PENGEKSTRAKAN ENTITI NAMA DATA KLINIKAL BAGI GRAF PENGETAHUAN DENGAN MENGGUNAKAN PELARASAN HALUS MODEL BIODISCHARGESUMMARYBERT

Lim Teck Huat, Lailatul Qadri Binti Zakaria

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor Malaysia.

p107443@siswa.ukm.edu.my, lailatul.qadri@ukm.edu.my

ABSTRAK

Graf pengetahuan perubatan sumber terbuka menggunakan perisian bukan proprietari tidak terhad kepada SQL atau skema perubatan. Graf pengetahuan boleh diskalakan dan mudah difahami oleh pengguna awam serta penyedia jagaan kesihatan untuk membolehkan mereka menambah pengetahuan perubatan. Pangkalan pengetahuan sedia ada bukan sumber terbuka wujud, namun begitu pangkalan pengetahuan sumber terbuka lazimnya kompleks seperti Bioportal. Selain daripada itu, terdapat juga pangkalan pengetahuan yang khusus kepada penyakit sahaja seperti Jensenlab atau menggunakan perisian proprietari seperti Neo4j. Kajian ini bertujuan untuk membangunkan graf pengetahuan perubatan untuk data klinikal berasaskan kepada entiti nama. Perisian Protege digunakan untuk visualisasikan data dan menguji capaian maklumat. Dataset yang digunakan dalam kajian ini ialah n2c2 2010 dan MedlinePlus. Entiti nama data klinikal diekstrak menggunakan pelarasan halus BioDischargeSummaryBERT. Entiti nama tersebut digunakan untuk menjana pangkalan data pengetahuan. Kajian ini dibahagikan kepada dua peringkat utama. Peringkat pertama bermula dengan peringkat pelarasan halus menggunakan set data n2c2 2010. Tujuan analisis di peringkat pertama adalah untuk mendapatkan nilai pelarasan halus hiperparameter yang paling optimum bagi mengekstrak entiti nama dengan menggunakan BioDischargeSummaryBERT. Peringkat kedua merupakan peringkat aplikasi yang menggunakan set topik data kesihatan MedlinePlus. Tujuan analisis di peringkat kedua adalah menguji keberkesanan model beserta nilai hiperparameter yang telah ditentukan di peringkat pertama. Keseluruhannya, terdapat enam fasa dalam kajian ini iaitu pengumpulan data, pengecaman entiti nama, penilaian, pasca-pemprosesan, penjanaan format RDF/RDFS dan penjanaan graf pengetahuan. Kajian semasa menunjukkan model pelarasan halus tanpa Medan Rawak Bersyarat (MRB) mempunyai skor F1 sebanyak 0.87 mengatasi skor F1 model BiLSTM dengan MRB sebanyak 0.84. Model BioDischargeSummaryBERT dan MRB mempunyai skor F1 bernilai 0.87 dan hampir setanding dengan model klinikal Stanza. Untuk pengujian data graf pengetahuan berasaskan entiti nama yang diekstrak, model kajian semasa mendapat skor median F1 0.91 untuk 246 topik kesihatan. Kajian ini membantu penyelidik untuk mengenal pasti maklumat penting dalam data klinikal.

Kata Kunci: Pengekstrakan Entiti Nama, n2c2 2010, MedlinePlus, data klinikal, graf pengetahuan, BioDischargeSummaryBERT

I. PENGENALAN

Shortliffe (2012) menekankan beberapa isu dimana kos penjagaan kesihatan yang melambung tinggi, jumlah maklumat yang besar dan meningkat, profesional penjagaan kesihatan yang tidak sekata di

kawasan bandar dan jangkaan pesakit yang lebih tinggi daripada profesional penjagaan kesihatan. Sistem sokongan keputusan seperti MYCIN boleh menyelesaikan sebahagian daripada masalah tersebut ini. Salah satu komponen penting dalam sistem sokongan keputusan ialah pangkalan pengetahuan untuk menyimpan pengetahuan seseorang pakar. Terdapat banyak model yang boleh digunakan untuk membangunkan pangkalan pengetahuan seperti ontologi dan graf pengetahuan. Graf pengetahuan pada asasnya ialah struktur data graf yang digunakan untuk menyimpan maklumat dan komponen penting graf pengetahuan ialah ia memerlukan nod dan hubungan antara nod. Terdapat banyak kaedah yang boleh digunakan untuk mengenal pasti nod dan perhubungan seperti berasaskan peraturan, berasaskan pembelajaran mesin dan berasaskan pembelajaran mendalam.

BioBERT ialah Perwakilan Pengekod Dwi Arah domain khusus daripada Transformers (BERT) untuk domain perlombongan teks bioperubatan. BioClinicalBERT dan BioDischargeSummaryBERT adalah variasi BioBERT yang dibangunkan oleh Alsentzer et al. (2019) di mana asalnya dilatih menggunakan keseluruhan set data MIMIC-III. Kemudian ianya dilatih hanya pada bahagian ringkasan pelepasan set data MIMIC-III sahaja. Melalui penggunaan BioDischargeSummaryBERT, bahagian nod graf pengetahuan hanya mengandungi intipati ringkas nota klinikal yang diekstrak berdasarkan label “masalah”, “ujian” dan “rawatan”. Label tersebut akan menjadi hubungan graf pengetahuan antara nod topik dan nod diekstrak.

