PENGECAMAN ENTITI NAMA BAGI ARTIKEL KESIHATAN BAHASA MELAYU

Adibah Syahzani Binti Safferi

Md. Jan Nordin

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Pengecaman Entiti Nama (NER) adalah salah satu pemprosesan bahasa tabii (NLP) untuk mengekstrak konsep dan maklumat penting daripada sesuatu dokumen bertulis. Pelbagai pendekatan boleh dilakukan untuk proses pengekstrakan maklumat, dan kajian ini lebih tertumpu terhadap pendekatan berasaskan peraturan. Pengecaman entiti nama untuk dokumen bahasa asing seperti Bahasa Melayu merupakan kajian yang kiat dijalankan kini dalam pemprosesan bahasa tabii (NLP), termasuklah kajian mengikut domain-domain yang lebih spesisfik seperti domain kesihatan. Kajian ini dilakukan untuk mengenalpasti entiti-entiti dan peraturan yang boleh ditambah untuk pembangunan model pengecam entiti nama khususnya bagi domain kesihatan.

1 PENGENALAN

Pemprosesan Bahasa Tabii atau dikenali sebagai Natural Language Processing (NLP) adalah subbidang penyelidikan kecerdasan buatan yang berkaitan dengan masalah membentuk perisian komputer yang dapat memproses bahasa tabii manusia seperti Bahasa Melayu, Bahasa Inggeris dan bahasa lain-lain (Zakree, 2015). Bahasa tabii, atau bahasa yang digunakan dalam kehidupan seharian manusia, mempunyai pembendaharaan kata yang meluas dan pelbagai. Selain itu, mesin dan komputer lebih cenderung untuk menghadapi beberapa masalah dalam memahami bahasa tabii berpunca kepada ralat-ralat yang tidak mampu dielakkan contohnya terdapat sesetengah perkataan yang sama tetapi mempunyai makna yang berbeza. Oleh itu, bidang ini bertujuan untuk mengurangkan jurang berkenaan kefahaman mesin dalam memahami bahasa tabii yang digunakan manusia.

Pada masa kini, terdapat banyak dokumen yang tidak berstruktur seperti teks berita, artikel blog, forum, tweet serta mikro blog dari rangkaian sosial. Dokumen-dokumen ini amat sukar untuk difahami oleh komputer. Bagi memperoleh maklumat daripada dokumen ini, capaian secara manual perlu dilakukan dan ini mengambil masa yang lama serta tidak praktikal. Oleh itu, kajian berkaitan pengekstrakan maklumat menjadi sangat penting bagi mengatasi masalah ini. Salah satu teknik yang semakin mendapat perhatian dalam bidang penyelidikan ialah Pengecaman Entiti Nama iaitu Named-Entity Recognition (NER) (Saad & Mansor, 2018).

Pengecaman Entiti Nama (NER) merupakan salah satu dari cabang penyelidikan Pengekstrakan Maklumat (IE) yang amat penting. Pengekstrakan maklumat ini bertujuan untuk mendapatkan senarai kata kunci yang relevan bagi sesuatu dokumen. Melalui NER, pengekstrakan maklumat dapat dilakukan dengan mengenalpasti kata nama serta mengklasifikasikan mengikut kategori individu, organisasi, lokasi, nilai kewangan, nilai peratusan dan tarikh atau masa (Saad & Mansor, 2018).

2 PENYATAAN MASALAH

Kajian berkaitan dengan teknik pengecaman entiti nama bagi dokumen bahasa Melayu masih baru jika dibandingkan dengan bahasa lain. Pada masa ini, aplikasi komersil bagi pengecaman entiti nama hanya terdapat dalam bahasa Inggeris. Oleh yang demikian, adalah menjadi keperluan penyelidik mengkaji dan menghasilkan model yang sesuai dengan bahasa Melayu, kerana pencarian maklumat dari sumber yang pelbagai akan memakan masa yang lama. Melalui penggunaan pengecaman entiti nama ini, entiti-entiti tertentu dapat diekstrak terlebih dahulu dan disimpan ke dalam pangkalan data bagi tujuan carian pada masa akan datang (Saad & Mansor, 2018).

Walaupun kajian NER Bahasa Melayu sudah dijalankan, tetapi terdapat beberapa peraturan yang dihasilkan masih tidak mencukupi dan tidak menyeluruh dalam mengecam entiti nama, terutamanya mengikut domain-domain yang tertentu (Nadia & Omar, 2019). Kajian NER bagi

domain kesihatan di dalam bahasa Melayu masih kurang diberi perhatian. Selain itu, terdapat beberapa kekurangan di dalam korpus Bahasa Melayu yang dibina terutamanya dalam bidang berkaitan dengan kesihatan. Ini salah satu faktor untuk membangunkan model prototaip sistem NER bagi dokumen kesihatan bahasa Melayu (Saad & Mansor, 2018).

