061j, D062j, D063j, D064j, D065j, D066j, D067f, D068f, D069f, D070f, D071f, D072f dan D073b.

Kajian ini telah mengikut tetapan kajian yang dilakukan dalam kajian terdahulu oleh (Binwahlan, Salim, & Suanmali, 2010) di mana mereka telah mengusulkan model peringkasan yang mengintegrasikan MMR dengan strategi kerumunan kabur untuk mempelbagaikan output ringkasan mereka. Mereka telah membandingkan model mereka dengan keputusan sistem terbaik (sys19) (Harabagiu & Lacatusu, 2002) dan tercorot (sys30) (Zajic, Dorr, & Schwartz, 2002) yang menyertai persidangan DUC 2002.

Keputusan eksperimen diberikan dalam Jadual 10 di mana prestasi terbaik dicapai oleh ringkasan tanda aras yang dihasilkan oleh manusia iaitu (H2-H1). Kaedah M4 yang dihasilkan oleh (Binwahlan et al., 2010) dan model MYTextSumBASIC dilihat telah menghasilkan keputusan yang kompetitif menggunakan skor ROUGE-1 (R1) ke atas purata dapatan semula iaitu 0.43962 berbanding 0.43896. Walau bagaimanapun, salah satu penemuan menarik dari eksperimen ini dapat dilihat pada skor ROUGE-2 (R2) dalam Jadual 10, iaitu prestasi MYTextSumBASIC adalah lebih baik jika berbanding kaedah M4. Keputusan daripada eksperimen ini juga menunjukkan kaedah MYTextSumBASIC berprestasi lebih baik berbanding sistem terbaik dan tercorot di DUC 2002.

Di sini kita dapat membuat kesimpulan bahawa model MYTextSumBASIC yang menggunakan perwakilan fitur FASP tidak dipengaruhi oleh kekangan bahasa, domain dan juga mempunyai potensi besar dalam bidang ATS.

JADUAL 10. Keputusan perbandingan MYTextSumBasic menggunakan set data bahasa Inggeris dengan nilai purata panggil balik ROUGE-1 (R1) dan ROUGE-2 (R2)

Kaedah	**R1**	**R2**
H2-H1 (Tanda Aras)	0.49657	0.20957
M4 (Binwahlan et al., 2010)	0.43962	0.19702
***MYTextSumBasic**	0.43896	0.19918
Sys19 (Terbaik DUC2002)	0.40259	0.1842
Sys30 (Tercorot DUC2002)	0.06705	0.03417

PENGGUNAAN PERWAKILAN FASP DALAM TEKS DWIBAHASA

Sampel penggunaan perwakilan FASP sebagai fitur ayat dalam teks dwibahasa bagi domain Bencana Alam berkaitan topik banjir di berikan dalam Jadual 11. Tanda * merujuk kepada urutan FASP yang terpanjang yang di jana dalam set data tersebut.

JADUAL 11. Sampel FASP bagi set data Bahasa Inggeris dan Bahasa Melayu untuk topik banjir

FASP Bahasa Melayu	**FASP Bahasa Inggeris**
banjir	flood
kilat	flood level
banjir kilat tanah runtuh	flood prevention
berjaya diselamatkan pasukan penyelamat	flood prevention control
kawasan selamat dibuka	flood situation remains
kejadian banjir tanah runtuh berlaku	flood victims
mangsa	masses flood victims
mangsa banjir	*national weather service said
mangsa banjir diselamatkan	operations center
mangsa dipindahkan	reports deaths injuries
mangsa diselamatkan	rescue relief work
*mesyuarat khas bencana banjir kilat tanah	search rescue
operasi mencari menyelamat mangsa	severe thunderstorm warning
pelbagai bencana banjir kilat	thunderstorms
pusat pemindahan banjir	victims