Kajian ini bertujuan untuk membangunkan graf pengetahuan perubatan data klinikal. Perisian Protege digunakan untuk visualisasi data dan menguji pencapaian maklumat. Dataset yang digunakan dalam kajian ini ialah n2c2 2010 dan MedlinePlus. Objektif utama kajian adalah seperti berikut dan *mapping* pernyataan masalah dan objektif kajian untuk menyelesaikan masalah tersebut ditunjukkan dalam Jadual 1:

1. Untuk menambah baik pengekstrakan entiti nama data klinikal menggunakan pelarasan halus BioDischargeSummaryBERT.

2. Untuk menjana pangkalan data pengetahuan berdasarkan entiti nama yang kenalpasti.

Jadual 1 Mapping pernyataan masalah dan objektif kajian

Penyataan Masalah	Objektif
Penyelidik terdahulu telah menggunakan pelbagai variasi BERT untuk mengekstrak entiti nama dari teks kesihatan. Namun begitu, kajian mengenai model BioDischargeSummaryBERT masih kurang dan boleh ditambah baik dengan menggubah suai nilai pelarasan halus.	Untuk menambah baik pengekstrakan entiti nama data klinikal menggunakan pelarasan halus BioDischargeSummaryBERT.

Terdapat isu keperluan dan kekurangan data jenis sumber terbuka dalam domain perubatan.	Untuk menjana pangkalan data pengetahuan berdasarkan entiti nama yang kenalpasti.

II. KAJIAN LITERASI

Berdasarkan Jadual 2, variasi BERT yang paling banyak digunakan untuk mengekstrak entiti nama ialah BioBERT. Kajian ini akan menggunakan variasi BERT iaitu BioDischargeSummaryBERT kerana model ini dilatih pada bahagian ringkasan pelepasan set data MIMIC-III menjadikannya sesuai untuk data aplikasi MedlinePlus.

