RINGKASAN TEKS BAHASA MELAYU MENGGUNAKAN PEMBENAMAN PERKATAAN

IERA AMIERA BINTI ZULKEFLY
DR. SAIDAH BINTI SAAD

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Ringkasan teks secara automatik adalah satu teknik yang mengaplikasi Pengkomputeran Bahasa Tabii yang mampu memberi impak yang besar dalam kehidupan seharian kita. Ringkasan teks ini memudahkan proses memahami teks yang ditulis. Ringkasan teks atau artikel merupakan proses dimana dokumen yang lebih ringkas dan berinformasi dihasilkan berdasarkan dokumen teks asal tanpa mengubah makna dokumen asli . Ringkasan teks secara umumnya boleh dibahagikan kepada dua kategori iaitu ringkasan jenis ekstraktif dan ringkasan jenis abstraktif. Secara asasnya tujuan peringkasan teks dan artikel ini adalah untuk menghasilkan satu teks ringkas dan berinformasi, maka penentuan teknik yang betul adalah sangat penting. Kajian ini dilakukan bagi mengimplementasikan teknik pembenaman kata di dalam proses ringkasan teks dari sumber dokumen berita Bahasa Melayu. Objektif kajian ialah untuk mengekstrak informasi yang releven dalam sesuatu dokumen atau artikel kepada ringkasan yang padat. Pengkajian ini dilakukan dengan meringkaskan korpus teks Bahasa Melayu yang diperolehi daripada sumber arkib berita BERNAMA. Teks yang telah diringkaskan ini kemudiannya akan dibandingkan dengan ringkasan yang dilakukan secara manual. Hasil daripada perbandingan menunjukkan hasil yang baik iaitu dengan dengan nilai dapatan bagi *recall* sebanyak 0.3 manakala bagi *precision* sebanyak 0.138 dan *F-measure* sebanyak 0.179. Hasil daripada kajian ini diharap dapat menyumbang kepada pengetahuan mengenai proses peringkasan Bahasa Melayu khususnya,seterusnya dapat membantu para penyelidik dan pengguna untuk meringkaskan teks yang panjang. Metodologi yang digunakan untuk kajian ini ialah *extreme programming Extreme programming* ialah suatu model yang termasuk dalam pendekatan agile yang cenderung kepada pendekatan Object-Oriented. Kajian mendapati bahawa teknik ini boleh digunakan untuk meringkaskan teks ringkasan Bahasa Melayu. Walau bagaimanapun skor dan ketepatan boleh ditingkatkan dengan melakukan proses pemprosesan teks yang lebih teliti dan menggunakan lebih banyak dataset variasi artikel.

1 PENGENALAN

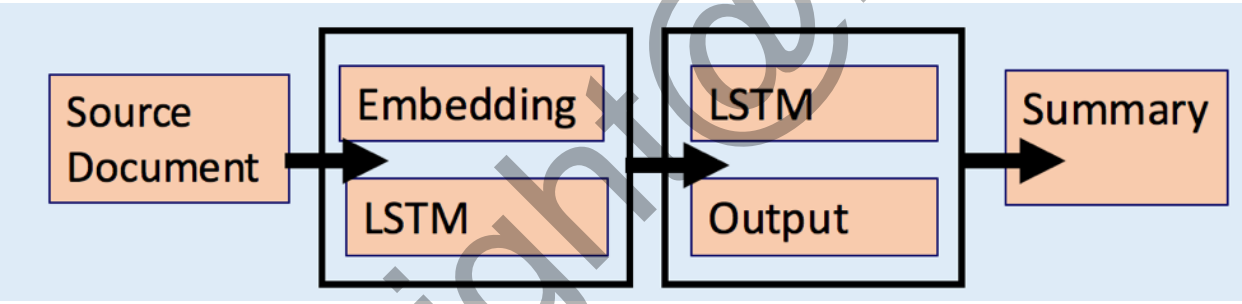
Ringkasan teks adalah satu aplikasi Pengkomputeran Bahasa Tabii yang mampu memberi impak yang besar dalam kehidupan seharian kita. Di era yang penuh dengan ledakan maklumat di hujung jari ini semua orang inginkan sesuatu yang memudahkan mereka. Tambahan pula dengan kekangan masa , tiada siapa mempunyai masa untuk membaca satu persatu artikel, dokumen dan buku untuk menentukan sama ada perkara tersebut penting atau tidak. Teknologi ringkasan automatik dokumen ini akan mampu untuk menyelesaikan masalah kepada ledakan maklumat yang berlebihan.