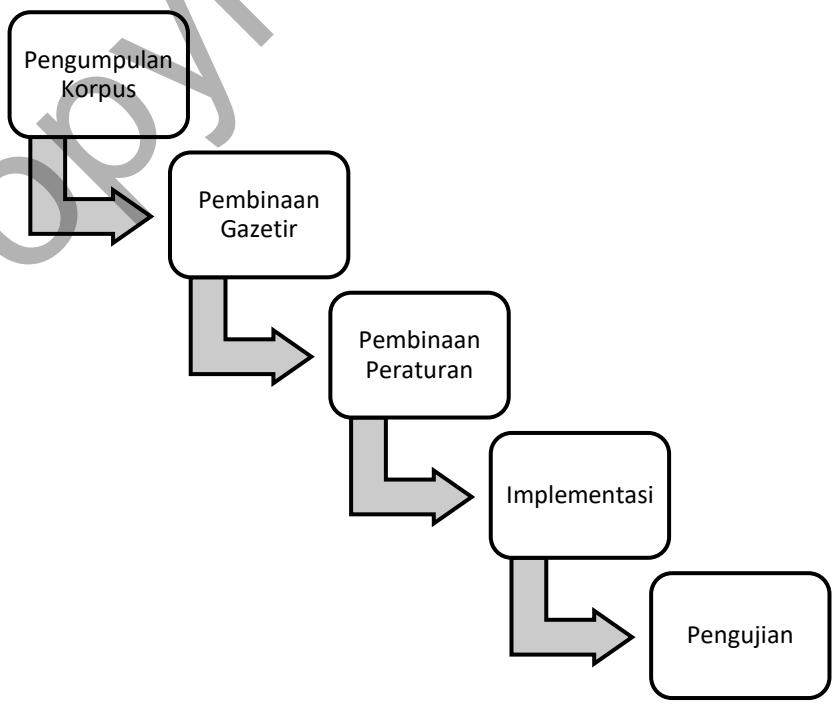
3 OBJEKTIF KAJIAN

Matlamat utama projek ini adalah:

- Mengkaji model yang mampu mengecam entiti nama kesihatan khususnya dalam Bahasa Melayu
- Mempelajari cara membangunkan model NER menggunakan pencarian leksikal dan pendekatan berasaskan peraturan

4 METOD KAJIAN

Metodologi yang digunakan bagi pembangunann model prototaip ini melibatkan lima fasa iaitu fasa pengumpulan korpus, fasa pembinaan gazetir, fasa pembinaan peraturan, fasa implementasi dan fasa pengujian.



RAJAH 1 Fasa metodologi untuk kajian ini

4.1 Fasa Pengumpulan Korpus

Pada fasa ini, data-data yang berkenaan dengan domain kesihatan dikumpulkan untuk menjana korpus bagi kajian ini. Artikel-artikel yang mempunyai segala maklumat relevan berkaitan dengan kesihatan dikumpulkan daripada laman web Hello Doktor, Majalah Sains dan laman sesawang rasmi Kementerian Kesihatan Malaysia (MyHEALTH).

4.2 Fasa Pembinaan Gazetir

Pada fasa ini, senarai kata kunci yang berkenaan dengan kesihatan dikumpulkan untuk dijadikan gazetir. Senarai kata kunci seperti penyakit, organ, indikator dan rawatan diperoleh daripada laman web rasmi Kementerian Kesihatan Malaysia (MyHEALTH).

diabetes, tuberculosis, selesema, denggi, loya, demam, leptospirosis, jaundis, ulser, ruam, radang, kudis

RAJAH 2.1 Senarai kata kunci penyakit

jantung, hidung, telinga, tekak, otak, kulit, tangan, dada, pinggang, hati, gigi, perut, kepala, bibir, lidah, gusi

RAJAH 2.2 Senarai kata kunci organ

masalah, penghidap, serangan, komplikasi, sakit, kesakitan, kesukaran, penyakit, kerosakan