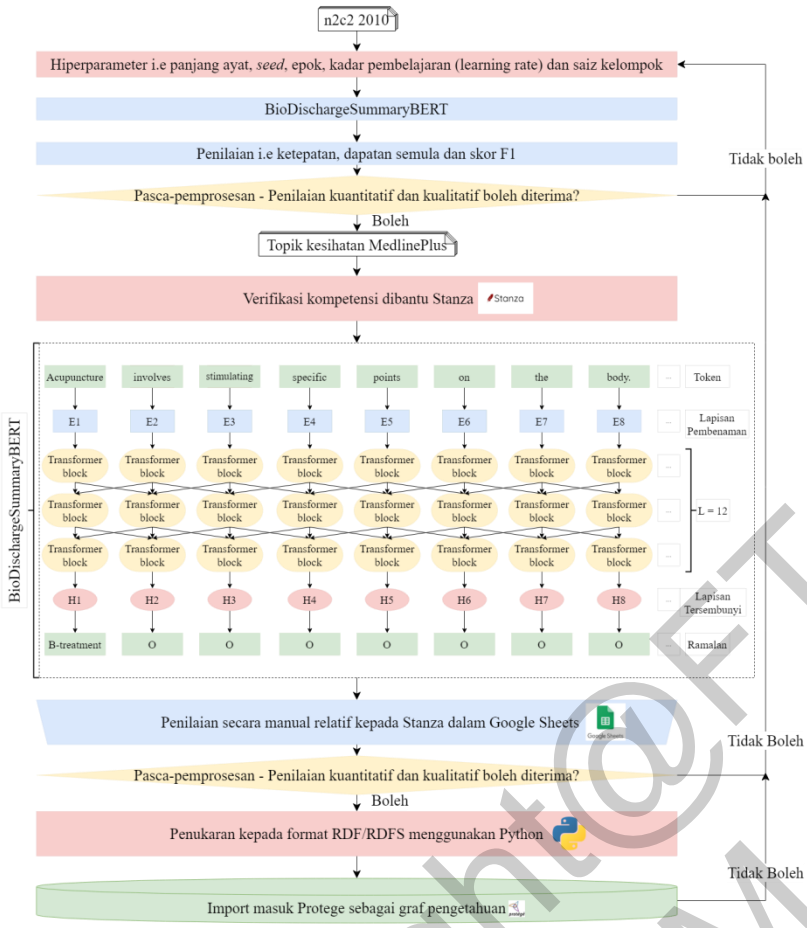
Jadual 2 Ringkasan kajian literasi

Bil	Kajian	Data	Teknik	Entiti Nama
1	Bozkurt et al. (2016)	300 laporan mamografi	Rangkaian Bayesian	Deskriptor BI-RADS dan maklumat klinikal pesakit
2	Chen et al. (2021)	*Peking University Hospital of Stomatology Electronic Dental Records*	BERT	Atribut gigi, nilai atribut gigi dan kedudukan gigi
3	Das et al. (2020)	Data Penyelidikan Terbuka COVID-19 (CORD-19)	Berasaskan BERT	Nod kertas penyelidikan
4	Harnoune et al. (2021)	505 nota klinikal dari MIMIC-III	Berasaskan BERT	Preskripsi ubat
5	Jin et al. (2019)	BC2GM, CoNLL 2003, MedNLI dan SNLI	Berasaskan BERT	Nama gen
10	Kim et al. (2021)	Artikel wiki yang dikikis dan set data berita	Berasaskan BERT	Covid-19
11	Kong et al. (2021)	MedMention, COMETA dan 3DNotes	Seni bina Siam	Perubatan
12	Lee et al. (2020)	PubMed, penyakit NCBI, GAD, BioASQ dan lain-lain	Berasaskan BERT	Penyakit, ubat/bahan kimia, gen/protein dan lain-lain
13	Liu et al. (2021)	PubMed	BERT-BiLSTM-CRF	Covid-19
15	Meng et al. (2021)	Sfull, PubMedQA, Inferens Bahasa Semulajadi, MedNLI dan lain-lain	Berasaskan BERT	Bioperubatan perkataan tunggal, gabungan atau frasa pendek seperti demam, sars-cov-2 atau telah mencari tapak
16	Mondal, I. (2021)	DrugBank, BioSNAP dan UniProt	Berasaskan BERT	Ubat, sasaran dan penyakit
17	Müller et al. (2019)	ICD-10 dan RX-Norm	Bayesian	Penyakit ICD-10, ubat-ubatan RX-Norm dan pemerhatian unik SNOMED atau LOINC
18	Naseem et al. (2021)	BC5DR, BC4CHEMD, penyakit NCBI dan lain-lain	Berasaskan BERT	Penyakit, kimia dan gen/protein

19	Prasad et al. (2021)	Data Penyelidikan Terbuka COVID-19 (CORD-19) untuk eksperimen dan BioNLP13cg untuk pelarasan halus	Berasaskan BERT	Bioperubatan seperti B-Kanser, I-Organisma dan lain-lain
22	Rodríguez-González et al. (2018)	MedlinePlus perubatan, Wikipedia dan lain-lain	MetaMap dan cTakes	Istilah diagnostik
23	Symeonidou et al. (2019)	*Text Analysis Conference 2017* (TAC2017), korpus *Adverse Drug Events* (ADE) dan set emas Elsevier	Berasaskan BERT	Reaksi Buruk Ubat
24	Wada et al. (2020)	Set data dalam bahasa Jepun yang dibangunkan sendiri bernama DocClsJp	Berasaskan BERT	Jangka sebutan
25	Wang et al. (2020)	*Text Analysis Conference Knowledge Base Population* (TAC KBP) dan Wikidata	Berasaskan BERT	Umum
26	Yang et al. (2021)	PubMed, *Traditional Chinese Medicine Database* (TCMID), *Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform* (TCMSP), *Encyclopedia of Traditional Chinese Medicine* (ETCM) dan lain-lain	Pembelajaran mendalam dan berasaskan peraturan	Penyakit, gen, simptom, ubat, laluan, herba, bahan kimia, ramuan dan ubat paten, jenis hubungan rawatan, sebab, kesan, asal entiti dan lain-lain
27	Yu et al. (2019)	2010 i2b2/VA	Berasaskan BERT	Masalah atau ujian
28	Zhu et al. (2020)	Korpus DDI, DrugBank dan Wikipedia	Berasaskan BERT	Dadah
29	Zhu et al. (2021)	BioCreative V dan Chemical Disease Relation Dataset	BERT	Sebutan format PMID

III. METODOLOGI

Rajah 1 menunjukkan rangka kerja untuk kajian ini yang dibahagikan kepada dua peringkat utama. Peringkat pertama bermula dengan peringkat pelarasan halus menggunakan set data n2c2 2010. Tujuan analisis di peringkat pertama adalah untuk mendapatkan nilai pelarasan halus hiperparameter yang paling optimum bagi mengekstrak entiti nama dengan menggunakan BioDischargeSummaryBERT. Peringkat kedua merupakan peringkat aplikasi menggunakan set data topik kesihatan MedlinePlus. Tujuan analisis di peringkat kedua adalah menguji keberkesanan model yang dijana di peringkat pertama. Keseluruhannya, terdapat 6 fasa dalam kajian ini iaitu pengumpulan data, pengecaman entiti nama, penilaian, pasca-pemprosesan, penjanaan format RDF/RDFS dan graf pengetahuan.



Rajah 1 Rangka Kerja Penyelidikan

A. Fasa 1 – Pengumpulan Data

Kajian ini menggunakan dua set data. Set data pertama adalah n2c2 2010 yang merupakan subset MIMIC-II. Set data ini merupakan set data piawai bagi Pengekstrakan Entiti Nama dalam bidang perubatan. Entiti nama yang menjadi fokus dalam kajian ini adalah "masalah", "ujian" dan "rawatan". Data kedua yang digunakan dalam kajian ini diperolehi secara manual dari MedlinePlus. Data ini akan digunakan untuk menguji keberkesanan model pengekstrakan entiti yang dibangunkan.

n2c2 2010 adalah set data subset nota klinikal MIMIC-II. Set data ini dipilih kerana ia mempunyai data untuk melatih model mengenal pasti entiti nama "masalah", "ujian" dan "rawatan" yang sesuai untuk topik kesihatan MedlinePlus. Dataset ini juga digunakan oleh Chalapathy et al. (2016). Jadual 3 menunjukkan bilangan label untuk "masalah" melebihi "ujian" dan "rawatan" dengan nisbah latihan/ujian 90/10. Set data yang tidak seimbang, skor F1 ialah metrik yang lebih sesuai berbanding *Accuracy*.