Ringkasan teks atau artikel merupakan proses dimana dokumen yang lebih ringkas dan berinformasi dihasilkan berdasarkan dokumen asal tanpa mengubah makna dokumen asal . Ringkasan teks secara umumnya boleh dibahagikan kepada dua kategori iaitu ringkasan jenis ekstraktif dan ringkasan jenis abstraktif.

Ringkasan jenis ekstraktif , kaedah ini berfungsi dengan mengekstrak beberapa bahagian, seperti ungkapan dan ayat-ayat , daripada teks dan dihimpunkan bersama menghasilkan suatu ringkasan. Kaedah ekstraktif ini menggunakan perkataan yang sudah wujud di dalam sesuatu ringkasan tersebut dan memilih beberapa kombinasi perkataan yang memberi makna di dalam teks. Ia termasuk dalam proses memberi skor kepada setiap perkataan dan memilih perkataan yang relevan sahaja. Teknik ini juga membuang perkataan dan ayat yang bertindan. Antara teknik yang biasa digunakan ialah kaedah *Term Frequency-Inverse Document Frequency* (TF-IDF), kaedah berdasarkan kluster, pendekatan *graph theoretic*, pendekatan pembelajaran mesin dan seumpanya adalah merupakan teknik peringkasan ekstraktif.

Ringkasan jenis abstraktif pula merupakan kaedah yang menggunakan teknik pengkomputeran bahasa tabii yang lebih maju untuk meringkaskan keseluruhan teks. Ia menggunakan kaedah linguistik untuk meringkaskan teks dan diterjemahkan kepada teks baharu

yang pendek dengan mengekstrak hanya perkataan yang bermakna dan penting. Pendekatan ini boleh dikelaskan kepada dua kategori iaitu kaedah berasaskan struktur(berdasarkan petua(rule based), berasaskan pepohon, ontologi dan seumpanya) atau berasaskan semantik (kaedah berasaskan item maklumat, kaedah geraf semantik, model perwakilan teks semantik dan seumpanya. Antara teknik abstraktif ini ialah ianya memproses data dengan membina *word embedding* sesuatu perkataan. Input embedding akan menukarkan teks kepada nombor. Sebuah terjemahan nombor dalam bentuk data menggunakan *encoder-decoder network* akan dihasilkan. *Word2vec* dan *Global-Vectors (Glove)* merupakan contoh *word embedding*. *Word2vec* merupakan algoritma yang menggabungkan continuous bag of words *(CBOW)* dan *skip-gram* model untuk menghasilkan perwakilan vektor perkataan. *Glove* pula menggunakan perwakilan vektor perkataan dan melatih perkataan yang sama dari kamus.



Rajah 1.1 *Encoder-decoder recurrent neural network model*

Senibina *encoder-decoder recurrent neural network* telah terbukti efektif apabila diaplikasikan kepada ringkasan penulisan. Senibina ini terdapat dua komponen iaitu *encoder* dan *decoder*. Struktur *encoder* akan membaca semua input dan akan menghasilkan input dan menjana output iaitu ringkasan. *Encoder* dan *decoder sub-model* akan dilatih bersama bermakna kedua duanya diperlukan agar output yang terhasil akan dimasukkan ke dalam struktur lagi sebagai input. Abstraktif bagaimanapun agak lebih kompleks dari kaedah ekstraktif.

2 PERNYATAAN MASALAH

a. Pengenalpastian ayat paling penting dalam sesebuah artikel

b. Penyingkiran maklumat yang tidak relevan dalam artikel

3 OBJEKTIF KAJIAN

Objektif kajian adalah untuk :

a. Mengenalpasti teknik yang akan digunakan bagi proses peringkasan yang akan dijalan.
b. Menguji pengekstrak informasi sesuatu dokumen atau artikel menjadi suatu ringkasan yang padat, signifikan dan bersesuaian.

4 METOD KAJIAN

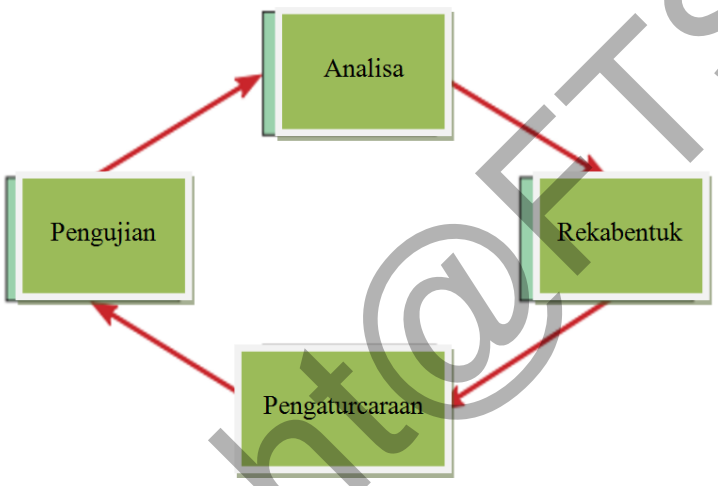
Sistem Pembangunan Metodologi ialah suatu proses standard yang perlu dilakukan untuk menjalankan semua langkah yang penting bagi menganalisa, merekabentuk, implementasi dan untuk mengukuhkan sesuatu sistem.