Dari Jadual 11, kajian ini melaporkan bahawa FASP yang ditemui dapat membantu dalam mengenalpasti maklumat utama dan relevan di dalam sesebuah artikel dengan memastikan aliran semantik sesebuah penceritaan dapat dikekalkan. Sebagai contoh, jika topik tersebut adalah berkaitan dengan banjir, maka istilah penting yang relevan yang ditemui sebagai FASP adalah seperti "mangsa banjir" or "*flood victims*" berdasarkan urutan ayat. Selain daripada itu, dalam ringkasan automatik yang akan dijana nanti dapat juga disertakan maklumat penting bahawa terdapat usaha bagi pemuliharaan bencana alam berdasarkan perwakilan FASP "*operasi mencari menyelamat mangsa*" atau "*rescue relief work*" yang ditemui dalam artikel dalam set data tersebut. Oleh yang demikian, dari dapatan ini, kajian ini telah menengetengahkan kepentingan penggunaan perwakilan FASP bersama kekangan corak tekstual sebagai fitur untuk mewakili isi penting sesebuah dokumen atau ayat di dalam bidang peringkasan teks ekstraktif dwibahasa.

KESIMPULAN

Dalam kajian ini, kami telah menerangkan proses pembangunan model peringkasan dwibahasa dinamakan MYTextSumBASIC. Kami telah memperkenalkan satu fitur perwakilan teks dinamakan FASP berdasarkan pertumbuhan-corak tersusun yang telah ditambahbaik dengan kekangan corak tekstual yang dikenali sebagai kekangan item kata, kekangan kata urutan bersebelahan dan kekangan saiz urutan kata bagi mengekstrak maklumat dan ayat yang terpenting untuk disertakan dalam sebuah ringkasan. Model peringkasan MYTextSumBASIC kami telah menunjukkan keputusan yang memberangsangkan di mana ia mengatasi model peringkasan Baseline (Lead) dan OTS dengan persetujuan nilai purata panggil balik sebanyak 58% menggunakan set data bahasa Melayu. Kajian yang dilakukan ke atas peringkas bahasa Inggeris juga menunjukkan keputusan yang memberangsangkan di mana MYTextSumBASIC berjaya mengatasi model-model terdahulu menggunakan data DUC2002.

Ini menunjukkan fitur perwakilan teks FASP bersama kekangan corak tekstual yang digunakan oleh model kami tidak bergantung kepada jenis bahasa yang digunakan dan mampu bersaing secara kompetitif dengan model peringkasan teks yang lain. Model peringkasan ini juga telah diuji penggunaannya oleh pakar bahasa Melayu di peringkat sekolah di mana maklumbalas positif berkaitan kebolehbacaan ringkasan yang dihasilkan sangat membantu pengguna akhir dalam mendapatkan maklumat yang penting secara automatik. Rancangan seterusnya adalah untuk memperluaskan kajian peringkasan teks secara kaedah abstraktif bagi meliputi pelbagai jenis data seperti set data daripada media sosial dan kajian literatur.

RUJUKAN

Alias, S., Mohammad, S. K., Hoon, G. K., & Ping, T. T. (2016). A Malay Text Corpus Analysis for Sentence Compression Using Pattern-Growth Method. *Jurnal Teknologi. 78*(8), 197-206.

Alias, S., Mohammad, S. K., Hoon, G. K., & Ping, T. T. (2018). A text representation model using Sequential Pattern-Growth method. *Pattern Analysis and Applications*. 1-15. doi:10.1007/s10044-017-0624-9

Baralis, E., Cagliero, L., Jabeen, S. & Fiori, A. (2012). *Multi-document summarization exploiting frequent itemsets.* Paper presented at the 27th Annual ACM Symposium on Applied Computing, Trento, Italy.

Binwahlan, M. S., Salim, N. & Suanmali, L. (2010). Fuzzy swarm diversity hybrid model for text summarization. *Information Processing & Management. 46*(5), 571-588.