Jadual 3 Statistik Data n2c2 2010 Kajian Semasa

Label	Latihan	Ujian
ayat	3110	251
masalah	3171	334
ujian	1546	159
rawatan	1865	121
O	22623	1855

Set data aplikasi ialah gabungan set data yang dimuat turun daripada Bioportal dan MedlinePlus. Set data ini mengandungi senarai topik kesihatan dari abjad A hingga Z. Set data dipilih berdasarkan panjang ayat iaitu tidak lebih dari 128 perkataan. Set data yang dipilih mengandungi 246 topik kesihatan dan dalam kajian ini, hanya bahagian ringkasan set data digunakan.

B. Fasa 2 – Pengecaman Entiti Nama Menggunakan BioDischargeSummaryBERT

Pelarasan halus hiperparameter dilakukan untuk mendapatkan nilai paling optimal untuk memastikan model yang digunakan dapat menghasilkan hasil terbaik. Proses pemilihan nilai hiperparameter dilakukan secara manual bermula dengan penggunaan nilai *intial* yang sebahagiannya diperolehi daripada kajian Harnoune et al. (2021) dan yang lain adalah nilai lalai pakej *Simple Transformers*. Antara hiperparameter yang dikaji ialah *seed*, panjang ayat, epok, kadar pembelajaran, dan saiz kelompok. Jadual 4 menunjukkan nilai *initial* dan nilai optimum hiperparameter yang digunakan dalam kajian ini untuk peringkat pelarasan halus dan hanya nilai optimum hiperparameter digunakan untuk peringkat aplikasi. Nilai optimum ini adalah nilai optimum yang telah diperolehi berdasarkan skor F1 terbaik. Hiperparameter yang dikaji akan diubah manakala hiperparameter yang lain akan kekal dengan nilai masing-masing.

Jadual 4 Nilai Initial dan Nilai Optimum Hiperparameter

Kunci	Nilai *Initial*	Nilai Optimum	Deskripsi
manual_seed	1	4	Nilai supaya hasil model latihan boleh dihasilkan semula dan tidak rambang
reprocess_input_data	True	True	Memproses semula data yang dimasukkan jika nilai ialah *True*
overwrite_output_dir	True	True	Menggantikan hasil daripada model latihan terdahulu jika nilai ialah *True*
sliding_window	False	False	Tidak pecahkan teks menggunakan kaedah *chunk* yang mempunyai label dari teks asal jika nilai ialah *False*

max_seq_length	128	128	Nilai maksimum untuk panjang ayat
num_train_epochs	5	2	Nilai epok untuk latihan
train_batch_size	17	2	Nilai saiz kelompok untuk latihan
eval_batch_size	17	2	Nilai saiz kelompok untuk penilaian
learning_rate	4e-5	4e-5	Nilai kadar pembelajaran
fp16	True	True	Kejituan *floating point* sebanyak 16-bit jika nilai ialah *True*
output_dir	./outputs	./outputs	Direktori untuk hasil
best_model_dir	./best_model	./best_model	Direktori untuk model terbaik
evaluate_during_training	True	True	Menjalankan penilaian semasa melatih model jika nilai ialah *True*
do_lower_case	False	False	Tidak menjadikan data yang dimasukkan sebagai huruf kecil jika nilai ialah *False*