Metodologi yang digunakan disini ialah metodologi rekabentuk *extreme programming*. Extreme programming terkenal dengan kitaran pembangunan yang pendek, dan pendekatan perancangan *incremental*. Ia menfokuskan ujian automatik yang ditulis oleh programmer untuk memantau proses pembangunan sistem mereka. *Extreme programming* juga mempunyai beberapa kebaikan seperti :

1) Perancangan, analisis, rekabentuk dan pembangunan sistem disatukan di dalam satu fasa aktiviti

2) Caranya yang unik untuk menggambarkan dan menyampaikan keperluan sistem dan spesifikasi rekabentuk.

Kesemua fasa sistem kitaran ini digabungkan di dalam satu siri aktiviti berdasarkan proses asas pengaturcaraan , pengujian, pendengaran dan merekabentuk. Di dalam pendekatan ini, pengaturcaraan dan pengujian berada di dalam bahagian yang berkait dan di dalam proses yang sama. Apa yang ingin disampaikan ialah di fasa ini pengujian yang dilakukan ialah menguji sistem yang mungkin gagal bukan melakukan pengujian kepada keseluruhan pengaturcaraan. Setelah kesemua coding ditulis barulah ia akan diuji semula kesemuanya. Jika kesemua pengujian ke atas pengaturcaraan berjalan dengan jayanya, barulah proses pembangunan dijalankan. Jika tidak, coding akan dibaiki sehingga pengujian dijalankan ke atasnya berjaya.



Rajah 4.1 Proses di setiap fasa pembangunan *extreme programming*

5 HASIL KAJIAN

Setelah mendapatkan hasil ringkasan daripada sistem dan hasil ringkasan manusia, perbandingan diantara kedua-dua ringkasan akan dijalankan. Perbandingan ringkasan dihasilkan menggunakan penilaian Metric ROUGE-N. Metric ROUGE(Recall-Oriented Understudy for Gisting Evaluation)(Lin 2004) telah digunakan secara meluas dalam dunia ringkasan. ROUGE merupakan recall-based metric yang digunakan untuk panjang ringkasan yang telah ditentukan oleh n-gram. Teknik ini akan diguna pakai untuk membandingkan ringkasan penulisan yang telah dijana oleh sistem dan ringkasan penulisan yang ditulis oleh manusia.

ROUGE-N(ROUGE-1,ROUGE-2) tekah digunakan untuk mengira *recall*, *precision* dan *f-measure.* Teknik ini akan membantu untuk mengira persamaan unit di antara ringkasan sistem dan ringkasan manusia di dalam bentuk *unigram* atau *bigrams* seperti di dalam formula di bawah.

$$\text{Recall} = \frac{gram_{ref} \cap grams_{gen}}{grams_{ref}}$$

$$\text{Precision} = \frac{gram_{ref} \cap grams_{gen}}{grams_{gen}}$$

$$\text{F} - 1\ score = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

$$\text{ROUGE} - \text{N} = \frac{\sum_{S \in References\ \ summaries} \ \sum_{N-gram \in S} Count_{match\,(N-gram)}}{\sum_{S \in References\ \ summaries} \ \sum_{N-gram \in S} Count\,(N-gram)}$$

Ringkasan Sistem vs Ringkasan Manusia

Di dalam penilaian ini, ringkasan penulisan manusia telah dijadikan rujukan untuk dibandingkan dengan ringkasan yang dijana oleh sistem. Hasil keputusan adalah seperti di bawah.