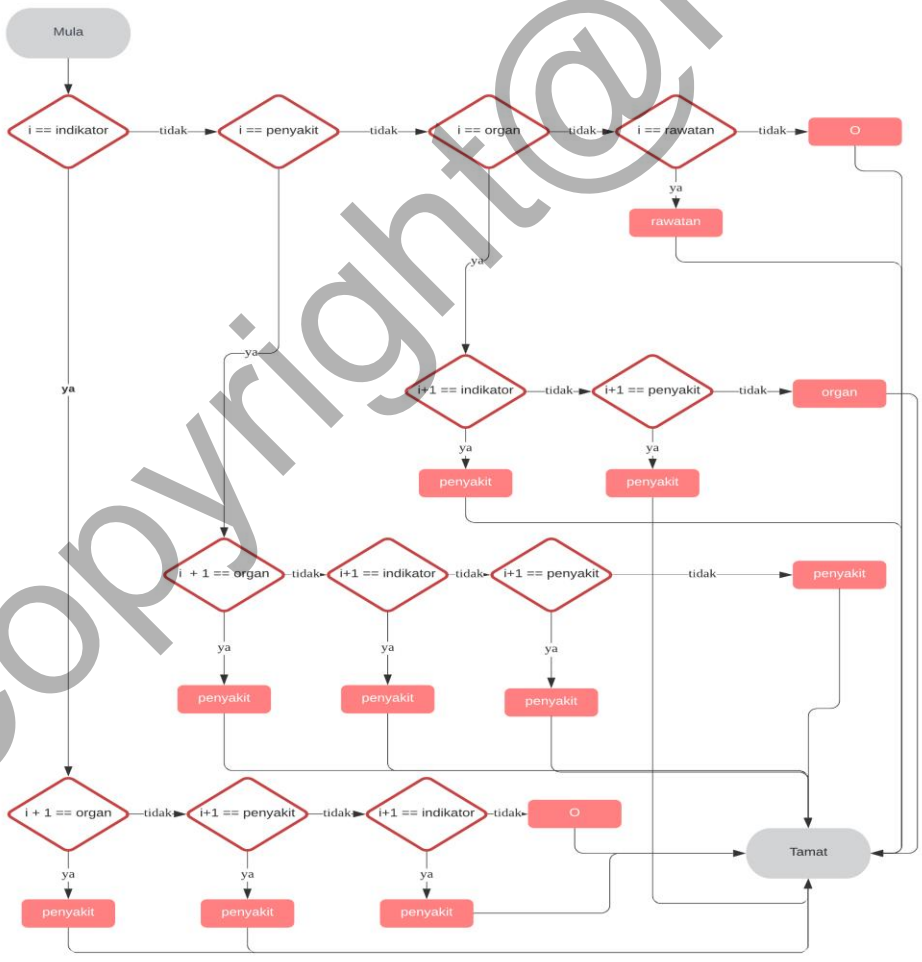
RAJAH 2.3 Senarai kata kunci indikator

pembedahan, laser, lasik, imunoterapi, vascepa, placebo, entresto, valsartan, avocado, susu

RAJAH 2.4 Senarai kata kunci rawatan (ubat/makanan/terapi)

4.3 Fasa Pembinaan Peraturan

Pada fasa ini, peraturan dibina untuk menentukan entiti nama bagi penyakit, organ, dan rawatan menggunakan gazetir yang telah dibangunkan. Cadangan peraturan yang dibangunkan untuk menentukan entiti penyakit, organ, dan rawatan adalah seperti rajah di bawah.



RAJAH 3 Carta alir peraturan entiti nama bagi penyakit, organ, dan rawatan

```
if token is in disease then
    entity = disease
    next_token = token + 1
    if next_token is in disease then
        new_sentence = token + next_token
        entity = disease
    if next_token is in indicator then
        new_sentence = token + next_token
        entity = disease
    if next_token is in organ then
        new_sentence = token + next_token
        entity = disease
    else
        new_sentence = token
        entity = entity
return new sentence, entity
```

RAJAH 4.1 Pseudokod peraturan entiti penyakit berdasarkan gazetir penyakit

```
if token is in indicator then
    entity = indicator
    next_token = token + 1
    if next_token is in organ then
        new_sentence = token + next_token
        entity = disease
    if next_token is in disease then
        new_sentence = token + next_token
        entity = disease
    if next_token is in indicator then
        new_sentence = token + next_token
        entity = disease
    else
        new_sentence = token
        entity = other
return new sentence, entity
```

RAJAH 4.2 Pseudokod peraturan entiti penyakit berdasarkan gazetir indicator

```
if token is in organ then
    entity = organ
    next_token = token + 1
    if next_token is in indicator then
        new_sentence = token + next_token
        entity = disease
    if next_token is in disease then
        new_sentence = token + next_token
        entity = disease
    else
        new_sentence = token
        entity = organ
return new sentence, entity
```

RAJAH 4.3 Pseudokod peraturan entiti penyakit dan organ berdasarkan gazetir organ

```
if token is in treatment then
    entity = treatment
    new_sentence = token
return new sentence, entity
```