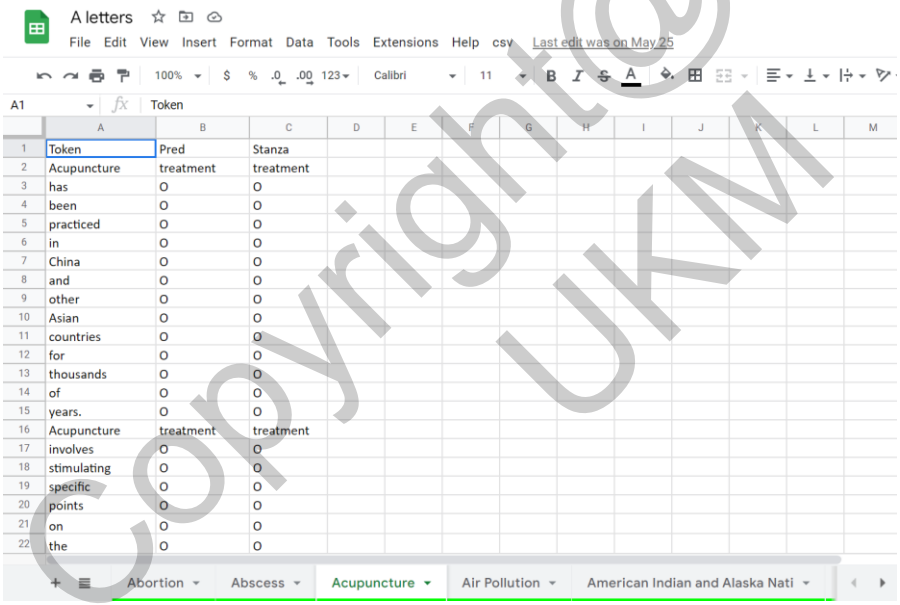
Pengecaman Entiti Nama boleh dijelaskan selanjutnya dengan menggunakan Rajah 1, untuk topik kesihatan "Acupuncture", ayat pertama untuk rumusan topik tersebut adalah "Acupuncture involves stimulating specific points on the body." Dalam BioDischargeSummaryBERT, ayat ini akan ditukar menjadi token dalam kotak berwarna hijau cair iaitu "Acupuncture", "involves", "stimulating", "specific", "points", "on", "the" dan "body.". Selepas itu, token-token ini akan dimasukkan ke dalam kotak lapisan pembenaman yang berwarna biru cair untuk ditukarkan kepada format nombor yang dikenali sebagai vektor. 12 lapisan *Transformer block* dalam bentuk bujur yang berwarna kuning cair bertujuan untuk proses vektor daripada lapisan pembenaman untuk mendapatkan nilai skor perhatian. Skor perhatian adalah penting untuk hanya memberi perhatian kepada label yang berkenaan dengan kajian ini iaitu "masalah", "ujian" dan "rawatan". Selepas itu, skor perhatian akan melalui *activation function* di dalam lapisan tersembunyi bentuk bujur yang berwarna merah cair di mana ramalan dalam kotak yang berwarna hijau cair akan keluar sebagai entiti nama berlabel, entiti nama yang keluar dari sini akan membawa label sama ada "masalah", "ujian" atau "rawatan" dalam format *Beginning, Inner* dan *Outer* (BIO) seperti entiti nama "Acupuncture" dalam Rajah 1 yang membawa label "B-rawatan" dan entiti nama yang lain membawa label O.

Dalam Rajah 1, hanya entiti nama "Acupuncture" diramalkan sebagai label "B-treatment" yang merupakan label yang betul, ini kerana model BioDischargeSummaryBERT telah dilaras halus sehingga mempunyai skor F1 sebanyak 0.87 dan nilai optimum hiperparameternya ditunjukkan dalam Jadual 4 namun jika nilai initial hiperparameter digunakan yang mempunyai skor F1 bernilai 0.81 entiti nama "Acupuncture" yang sebelum ini membawa label "B-rawatan" sebagai contoh mungkin akan membawa label lain seperti label O atau "B-masalah".

C. Fasa 3 - Penilaian

Dalam kajian ini, terdapat dua jenis penilaian iaitu kualitatif dan kuantitatif, hanya penilaian kuantitatif dijalankan untuk peringkat pelarasan halus sementara kedua-dua penilaian kuantitatif dan kualitatif dijalankan untuk peringkat aplikasi.

Nilai kejituan, dapatan semula dan skor F1 digunakan untuk penilaian kuantitatif. Penilaian set data peringkat pertama yang menggunakan set data n2c2 2010 dikira selepas entiti nama diekstrak. Label entiti nama yang diekstrak dibandingkan dengan nilai label entiti nama yang telah dilabelkan dalam dataset tersebut. Penilaian set data pada peringkat kedua yang menggunakan set data topik kesihatan MedlinePlus dijalankan selepas entiti nama diekstrak menggunakan BioDischargeSummaryBERT dengan pelarasan halus yang optimum. Label entiti nama yang dikenalpasti kemudiannya dibandingkan dengan label entiti nama yang diekstrak dengan menggunakan aplikasi Stanza seperti dalam Rajah 2.



	Token	Pred	Stanza
2	Acupuncture	treatment	treatment
3	has	O	O
4	been	O	O
5	practiced	O	O
6	in	O	O
7	China	O	O
8	and	O	O
9	other	O	O
10	Asian	O	O
11	countries	O	O
12	for	O	O
13	thousands	O	O
14	of	O	O
15	years.	O	O
16	Acupuncture	treatment	treatment
17	involves	O	O
18	stimulating	O	O
19	specific	O	O
20	points	O	O
21	on	O	O
22	the	O	O

Rajah 2 Entiti yang Diekstrak Menggunakan Biodischargesummarybert dan Stanza Dicantum dalam Google Sheets

Penilaian kualitatif dilakukan bagi menguji keberkesanan graf pengetahuan yang dijana berdasarkan entiti nama yang telah diekstrak. Jadual 5 menunjukkan soalan kecekapan yang digunakan untuk menilai secara kualitatif graf pengetahuan.

Jadual 5 Soalan Kecekapan untuk Set Data Aplikasi

Soalan kecekapan set data MedlinePlus
1. Apakah entiti nama masalah bagi abjad A, N, M dan W?
2. Apakah entiti nama ujian untuk abjad A, N, M dan W?

3. Apakah entiti nama rawatan untuk abjad A, N, M dan W?
4. Apakah entiti nama masalah, ujian dan rawatan untuk COVID-19?
5. Berapakah bilangan entiti nama masalah?
6. Berapakah bilangan entiti nama ujian?
7. Berapakah bilangan entiti nama rawatan?