Jadual 4.1 Ringkasan Manusia dan Ringkasan Sistem

Dataset	Ringkasan Manusia	Ringkasan Sistem
Dataset 1	Satu berita baik bagi rakyat Malaysia apabila jumlah pesakit covid-19 telah berkurangan dan makin ramai pesakit telah didiscaj. Jumlah pesakit baru telah berkurangan daripada mencecah tiga angka kepada 2 angka. Terkini, jumlah pesakit yang pulih adalah dua kali ganda daripada jumlah kes baharu. Daripada 19 orang pesakit yang positif covid-19, hanya 8 sahaja pesakit yang memerlukan bantuan pernafasan dan tiada kes kematian telah dicatatkan sehingga kini. Namun demikian, Ketua Pengarah Kesihatan menyatakan bahawa Malaysia belum bersedia untuk memasuki pelan pasca krisis COVID-19 kerana jumlah kes di Malaysia masih mencecah dua angka. Malaysia juga masih berada dalam fasa pemantauan kes harian dan mungkin akan berlanjutan selama dua minggu lagi melalui Perintah Kawalan Pergerakan Bersyarat (PKPB).	covid19 jumlah baharu mei dua khamis malaysia positif pesakit pulih kesihatan setakat pukul 12 tengah hari menjadikan 2 semalam mencatat minggu fasa 6 1 4 dicatat peratus keseluruhan kematian harian pelan pasca krisis kuala lumpur bernama melegakan tiga penghujung lepas didiscaj terus meningkat seiring kenyataan kementerian kkm negara pemulihan 7 39 direkod 467 aktif kebolehjangkitan 584 kes rabu 45 selasa 5 30 isnin 55 berbanding 122 ahad 3 105 sabtu sebanyak 74 ganda kumulatif sepenuhnya 776 73 9 seramai 19 dirawat unit rawatan rapi tersebut lapan 8 memerlukan bantuan pernafasan tiada dilaporkan kekal 107 65 ketua pengarah datuk noor hisham abdullah menjawab soalan wartawan sidang media menyatakan memasuki exit strategy rekod berkadar sebarang usaha merangka dibuat bersandarkan garis panduan pertubuhan sedunia who peringkat pemantauan sekurangkurangnya tempoh menerusi pendekatan perintah kawalan pergerakan bersyarat pkpb

Dataset 2	Kerajaan India sedang giat mencari ubatan-ubatan tradisional tempatan sebagai penawar penyakit COVID-19 dan disebabkan India mempunyai sejarah dalam perubatan tradisional maka Kementerian AYUSH yang bertanggungjawab dalam bidang ini sedang menjalankan "ujian klinikal" dengan kerjasama jabatan lain bagi menangani penyakit COVID-19. Walaupun begitu, Majlis Penyelidikan Perubatan Inida (ICMR) yang pada mulanya memberi sokongan kepada Kementerian AYUSH, telah menolak cadangan kerajaan India untuk menjalakan kajian ke atas manfaat kegunaan air sungai Ganga untuk merawat penyakit COVID-19 melainkan ada data saintifik yang membolehkan kajian dilaksanakan.	kementerian majlis india perubatan ayush kajian kerajaan tradisional covid19 kesihatan berkata klinikal penyelidikan saintifik icmr new delhi 8 mei usaha mencari ubatubatan tempatan dijadikan penawar penyakit menteri dr harsh vardhan sejarah tadisional ayurveda yoga naturopati unani siddha homeopati bertangggungawab ehwal melaksanakan ujian menangani masalah pandemik covid 19 dijalankan kerjasama kesejahteraan keluarga sains teknologi melalui industri csir sokongan dalam perkembangan menolak cadangan menjalankan manfaat air sungai ganga merawat pesakit data diperlukan sebarang mengenainya dilaksanakan
Dataset 3	Sepanjang Perintah Kawalan Pergerakan Bersyarat (PKPB) dilaksanakan, kerajaan negeri Melaka tidak membenarkan sebarang aktiviti sukan dan makan di premis dijalankan malah hanya industri pembuatan sahaja yang dibenarkan beroperasi mengikut panduan prosedur operasi standard (SOP). Ini adalah berikutan teks ucapan perdana menteri Malaysia yang menyatakan sebarang kelonggaran adalah bergantung pada kehendak negeri itu. Keputusan ini dibuat bagi mengelakkan lebih ramai pekerja daripada sektor pembuatan kehilangan pekerjaan mereka. Pengoperasian sektor ekonomi yang lain masih sedang diperhalusi kerana mengambil kira perkembangan semasa kes positif COVID-19 bukan sahaja di Melaka malah di seluruh negara.	melaka dibenarkan negeri makan hari berkata industri pembuatan ekonomi khususnya sektor covid19 kerajaan membenarkan aktiviti perintah kawalan pergerakan pkpb sulaiman beliau beroperasi tempoh operasi sop dilaksanakan jam pkp sosial katanya keputusan dilakukan positif peningkatan sekiranya kedai 4 mei sebarang sukan premis sepanjang bersyarat dilaksana ketua menterinya datuk md sebaliknya pengoperasinya mengikut garis panduan prosedur standard ditetapkan kementerian perdagangan antarabangsa mematuhi dikeluarkan perdana menteri sri muhyiddin yassin pelaksanaan mestilah bersesuaian acuan kehendak kelonggaran justeru menetapkan 24 berbanding 12 manakala buat masa pemberita sidang akhbar bertujuan mengelak ramai pekerja hilang pekerjaan ekoran kerugian ditanggung pihak kilang dalan pengoperasian peruncitan diperhalusi diputuskan jawatankuasa khas menangani diumumkan sehari dua tepat mengambil kira perkembangan semasa negara menunjukkan trend pelanggan berlaku covid 19 pelan