RAJAH 4.4 Pseudokod peraturan entiti rawatan

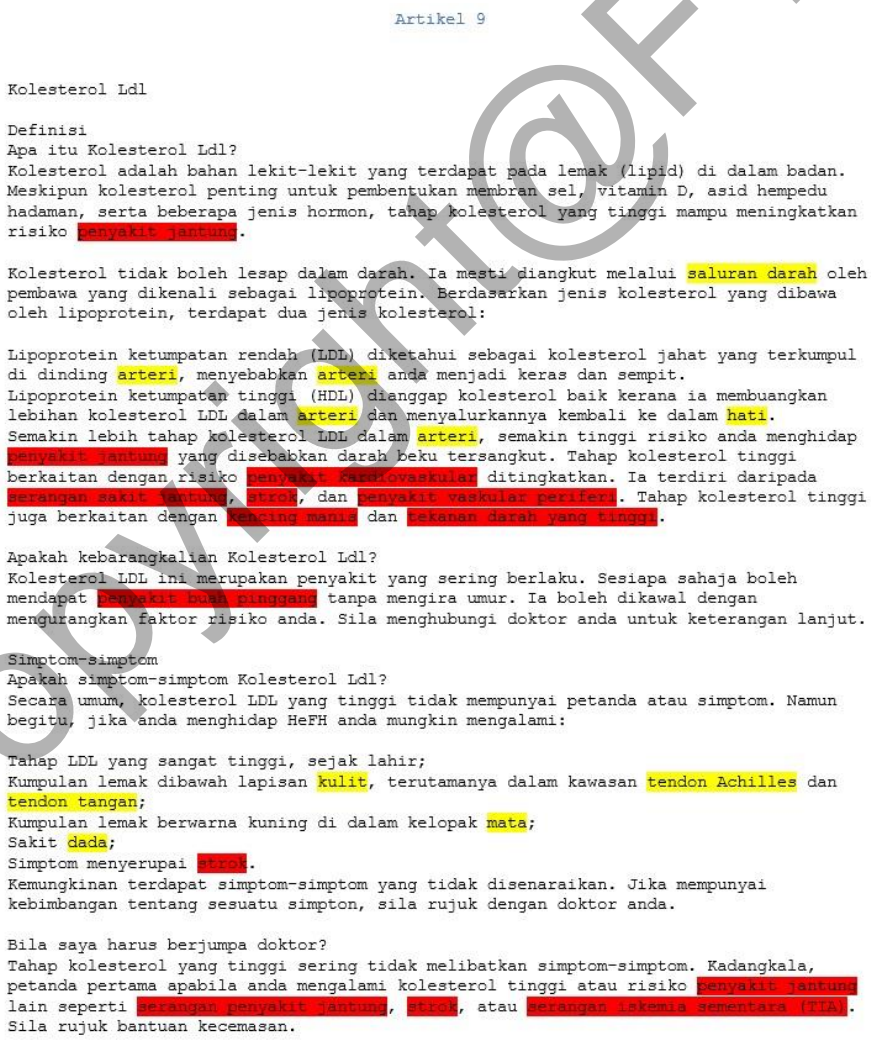
4.4 Fasa Implementasi

Pada fasa ini, model diwujudkan dengan mengimplementasikan pencarian leksikal dan pendekatan berasakan peraturan. Bahasa pengaturcaraan yang digunakan dalam kajian ini adalah Python, melalui aplikasi Jupyter Notebook.