Boudin, F. & Morin, E. (2013, 2013). *Keyphrase Extraction for N-best reranking in multi-sentence compression.* Paper presented at the North American Chapter of the Association for Computational Linguistics (NAACL).

Clarke, J., & Lapata, M. (2008). Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research, 31*, 399-429.

Conroy, J. M., Schlesinger, J. D., O'leary, D. P. & Goldstein, J. (2006, November). *Back to basics: CLASSY 2006.* Paper presented at the Proceedings of DUC.

Edmundson, H. P. (1969). New Methods in Automatic Extracting. *Journal of the ACM (JACM). 16*(2), 264-285. doi:10.1145/321510.321519

Ferreira, R., de Souza Cabral, L., Freitas, F., Lins, R. D., de França Silva, G., Simske, S. J., & Favaro, L. (2014). A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications. 41*(13), 5780-5787.

Ferreira, R., de Souza Cabral, L., Lins, R. D., e Silva, G. P., Freitas, F., Cavalcanti, G. D. & Favaro, L. (2013). Assessing sentence scoring techniques for extractive text summarization. *Expert Systems with Applications. 40*(14), 5755-5764.

Gambhir, M. & Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review. 47*(1), 1-66. doi:10.1007/s10462-016-9475-9

Ganesan, K., Zhai, C. & Han, J. (2010). *Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions.* Paper presented at the Proceedings of the 23rd international conference on computational linguistics

García-Hernández, R. A. & Ledeneva, Y. (2009). *Word Sequence Models for Single Text Summarization.* Paper presented at the 2009 Second International Conferences on Advances in Computer-Human Interactions.

Harabagiu, S. M. & Lacatusu, F. (2002, July). *Generating single and multi-document summaries with gistexter.* Paper presented at the Document Understanding Conferences.

Jones, K. S. (2007). Automatic summarising: The state of the art. *Information Processing & Management. 43*(6), 1449-1481. doi:10.1016/j.ipm.2007.03.009

Jusoh, S., Masoud, A. M. & Alfawareh, H. M. (2011). Automated text summarization: sentence refinement approach. In P. J. Snasel V., El-Qawasmeh E. (Ed.), *Digital Information Processing and Communications. Communications in Computer and Information Science* (Vol. 189, pp. 207-218): Springer, Berlin, Heidelberg.

Khan, A., Salim, N., Reafee, W., Sukprasert, A. & Kumar, Y. J. (2015). A Clustered Semantic Graph Approach For Multi-Document Abstractive Summarization. *Jurnal Teknologi, 77*(18).

Kim, H. D., Park, D. H., Lu, Y. & Zhai, C. (2012). Enriching text representation with frequent pattern mining for probabilistic topic modeling. *Proceedings of the American Society for Information Science and Technology. 49*(1), 1-10. doi:10.1002/meet.14504901209

Le, Q. V. & Mikolov, T. (2014). *Distributed representations of sentences and documents.* Paper presented at the Proceedings of the 31st International Conference on Machine Learning (ICML-14).

Ledeneva, Y., Gelbukh, A. & García-Hernández, R. (2008). *Terms Derived from Frequent Sequences for Extractive Text Summarization.* Paper presented at the International Conference on Intelligent Text Processing and Computational Linguistics.

Litvak, M. & Last, M. (2013). Cross-lingual training of summarization systems using annotated corpora in a foreign language. *Information Retrieval. 16*(5), 629-656.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development. 2*(2), 159-165. doi:10.1147/rd.22.0159

M. Denil, A. D., N. Kalchbrenner, P. Blunsom, N. de Freitas. (2014). *Modelling, Visualising and Summarising Documents with a Single Convolutional Neural Network.* Paper presented at the 26th Int. Conf. Computational Linguistics.