		kontingensi laksanakan melihat aspek penguatkuasaan membendung penularan segi toleransi isu
Dataset 4	Britain telah mengatasi Itali sebagai negara paling teruk terjejas dengan wabak penyakit COVID-19. Setiasusaha Luar Negeri Dominic Raab mengatakan bahawa satu keputusan tepat dan komprehensif mengenai semua punca kematian adalah sukar untuk didapati dengan hanya membuat perbandingan antara negara yang terjejas. Beliau mengatakan bahawa Britain akan mengalami kesusahan untuk menyesuaikan diri dengan kehidupan "normal baharu" selepas ini. Setiasusaha Kesihatan Matt Hancock menyatakan tentang aplikasi mengesan wabak yang masih dalam proses ujian dan pihak kerajaan berharap agar aplikasi tersebut boleh dilancarkan dalam masa terdekat. Perdana Menteri Britain, Boris Johnson berkata pelan komprehensif akan diumumkan mengenai cara memerangi penyakit kerana negara tersebut sudah melepasi waktu puncak wabak itu dan pada masa yang sama merancang untuk membuka semula ekonomi negara.	wabak berkata negara britain itali teruk setiausaha negeri raab selasa menjadikan jumlah 29 data korban berakhir komprehensif aplikasi london 6 mei mengatasi terjejas eropah akibat covid19 lapor agensi berita xinhua luar dominic taklimat media downing street seramai 693 pesakit maut keseluruhan 427 terdahulu dikeluarkan pejabat statistik kebangsaan ons menunjukkan melebihi 32 000 diancam diikuti angka berjumlah 315 keadaan satu tragedi besar sukar membandingkan fikir peroleh keputusan tepat terutamanya antarabangsa punca kematian katanya beliau fasa seterusnya mudah rakyat menyesuaikan kehidupan normal baharu sementara kesihatan matt hancock mengesan menjalani ujian isle of wight kerajaan berharap melancarkan pertengahan bulan pada khamis perdana menteri boris johnson melepasi waktu puncak pelan diumumkan akhir minggu terus memerangi penyakit membuka ekonomi
Dataset 5	Angka korban di Amerika Syarikat telah meningkat berlipat kali ganda mengikut data yang dikeluarkan John Hopkins University. Virus itu telah merebak hampir ke semua pelusuk negara itu dan menjadi negara tertinggi dalam kes.	korban kematian 1 250 jumlah negara china ankara 6 mei covid19 dunia melebihi 000 isnin menurut johns hopkins university berpusat amerika syarikat agensi berita anadolu satu laporan rabu mengatakan 134 dicatatkan pemulihan masingmasing 3 562 919 144 454 as teruk dilanda wabak global juta 68 300 itali kedua tertinggi 29 079 diikuti united kingdom 28 809 sejak minggu dilaporkan kekal seramai 4 637 secara keseluruhan virus merebak 187 muncul disember
Dataset	Ejen rahsia Amerika Syarikat juga telah dijangkiti Covid-19. Pekerja	seramai dijangkiti agensi ejen covid19 jumaat dhs presiden trump pence negara miller