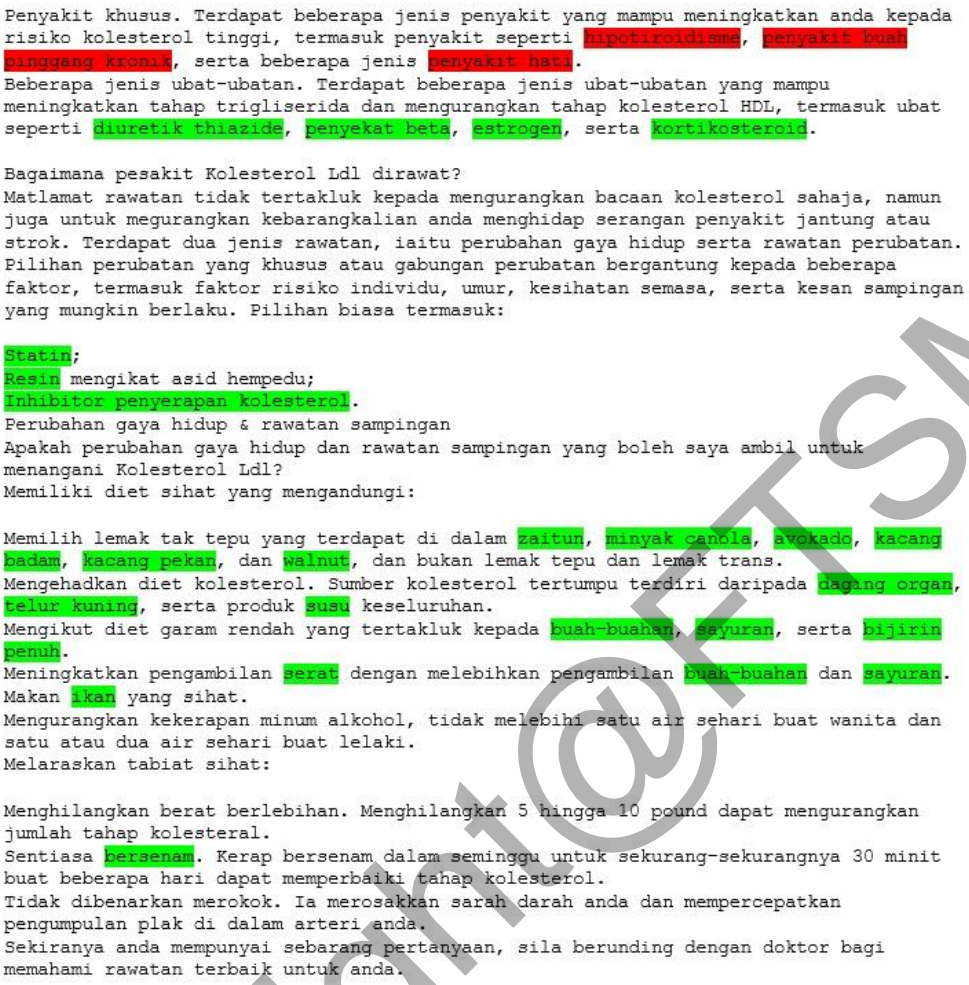
4.5 Fasa Pengujian

Pada fasa ini, model diuji dengan menggunakan artikel-artikel yang dikumpulkan sebagai input. Pengiraan terhadap jadual pengujian dikira menggunakan penilaian dapatan (recall), kejituan (precision) dan F-measure (Saad & Mansor, 2018).

Sebelum korpus artikel diuji, artikel-artikel ini perlu ditanda entitinya terlebih dahulu menggunakan kepakaran manusia. Contoh penandaan entiti menggunakan kepakaran manusia boleh dilihat berdasarkan rajah di bawah.



RAJAH 4.5 Entiti penyakit dan organ yang ditanda dengan kepakaran manusia



Penyakit khusus. Terdapat beberapa jenis penyakit yang mampu meningkatkan anda kepada risiko kolesterol tinggi, termasuk penyakit seperti hipotiroidisme, penyakit buah pinggang kronik, serta beberapa jenis penyakit hati.
Beberapa jenis ubat-ubatan. Terdapat beberapa jenis ubat-ubatan yang mampu meningkatkan tahap trigliserida dan mengurangkan tahap kolesterol HDL, termasuk ubat seperti diuretik thiazide, penyekat beta, estrogen, serta kortikosteroid.

Bagaimana pesakit Kolesterol Ldl dirawat?
Matlamat rawatan tidak tertakluk kepada mengurangkan bacaan kolesterol sahaja, namun juga untuk megurangkan kebarangkalian anda menghidap serangan penyakit jantung atau strok. Terdapat dua jenis rawatan, iaitu perubahan gaya hidup serta rawatan perubatan. Pilihan perubatan yang khusus atau gabungan perubatan bergantung kepada beberapa faktor, termasuk faktor risiko individu, umur, kesihatan semasa, serta kesan sampingan yang mungkin berlaku. Pilihan biasa termasuk:

Statin;
Resin mengikat asid hempedu;
Inhibitor penyerapan kolesterol.
Perubahan gaya hidup & rawatan sampingan
Apakah perubahan gaya hidup dan rawatan sampingan yang boleh saya ambil untuk menangani Kolesterol Ldl?
Memiliki diet sihat yang mengandungi:

Memilih lemak tak tepu yang terdapat di dalam zaitun, minyak canola, avokado, kacang badam, kacang pekan, dan walnut, dan bukan lemak tepu dan lemak trans.
Mengehadkan diet kolesterol. Sumber kolesterol tertumpu terdiri daripada dagang organ, telur kuning, serta produk susu keseluruhan.
Mengikut diet garam rendah yang tertakluk kepada buah-buahan, sayuran, serta bijirin penuh.
Meningkatkan pengambilan serat dengan melebihkan pengambilan buah-buahan dan sayuran.
Makan ikan yang sihat.
Mengurangkan kekerapan minum alkohol, tidak melebihi satu air sehari buat wanita dan satu atau dua air sehari buat lelaki.
Melaraskan tabiat sihat:

Menghilangkan berat berlebihan. Menghilangkan 5 hingga 10 pound dapat mengurangkan jumlah tahap kolesteral.
Sentiasa bersenam. Kerap bersenam dalam seminggu untuk sekurang-sekurangnya 30 minit buat beberapa hari dapat memperbaiki tahap kolesterol.
Tidak dibenarkan merokok. Ia merosakkan sarah darah anda dan mempercepatkan pengumpulan plak di dalam arteri anda.
Sekiranya anda mempunyai sebarang pertanyaan, sila berunding dengan doktor bagi memahami rawatan terbaik untuk anda.