D. Fasa 4 – Pasca-Pemprosesan

Fasa pasca-pemprosesan akan memastikan hasil penilaian skor F1 bagi peringkat pertama dan peringkat kedua melebihi ambang nilai 0.84 (Chalapathy et al. 2016) dan 0.87 (Harnoune et al. 2021). Ini kerana ada persamaan dari segi data dan model yang digunakan dari segi data dan model. Sekiranya analisis mendapat nilai skor F1 yang lebih rendah dari nilai-nilai tersebut, nilai hiperparameter akan dilaras sehingga analisis berjaya mencapai nilai skor F1 yang lebih tinggi bagi mendapatkan nilai hiperparameter yang optimum untuk kajian ini.

Tujuh soalan kompetensi seperti dalam Jadual 5 digunakan semasa peringkat penilaian iaitu sebelum set data MedlinePlus dimasukkan ke dalam BioDischargeSummaryBERT dibantu Stanza dan selepas import masuk Protege dengan menggunakan SPARQL. Proses sebelum dimasukkan ke dalam BioDischargeSummaryBERT adalah dengan mencipta *Gold Standard* untuk setiap soalan kecekapan menggunakan pengetahuan saya sebagai jurutera ontologi dan dibantu Stanza manakala proses selepas import masuk Protege adalah dengan menukarkan soalan kecekapan dalam Jadual 5 kepada format SPARQL. Soalan kecekapan digunakan untuk menilai keberkesanan graf pengetahuan yang dibina. *Gold Standard* yang dijana dibantu oleh Stanza adalah untuk verifikasi dan perbandingannya iaitu SPARQL adalah untuk validasi. Jika jawapan tidak memuaskan maka nilai optimum hiperparameter dalam Jadual 4 dari peringkat pelarasan halus model ini akan dilaras sehingga jawapan yang memuaskan diperolehi.

E. Fasa 5 – Penjanaan Format RDF/RDFS

Fasa ini hanya dilakukan semasa peringkat kedua. Entiti berlabel “masalah”, “ujian” dan “rawatan” yang telah diekstrak menggunakan BioDischargeSummaryBERT seperti dalam Rajah 2 ditukar kepada format RDF/RDFS yang memerlukan *prefix* ontologi, kelas topik kesihatan, individu, *properties* data dan *triple* mengikut label masing-masing.

Entiti nama yang telah diekstrak dalam format CSV seperti Rajah 2 dalam seksyen III. subseksyen C akan memenuhi bahagian objek dalam set *triple* yang merangkumi subjek, predikat dan objek. Contohnya, entiti nama untuk triple pertama yang diekstrak untuk topik kesihatan

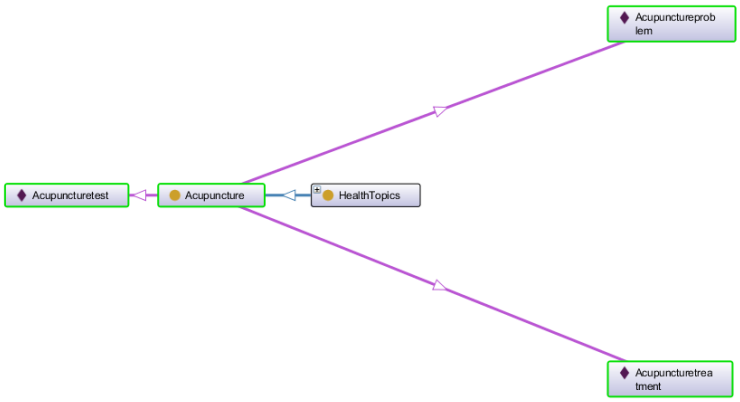
“Acupuncture” yang berlabel rawatan ialah “Acupuncture”. Entiti nama “Acupuncture” akan memenuhi bahagian objek iaitu "Acupuncture"^^xsd:string dalam *triple* yang merangkumi mlplusht:Acupuncturetreatment mlplusht:hasTreatment "Acupuncture"^^xsd:string. Dalam *triple* yang ditunjukkan mlplusht:Acupuncturetreatment ialah subjek *triple* tersebut manakala mlplusht:hasTreatment pula ialah predikat *triple* tersebut.