6	yang mempunyai kontak rapat dengan ejen yang dijangkiti juga dikuarantin sepenuhnya. Walaupun begitu, egensi penguatkuasa undang-undang yang dijangkiti sebelum ini sudah sembuh.	washington 9 mei 11 rahsia amerika syarikat memetik yahoo news lapor berita anadolu menurut jabatan keselamatan negeri 23 petugas penguat kuasa undangundang sembuh 60 pekerja dikuarantin dipastikan donald naib mike kontak rapat bertanggungjawab melindungi pemimpin ketua melawat pada berkata jurucakap katie isteri penasihat kanan stephen disahkan
Dataset 7	Di Timur Tengah, Arab Saudi telah mencatatkan lapan kematian akibat COVID-19 manakala di Emiriah Arab Bersatu mencatatkan kes kematian berjumlah 126. Perkara yang sama telah berlaku di Kuwait iaitu seramai 38 orang maut manakala Mesir mencatatkan sebanyak 14 kes kematian. Algeria pula melaporkan sebanyak 4 kematian sementara di Iraq mencatatkan 2 kematian. Bagi Maghribi dan Kuwait, hanya berlaku peningkatan kes wabak COVID-19 sahaja dan tiada kes kematian dilaporkan setakat ini.	jumlah keseluruhan 4 kematian 1 manakala pesakit baharu menjadikan kementerian kesihatan seramai setakat arab covid19 sebanyak dikesan sembuh di maut 14 sementara pulih istanbul mei saudi mencatatkan lapan akibat 184 ahad lapor agensi berita anadolu 552 tempoh 24 jam menyaksikan jangkitan 27 011 137 emiriah bersatu uae 126 bertambah tujuh melonjak 163 564 kuwait 38 lima dicatatkan meningkat 983 mencecah 776 679 qatar 15 551 644 discaj mesir mengatakan meninggal dunia berjumlah 429 negara mengesahkan 272 6 465 algeria 463 korban wabak empat dilaporkan 474 936 iraq melaporkan dua 97 maghribi 903 438
Dataset 8	Kes kematian akibat covid-19 yang kebanyakannya berlaku di Malaysia adalah disebabkan oleh penyakit kronik dan penyakit berjangkit. Kajian di China juga mendapati bahawa separuh daripada kes menderita penyakit kronik.	penyakit peratus covid19 beliau tinggi fizikal kematian satu ncd berkata imun tahun dr lee risiko kronik diabetes dewasa https www who makanan meningkatkan jangkitan sistem berusia intensiti katanya hipertensi malaysia dijangkiti lemah masalah jantung tidur terkena penggunaan diet sihat 9 juta menjaga membantu melawan aktiviti minit aerobik sederhana vitamin badan kajian 65 pernafasan menunjukkan kuala lumpur april 29 70 berjangkit pakar perubatan senaman prof chee pheng imunnya keadaan paruparu tekanan emosi keletihan menjadikan mudah berikutan terdapatnya faktor tembakau ketidakaktifan alkohol berbahaya semasa kanakkanak remaja kaitan signifikan perkembangan masa kolumnya disiarkan bernama com pembunuh utama dunia menurut data pertubuhan kesihatan sedunia intghoncdmortalitymorbidityen 56 global 2016 40 5 71 disebabkan imuniti pemakanan

		memetik garis panduan 2010 intdietphysicalactivityfactsheetadultsen menyatakan 18 64 melakukan sekurangkurangnya 150 seminggu 75 minggu gabungan setara 36 aktif tahap aktifnya kalangan wanita 50 berbanding lelaki 24 7 mengenai pengambilan kaya melalui peningkatan selsel antioksidan berperanan penting mengawal fungsi tubuh terhidrasi cukup membendung namun diingat berfungsi memerangi diketahui dalam catatannya mengetengahkan pusat pengawalan pencegahan china membuat kesimpulan tua cenderung cdc govcoronavirus2019ncovneedextraprecautions peopleathigherrisk html itali perempat 23 rakyat negara ramai berpenyakit analisis amerika syarikat kadar cfr tertinggi berdasarkan laporan misi bersama whochina coronavirus 2019 individu berisiko berumur 60 kardiovaskular sementara pesakit dirawat wuhan 48 menderita akibat jumlah tersebut 30 menghidap diikuti 19 koronari lapan
Dataset 9	Angka pesakit pulih akibat Covid-19 di Malaysia telah mencatat angka tertinggi dan tiada kematian dilaporkan. Daripada jumlah berkenaan kebanyakannya merupakan bukan warganegara	jumlah covid19 dilaporkan dr noor hisham pulih hari berkata pertambahan menjadikan kumulatif peratus keseluruhan putrajaya 333 jangkitan tertinggi dicatatkan perintah kawalan pergerakan pkp dikuatkuasakan 18 mac ketua pengarah kesihatan datuk 307 membabitkan warganegara katanya 7 733 90 9 dalam jam 12 tengah 11 baharu negara 8 505 selain memaklumkan kematian berkaitan jumlahnya kekal 121 1 42
Dataset 10	Perkembangan terbaharu bagi kes Covid-19 ialah kes di Malaysia masih lagi dilaporkan kekal dua digit. Sebuah kluster berkaitan dengan kes kematian di Sabah juga dikenalpasti.	covid19 baharu dilaporkan kematian sidang media kluster dikenalpasti kekal negara warganegara malaysia pulih discaj jumlah kuala lumpur perkembangan terbaharu ketua pengarah kesihatan datuk dr noor hisham abdullah menyaksikan satu hari namun sepanjang tempoh 24 jam dua digit ini lima tumpuan 1 10 melibatkan penularan tiga tujuh 2 import jangkitan luar 3 140 dibenarkan menjadikan kumulatif sepenuhnya wad 7 873 4 akibat virus 121 5 sebuah berkaitan ke120 ke8403 sabah