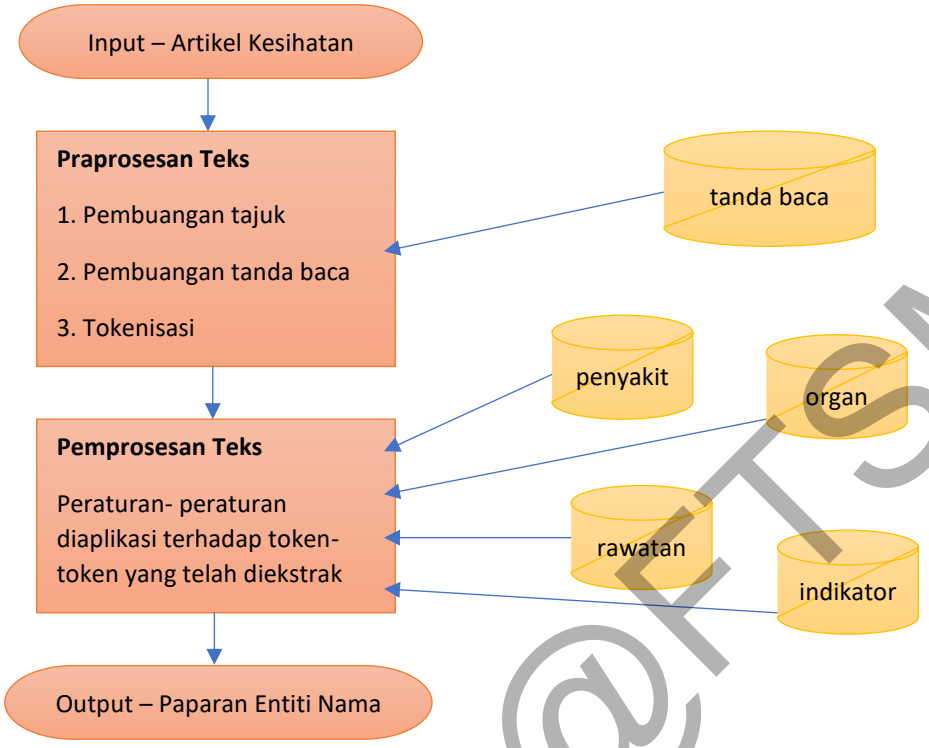
RAJAH 4.6 Entiti penyakit dan rawatan(ubat/makanan/terapi) dengan kepakaran manusia

Rajah 4.5 dan 4.6 merupakan sumber dari artikel yang sama. Setelah penandaan dilakukan, entiti-entiti tersebut disenaraikan di dalam jadual pengujian.

Jadual Pengujian

Artikel009.txt	Entiti Nama (Kepakaran Manusia)	Entiti Nama (Model)	T	ST	TT
Penyakit/ Kondisi	penyakit jantung penyakit kardiovaskular serangan sakit jantung strok penyakit vaskular periferi kencing manis tekanan darah yang tinggi penyakit buah pinggang sakit dada serangan iskemia sementara (TIA) hipotiroidisme penyakit hati				
Organ	jantung saluran darah arteri hati kulit buah pinggang dada tendon Achilles tendon tangan mata hati				
Rawatan (terapi/ makanan/ ubat)	diuretik thiazide penyekat beta estrogen kortikosteroid statin resin inhibitor penyerapan kolesterol zaitun minyak canola avocado kacang badam kacang pekan walnut daging organ telur kuning susu buah-buahan sayuran bijirin penuh serat ikan bersenam				

RAJAH 4.7 Entiti nama yang telah disenaraikan ke dalam jadual



RAJAH 4.8 Carta alir model pengecam entiti nama di fasa pengujian

5 HASIL KAJIAN

Pada fasa pengujian, entiti nama bagi penyakit, organ dan rawatan telahpun ditanda menggunakan kepakaran manusia terlebih dahulu sebelum teks artikel dibaca oleh model. Kemudian, entiti nama tersebut direkod dalam bentuk jadual untuk memudahkan perbandingan dilakukan. Hasil output model dibandingkan bersama dengan entiti nama kepakaran manusia seperti di rajah di muka sebelah.

Jadual Pengujian

Artikel009.txt	Entiti Nama (Kepakaran Manusia)	Entiti Nama (Model)	T	ST	TT
Penyakit/ Kondisi	penyakit jantung penyakit kardiovaskular serangan sakit jantung strok penyakit vaskular periferi kencing manis tekanan darah yang tinggi penyakit buah pinggang sakit dada serangan iskemia sementara (TIA) hipotiroidisme penyakit hati	penyakit jantung darah serangan sakit sakit jantung strok kencing manis tekanan darah kebimbangan kronik hipotiroidisme penyakit	3	4	3
Organ	jantung saluran darah arteri hati kulit buah pinggang dada tendon Achilles tendon tangan mata	jantung arteri hati pinggang kulit tangan mata dada	6	2	0
Rawatan (terapi/ makanan/ ubat)	diuretik thiazide penyekat beta estrogen kortikosteroid statin resin inhibitor penyerapan kolesterol zaitun minyak canola avocado kacang badam kacang pekan walnut daging organ telur kuning susu buah-buahan sayuran bijirin penuh serat ikan bersenam	diuretik thiazide kortikosteroid statin zaitun minyak canola avokado kacang badam pekan walnut daging telur susu buah sayuran bijirin penuh serat ikan bersenam	10	12	0

RAJAH 5 Hasil entiti direkod dalam jadual pengujian

Berdasarkan Rajah 5, jumlah ketepatan entiti nama model direkodkan dalam tiga ruangan di sebelah kanan, iaitu T, ST dan TT yang mewakili tepat, separa tepat dan tidak tepat. Hasil yang tepat mempunyai entiti nama yang sama seperti entiti oleh kepakaran manusia, manakala hasil yang separa tepat mempunyai sebahagian daripada padanan entiti oleh kepakaran manusia, dan hasil yang tidak tepat tidak mempunyai entiti seperti yang disenaraikan oleh kepakaran manusia.