F. Fasa 6 – Penjanaan Graf Pengetahuan

Fasa ini menjana graf pengetahuan berdasarkan entiti yang diekstrak daripada topik kesihatan MedlinePlus dengan menggunakan aplikasi Protege. Protege dipilih kerana ia adalah dari sumber terbuka berbanding Neo4j. Rajah 3 menunjukkan contoh RDF/RDFS yang mengandungi prefix ontologi, kelas topik kesihatan, individu, properties data dan triple yang dijana secara automatik manakala Rajah 4 menunjukkan contoh graf pengetahuan yang divisualisasikan dalam Protege menggunakan topik kesihatan MedlinePlus.

```
#Prefixes
@prefix mlplusht: <http://www.ftsm.ukm.my/mlplushtOnt#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

#Classes and SubClasses
mlplusht:HealthTopics rdf:type rdfs:Class .

#Acupuncture Class and SubClasses
mlplusht:Acupuncture rdf:type rdfs:Class .
mlplusht:Acupuncture rdfs:subClassOf mlplusht:HealthTopics .

#Individuals
mlplusht:Acupuncturetreatment rdf:type mlplusht:Acupuncture .
mlplusht:Acupunctureproblem rdf:type mlplusht:Acupuncture .
mlplusht:Acupuncturetest rdf:type mlplusht:Acupuncture .

#Properties
mlplusht:hasTreatment rdf:type rdf:Property .
mlplusht:hasTreatment rdfs:range xsd:string .
mlplusht:hasProblem rdf:type rdf:Property .
mlplusht:hasProblem rdfs:range xsd:string .
mlplusht:hasTest rdf:type rdf:Property .
mlplusht:hasTest rdfs:range xsd:string .

#Triples
mlplusht:Acupuncturetreatment mlplusht:hasTreatment "Acupuncture"^^xsd:string .
mlplusht:Acupuncturetreatment mlplusht:hasTreatment "Acupuncture"^^xsd:string .
mlplusht:Acupuncturetreatment mlplusht:hasTreatment "thin needles"^^xsd:string .
mlplusht:Acupuncturetreatment mlplusht:hasTreatment "skin,"^^xsd:string .
mlplusht:Acupuncturetreatment mlplusht:hasTreatment "acupuncture"^^xsd:string .
mlplusht:Acupunctureproblem mlplusht:hasProblem "nausea"^^xsd:string .
mlplusht:Acupunctureproblem mlplusht:hasProblem "vomiting"^^xsd:string .
mlplusht:Acupuncturetreatment mlplusht:hasTreatment "surgery"^^xsd:string .
mlplusht:Acupuncturetreatment mlplusht:hasTreatment "chemotherapy."^^xsd:string .
mlplusht:Acupunctureproblem mlplusht:hasProblem "pain."^^xsd:string .
mlplusht:Acupuncturetreatment mlplusht:hasTreatment "acupuncture"^^xsd:string .
mlplusht:Acupunctureproblem mlplusht:hasProblem "pain-killing chemicals."^^xsd:string .
mlplusht:Acupuncturetest mlplusht:hasTest "blood pressure"^^xsd:string .
```

Rajah 3 Contoh RDF/RDFS Sebelum Dimasukkan ke dalam Protege



Rajah 4 Contoh Graf Pengetahuan Topik Kesihatan Medlineplus Divisualisasikan Menggunakan Protege

IV. KEPUTUSAN DAN PERBINCANGAN

Secara keseluruhan, kajian ini telah memperolehi skor median F1 0.91 dengan markah skor F1 terendah dan tertinggi masing-masing ialah 0.71 dan 1.00. Hasil pelarasan halus menunjukkan bahawa, pelarasan halus membantu untuk menurunkan nilai False Positive (FP) dalam formula kejituan dan *False Negative* (FN) dalam formula dapatan semula yang akan meningkatkan skor F1 kerana skor F1 adalah purata kejituan dan dapatan semula.

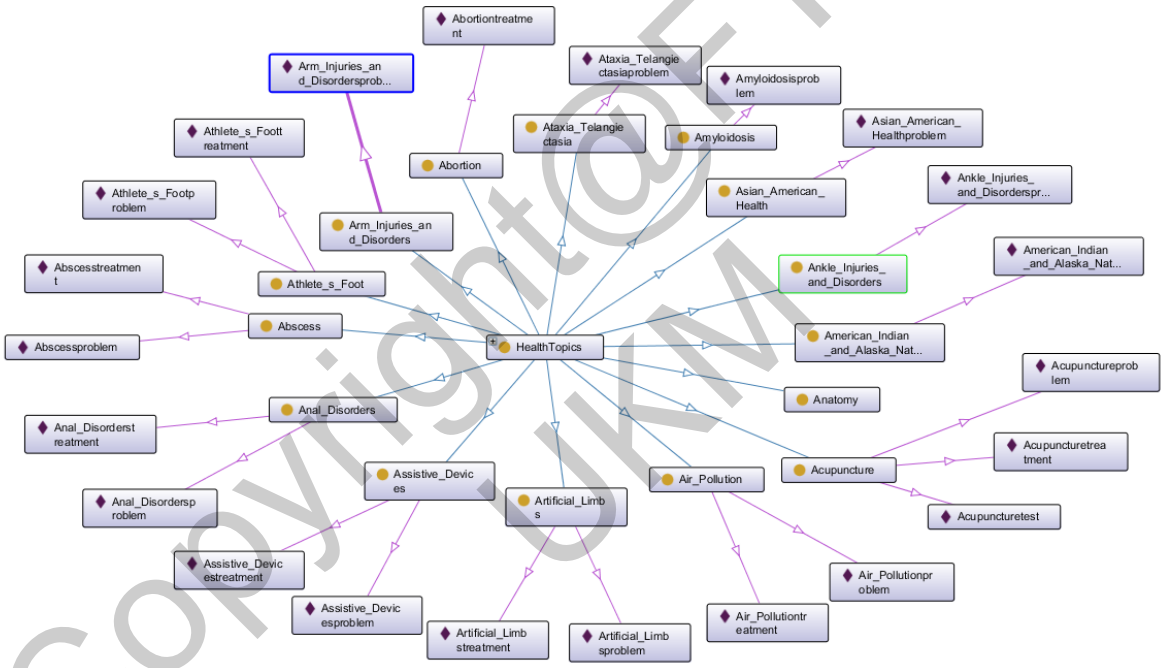
Hasil aplikasi menunjukkan bahawa terdapat senario bergantung kepada konteks di mana model kajian semasa adalah betul. Untuk topik kesihatan "*Abortion*", model kajian semasa adalah betul untuk tidak ramal "*An*", "*abortion*", "*an*" dan "*abortion*" sebagai label betul O dan sebaliknya sebagai label ramalan rawatan.