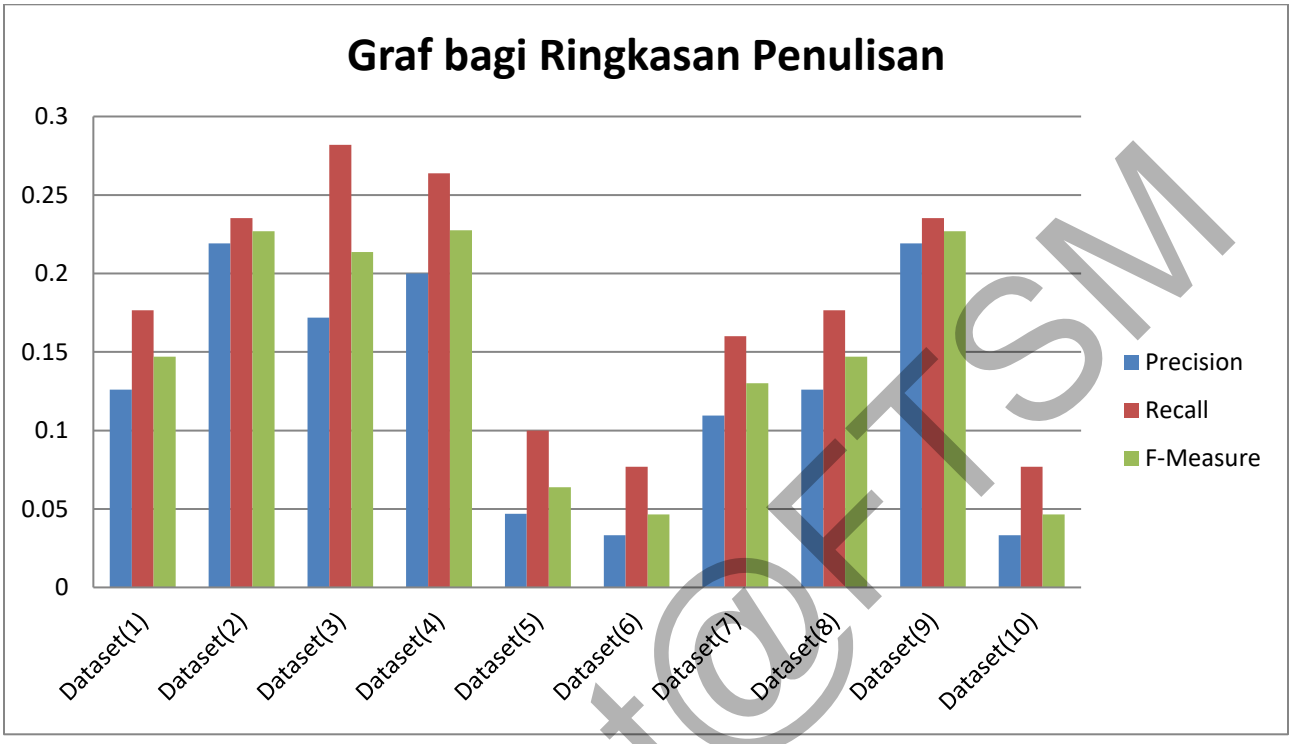
5.1 KEPUTUSAN PERBANDINGAN RINGKASAN

Ringkasan penulisan sistem dan juga ringkasan manusia sangat berkait rapat antara satu sama lain. Oleh itu, precision,recall dan F-measure telah digunakan untuk membandingkan kedua-dua ringkasan ini.