Penilaian dilakukan terhadap korpus ujian dengan beberapa pengiraan. Pertama sekali, pengiraan dilakukan dengan menjumlahkan entiti penyakit mengikut ketepatan (tepat, separa tepat, dan tidak tepat), dan perkara yang sama dilakukan terhadap entiti organ dan rawatan. Kemudian, keseluruhan entiti model dijumlahkan untuk mendapatkan jumlah entiti yang berjaya diekstrak oleh model supaya boleh dibandingkan dengan jumlah entiti yang menggunakan kepakaran manusia.

Hasil turut dinilai menggunakan penilaian dapatan, kejituan, dan ukuran-f. Definisi bagi penilaian tersebut berdasarkan kajian lepas (Saad & Mansor, 2018) ialah:

a) Dapatan(*recall*): bilangan pengecaman entiti nama yang tepat oleh sistem.

b) Separa tepat(*partial*): bilangan pengecaman entiti nama separa tepat oleh sistem.

c) PEN secara manual(*possible*): bilangan pengecaman entiti nama yang dilakukan secara manual oleh pakar bahasa.

d) Pen oleh sistem(*actual*): bilangan keseluruhan pengecaman entiti nama yang dilakukan oleh sistem termasuk pengecaman yang tepat, separa tepat dan tidak tepat.

Berdasarkan kajian lepas (Nadia & Omar, 2019), formula berikut digunakan untuk penilaian:

$$Dapatan = \frac{tepat + (0.5 * separa\ tepat)}{jumlah\ keseluruhan\ PEN\ secara\ manual}$$

$$Kejituan = \frac{tepat + (0.5 * separa\ tepat)}{jumlah\ keseluruhan\ PEN\ oleh\ sistem}$$

$$Ukuran - F = \frac{kejituan * dapatam}{0.5 * (kejituan + dapatan)}$$

Berdasarkan pengujian yang telah dilakukan terhadap korpus ujian, jumlah dapatan, kejituan, dan ukuran-f diringkaskan berdasarkan jadual di bawah.

Entiti	**Tepat(T)**	**Separa Tepat (ST)**	**Tidak Tepat (TT)**	**Jumlah Kepakaran Manusia**	**Dapatan**	**Kejituan**	**Ukuran-F**
Penyakit	97	68	83	236	55.50%	52.82%	54.13%
Organ	70	12	14	83	91.57%	79.17%	84.92%
Rawatan	57	36	25	96	78.12%	63.56%	70.09%
Jumlah	224	116	122	415	67.95%	61.04%	64.31%

JADUAL 5 Jadual keputusan penilaian model kajian ini

Berdasarkan penilaian yang telah dilakukan, hasil mendapati nilai keseluruhan kajian bagi dapatan ialah 67.95%, manakala kejituan ialah 61.04%, dan ukuran-f ialah 64.31%. Bagi entiti penyakit, nilai dapatan, kejituan, dan ukuran-f ialah di tahap sederhana iaitu 55.50%, 52.82% dan 54.13%. Hal ini kerana model mampu mengecam hanya dua token yang mampu menghasilkan entiti penyakit, tetapi tidak mampu mengesan tiga token yang sebenarnya merupakan sebuah entiti penyakit. Seperti contoh, model hanya mampu mengeluarkan “tekanan darah” dan “darah tinggi”, namun tidak mampu menghasilkan “tekanan darah tinggi”. Selain itu, model turut mempunyai kecenderungan untuk menggabungkan dua jenis penyakit, sebagai contoh, “strok” dan “serangan jantung” adalah entiti yang sama namun berbeza jenis, dan penggabungan terjadi apabila model mengeluarkan kedua-dua token iaitu “strok serangan” dan “serangan jantung”. Hal ini berkemungkinan turut mempengaruhi kejituan model, selain terdapat beberapa entiti nama yang tiada di dalam gazetir. Selain itu, model kajian mengeluarkan entiti penyakit seperti biasa.

Manakala, bagi entiti organ, nilai dapatan, kejituan, dan ukuran-f ialah 91.57%, 79.17% dan 84.92%. Kesilapan berlaku apabila organ yang ditanda menggunakan kemahiran manusia adalah lebih spesifik, contohnya “urat femoral”. Model hanya mampu mengeluarkan “urat” sebagai entiti

organ kerana ia hanya mengecam organ yang umum namun tidak yang spesifik. Selain itu, model turut mengeluarkan entiti yang tidak menepati konteks. Misalnya, artikel yang menerangkan mengenai cara memberus gigi secara terperinci mempunyai ayat seperti “kepala berus gigi”, di mana model mengeluarkan “kepala” daripada ayat tersebut di mana ianya tidak relevan dengan entiti organ jika dibaca mengikut konteks. Selain daripada itu, model kajian mengeluarkan entiti organ seperti biasa.