Mahajani, A., Pandya, V., Maria, I. & Sharma, D. (2019). A Comprehensive Survey on Extractive and Abstractive Techniques for Text Summarization. In T. S. Hu YC., Mishra K., Trivedi M. (Ed.). Y.-C. Hu, S. Tiwari, K. K. Mishra, & M. C. Trivedi (Series Eds.), Ambient Communications and Computer Systems, Advances in Intelligent Systems and Computing Ambient Communications and Computer Systems (Vol. 904, pp. 339-351): Springer Singapore.

Narayan, S., Cohen, S. B. & Lapata, M. (2018). *Ranking sentences for extractive summarization with reinforcement learning.* Paper presented at the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Nenkova, A. & McKeown, K. (2011). Automatic Summarization. *Foundations and Trends® in Information Retrieval. 5*(2–3), 103-233. doi:10.1561/1500000015

Nenkova, A. & McKeown, K. (2012). A survey of text summarization techniques. In Charu C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 43-76): Springer.

Nenkova, A. & Vanderwende, L. (2005). The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101*.

Ning, Z., Yuefeng, L. & Sheng-Tang, W. (2012). Effective Pattern Discovery for Text Mining. *Knowledge and Data Engineering, IEEE Transactions. 24*(1), 30-44. doi:10.1109/TKDE.2010.211

Noah, S. A. M., Ali, N. M. & Hasan, M. S. (2018). Penjanaan Ringkasan Isi Utama Berita Bahasa Melayu berdasarkan Ciri Kata (Generation of News Headline for Malay Language based on Term Features). *GEMA Online® Journal of Language Studies. 18*(4).

Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q. & Hsu, M.-C. (2004). Mining Sequential Patterns by Pattern-Growth: The PrefixSpan approach. *IEEE Transactions on Knowledge and Data Engineering. 16*(11), 1424-1440.

Qiang, J.-P., Chen, P., Ding, W., Xie, F. & Wu, X. (2016). Multi-document summarization using closed patterns. *Knowledge-Based Systems. 99*, 28-38.

Rotem, N. (2019). Open Text Summarizer (OTS). Retrieved from http://libots.sourceforge.net/

Van Lierde, H. & Chow, T. W. (2019). Query-oriented text summarization based on hypergraph transversals. *Information Processing & Management. 56*(4), 1317-1338.

Verma, V. K., Yadav, A. & Jain, T. (2019). *Key Feature Extraction and Machine Learning-Based Automatic Text Summarization.* Paper presented at the Emerging Technologies in Data Mining and Information Security. Advances in Intelligent Systems and Computing.

Xie, F., Wu, X. & Zhu, X. (2017). Efficient sequential pattern mining with wildcards for keyphrase extraction. *Knowledge-Based Systems. 115*, 27-39.

Zajic, D., Dorr, B. & Schwartz, R. (2002). *Automatic headline generation for newspaper stories.* Paper presented at the Workshop on Automatic Summarization.

Zamin, N. & Ghani, A. (2010, 2010). *A Hybrid Approach for Malay Text Summarizer.* Paper presented at the Proceedings of the International Multi-Conference on Engineering and Technological Innovation.

PENULIS

Suraya Alias (Ph.D) merupakan pensyarah Kanan di Fakulti Komputeran dan Informatik, UMS. Bidang kajian beliau adalah dalam Perlombongan Data dan perwakilan Teks. Pengkhususan dalam Peringkasan Teks Bahasa Melayu dan Inggeris.

Mohd Shamrie Sainin (Ph.D) merupakan pensyarah Kanan di Fakulti Komputeran dan Informatik, UMS. Bidang kajian beliau adalah dalam Pembelajaran Mesin dan Terjemahan Mesin dalam Bahasa Asli (KadazanDusun).

Siti Khaotijah Mohammad (Ph.D) merupakan pensyarah Kanan di Pusat Pengajian Sains Komputer, USM. Bidang kajian beliau adalah dalam linguistik dan leksikal Bahasa Melayu.

eISSN: 2550-2131
ISSN: 1675-8021