Rajah 4.2 Precision , Recall dan F-Measure berdasarkan Ringkasan

	Precision	Recall	F-Measure
Dataset(1)	0.3424657534246575	0.3048780487804878	0.3225806401781478
Dataset(2)	0.16585365853658537	0.2982456140350877	0.21316613960751182
Dataset(3)	0.20224719101123595	0.4	0.26865671195700613
Dataset(4)	0.2746478873239437	0.3333333333333333	0.3011582962048867
Dataset(5)	0.046875	0.05714285714285714	0.0776698985578285
Dataset(6)	0.1206896551724138	0.22580645161290322	0.15730336624668617
Dataset(7)	0.20430107526881722	0.2714285714285714	0.23312882945538047
Dataset(8)	0.2413793103448276	0.03317535545023697	0.058333331208680644
Dataset(9)	0.10714285714285714	0.2727272727272727	0.1538461497961868
Dataset(10)	0.13043478260869565	0.36	0.19148935779764606
Purata	0.184	0.256	0.120

Purata precision ringkasan ialah 0.184, purata recall ialah 0.256, purata f-measure ialah 0.120. Carta bar bagi precision , recall dan f-measure adalah seperti rajah di bawah



6 PERBINCANGAN DAN KESIMPULAN

Setelah meneliti hasil ringkasan yang dihasilkan oleh algoritma mendapati bahawa hasil ringkasan yang dihasilkan oleh sistem adalah rendah dan tidak mencapai *precision, recall dan f-measure* yang diinginkan. Ini adalah mungkin kerana terdapat banyak kekangan untuk menghasilkan ringkasan ini contohnya library yang wujud bagi Bahasa Melayu adalah sangat terhad. Walau bagaimanapun, library yang wujud digunakan sebaik mungkin untuk meringkaskan teks ini. *Tokenization* dan pemprosesan teks secara manual juga dilakukan disebabkan limitasi library di dalam python.

Tidak dapat dinafikan lagi bahawa hasil ringkasan yang dihasilkan manusia lebih bertepatan dan lebih difahami. Pada masa hadapan diharapkan lebih banyak library bagi Bahasa Melayu

dihasillkan agar ringkasan penulisan ini dapat dijalankan dengan sempurna. Dengan menambahkan lebih banyak dataset atau menggunakan teknik *word embedding* yang lain seperti *Glove* mungkin akan dapat meningkatkan ketepatan yang diinginkan.

7 RUJUKAN

Ji Eun Lee, H. S. (2013). Learning to Predict the Need of Summarization on News Article. *Science Direct* , 274-279.

Mr.S.A.Babar. (2013). Text Summarization : An Overview. *Research Gate* .

Sarkar, D. (2016). *Text Analytics with Python.* Apress.

Hutchins, J. (1987). Summarization : Some Problems and Methods. *University of East Aglia* , 151-173.

Ismail, N. H. (2017). Ringkasan Penulisan Bahasa Melayu.

Palomar, E. L. (2009). Challenging Issues of Automatic Summarization: Relevance Detection and Quality-based Evaluation. *Department of Software and Computing Systems , University of Alicante , Spain* , 29-35.

A.Khan, S. a. (2005). MRST: A new technique for information summarization . Dragonmir R Radev, E. H. (2002). Introduction to the special issue on summarization.

Computational linguistics 28 , 399-408.

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM (JACM)* , 264-289.

Elhadad, R. B. (1997). Using lexical chains in summarization . *Proceedings of the ACL/EACL '97 Workshop on Intelligent Scalable Text Summarization , Madrid, Spain* , 10-17.

H.Chen, J. K. (2008). Multidocument summary generation: using informative and event words.

ACM Trans Asian Lang Inf Process (TALIP) , 37-41.

Julien Kupiec, J. P. (1995). A trainable document summarizer . *In Proceeding of the 18th annual international ACM SIGIR conference on Research and Development in information retrieval* , 68-73.

K. Farshad, K. H. (2008). Optimizing text sumarization based on fuzzy logics. *Proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science, IEE, University of Shahid Bahanor , UK* .

L. Yulia, H. R. (2011). EM clustering algorithm for automatic text summarization . *Proceedings of the 10th Mexican International Conference on Advances in Artificial Intelligence* , 305-311.

L.Chun-He, P.-Y. Z. (2009). Automatic text summarization based on sentences clustering and extraction. *Computer Science and Information Technology* .

Lehal, V. G. (2010). A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web* , 258-268.

Liu, Y. a. (2001). Generic text summarization using relevance measure and latent semantic analysis. *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* , 19-25.

Luhn, H. P. (1958). Text automatic creation literature abstract. *IBM Journal of research and development* , 159-165.

Mohammad, M. a. (2012). Automated text summarization base on lexicales chain and graph using of word net and wikipedia knowledge base. *IJCSI International Journal of Computer Science Issues* .

S.T. Khushboo, R. D. (2010). Graph-based algorithms for text summarization. *Third International Conference on Emerging Trends in Engineering and Technology.*

Z.Teng, Y. a. (2008). Single document summarization based on local topic identification and word frequency. *Proceedings of the 2008 Seventh Mexican International Conference on Artificial Intelligence* , 37-41.