Bagi entiti rawatan pula, nilai dapatan, kejituan, dan ukuran-f ialah 78.12%, 63.56% dan 70.09%. Terdapat sedikit kesilapan di mana entiti rawatan yang dikeluarkan tidak menepati konteks. Contohnya, “buah pinggang” merupakan entiti organ, namun model mengecam “buah” di dalam ayat tersebut sebagai entiti rawatan iaitu makanan. Di samping itu, terdapat beberapa entiti yang tiada di dalam gazetir yang berkemungkinan menyebabkan ianya mempengaruhi nilai dapatan, kejituan dan ukuran-f bagi entiti rawatan ini.

6 KESIMPULAN

Terdapat beberapa kekangan yang dapat dikenalpasti setelah mengkaji sistem ini, seperti ketiadaan set data berlabel mahupun gazetir yang sedia ada khususnya bagi entiti yang berkaitan dengan artikel kesihatan untuk melaksanakan projek ini. Set data mahupun gazetir perlu dibangunkan sendiri dan hal ini memakan masa yang agak lama dan menyebabkan beberapa fasa perlu ditunda dan dilakukan dalam masa yang terhad seperti fasa pembangunan peraturan kerana pembangunan peraturan yang efektif turut memakan masa. Selain itu, kekangan yang lain melibatkan model kajian ini adalah beberapa kesilapan terhadap hasil entiti nama seperti ketidakmampuan mengekstrak entiti penyakit melebihi dua token dan pengecaman terhadap satu entiti yang mempunyai perkataan entiti yang lain. Hal ini boleh mengakibatkan sistem meletakkan sesuatu entiti pada kategori atau kelas yang kurang tepat.

Penambahbaikan yang boleh dilakukan ialah mewujudkan set data mahupun gazetir bagi entiti yang berkaitan dengan kesihatan, mahupun menambahkan lagi data atau token ke dalam set data atau gazetir yang sedia ada supaya model mampu mengecam lebih banyak entiti. Selain itu, penambahbaikan seperti menambahkan lagi kategori atau domain yang spesifik untuk memperluaskan lagi pengekstrakan maklumat dengan mempunyai kategori yang lebih meluas juga boleh dilakukan. Misalnya, dalam projek ini, kategori entiti yang dibangunkan ialah penyakit, organ, dan rawatan, dan di dalam kategori rawatan masih boleh dibahagikan kepada tiga lagi entiti iaitu ubat, makanan dan terapi. Entiti lain seperti simptom juga turut boleh ditambah ke dalam sistem buat masa akan datang.

Kesimpulannya, sistem ini bertujuan untuk membangunkan pengecam entiti nama Bahasa Melayu dalam domain kesihatan. Pembangunan sistem ini memerlukan perancangan rapi dengan mengaplikasi metodologi yang telah digariskan sesuai dengan pembangunan projek seperti yang telah dinyatakan di dalam bab-bab yang sebelumnya. Akhir sekali, sistem ini juga mempunyai banyak ruang untuk diperbaiki dan diperkembangkan lagi untuk masa akan datang.

7 RUJUKAN

Alfred, R., Leong, L. C., & On, C. K. (2014). Malay Named Entity Recognition Based on Rule-Based Approach. *International Journal of Machine Learning and Computing, 4*(3), 300-306. doi:10.7763/IJMLC.2014.V4.428

Darwich, M., & Mohd, M. (2015). Probabilistic Reference to Suspect or Victim in Nationality Extraction from Unstructured Crime News Documents. *Information and Knowledge Management, 5*(9), 64-75.

Lesser, U., & Hakenberg, J. (December, 2005). What makes a gene name? Named entity recognition in the biomedical literature. *Briefings in Informatics, 6*(4), 357-369.

Nadia, U., & Omar, N. (2019). Pengecaman Entiti Nama Bahasa Melayu Menggunakan Pendekatan Berasaskan Peraturan. *Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik*, *8*(1), 37-47.

Saad, S., & Mansor, M. K. (2018). Pendekatan Teknik Pengecaman Entiti Nama Bagi Capaian Berita. *GEMA Online: Journal of Language Studies*, 216-235.

Soomro, P. D., Kumar, S., Banbhrani, Shaikh, A. A., & Raj, H. (2017). Bio-NER: Biomedical Named Entity Recognition using Rule-Based and Statistical Learners. *(IJACSA) International Journal of Advanced Computer Science and Applications, 8*(12), 163-170. Retrieved 4 December, 2019, from www.ijacsa.thesai.org [4 Disember 2019]

Tome, E., KorousÏ, S. B., & Peter, K. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS ONE, 12*(6). Retrieved from https://doi.org/10.1371/journal.pone.0179488 [4 Disember 2019]

Zakree, M. (2015). *Pemprosesan Bahasa Tabii Dengan Python*. Retrieved from http://www.ukm.my/zakree/pemprosesan-bahasa-tabii-dengan-python/ [19 Oktober 2019]