Model kajian semasa adalah betul juga untuk tidak ramal token-token "*infected*", "*area,*", "*the*", "*damaged*", "*tissue,*" dan "*living*" sebagai label betul O manakala token "*white*", "*blood*" dan "*cells,*" tidak dapat diramalkan sebagai label betul ujian untuk topik kesihatan "Abscess". Selain itu, untuk topik kesihatan "*Acupuncture*", model kajian semasa adalah betul untuk tidak ramal token-token "*thin*", "*needles*", "*skin*" dan "*acupuncture*" sebagai label betul O manakala token "*acupuncture*" tidak dapat diramalkan sebagai label betul masalah.

Di samping itu, dalam konteks topik kesihatan "*Anatomy*", model kajian semasa adalah betul untuk tidak ramal token-token "*organ*" dan "*systems*" sebagai label betul ujian dan sebaliknya sebagai label ramalan O. Akhirnya, dalam konteks topik kesihatan "*Athlete's Foot*", model kajian semasa adalah betul juga untuk tidak ramal token "*flip-flops*" sebagai label betul O dan sebaliknya sebagai label ramalan rawatan.

Terdapat tiga senario iaitu padanan tepat, padanan separa dan padanan sifar. *Gold Standard* soalan satu, tiga dan empat berbanding SPARQL *query* tergolong dalam padanan separa. *Gold Standard* soalan dua berbanding SPARQL *query* pula tergolong dalam padanan sifar manakala *Gold Standard* soalan lima, enam dan tujuh berbanding SPARQL *query* pula tergolong dalam padanan tepat.

Graf pengetahuan yang divisualisasikan dalam Protege terdiri daripada empat variasi label untuk setiap topik kesihatan contohnya topik kesihatan "*Abortion*" dengan salah satu label "rawatan" sahaja, topik kesihatan "*Acupuncture*" dengan tiga label "masalah", "ujian" dan "rawatan", topik kesihatan "*Anatomy*" tanpa label dan topik kesihatan "*Athlete's Foot*" dengan dua kombinasi label "masalah" dan "rawatan".



Rajah 5 Topik Kesihatan Abjad A

V. RUMUSAN DAN CADANGAN

Kajian ini telah menggunakan BioDischargeSummaryBERT untuk mengekstrak entiti nama dalam bidang perubatan. Set data latihan yang digunakan ialah n2c2 2010. Tiga entiti nama yang difokuskan dalam kajian ini ialah masalah, ujian dan rawatan. Model yang dibina kemudiannya diuji dengan topik kesihatan dari MedlinePlus yang mengandungi ayat ringkas tidak lebih daripada 128 perkataan bagi setiap topik. Graf pengetahuan dijana dengan menggunakan perisian Protege. Soalan kecekapan digunakan untuk menilai keberkesanan graf pengetahuan yang dibina.

Objektif pertama kajian ini telah dicapai iaitu menambah baik pengekstrakan entiti nama data klinikal menggunakan pelarasan halus BioDischargeSummaryBERT. Objektif ini dicapai dengan menggunakan data sensitif yang terhad iaitu n2c2 2010 untuk pelarasan halus untuk mendapatkan nilai optimum bagi hiperparameter seperti *seed*, panjang ayat, epok, kadar pembelajaran dan saiz kelompok. Menggunakan data n2c2 2010, kajian semasa berjaya melatih model BioDischargeSummaryBERT tanpa Medan Rawak Bersyarat (MRB) dengan skor ujian F1 0.87 berbanding skor F1 bernilai 0.84 menggunakan model BiLSTM dengan MRB oleh Chalapthy et al. (2016), setanding model BioDischargeSummaryBERT dengan MRB yang mempunyai skor F1 bernilai 0.87 oleh Harnoune et al. (2021) dan hampir setanding dengan model klinikal Stanza yang mempunyai skor F1 bernilai 0.88 seperti dalam Jadual 6. Model yang dibina kemudiannya diuji dengan topik kesihatan dari MedlinePlus yang mengandungi ayat ringkas tidak lebih daripada 128 perkataan bagi setiap topik.

Jadual 6 Jadual Perbandingan Skor F1

Model	Skor F1
BioDischargeSummaryBERT tanpa MRB sebelum pelarasan halus (Kajian semasa)	0.81
BiLSTM dengan MRB (Chalapathy et al. 2016)	0.84
BioDischargeSummaryBERT tanpa MRB selepas pelarasan halus (Kajian semasa)	0.87
Harnoune et al. (2021)	0.87
Stanza (Zhang et al. 2021)	0.88

Objektif kedua kajian ini iaitu menjana pangkalan data pengetahuan berdasarkan entiti nama yang kenalpasti juga dicapai, graf pengetahuan dijana dengan menggunakan perisian sumber terbuka Protege yang mengandungi 246 topik kesihatan. Entiti nama yang telah diekstrak dalam format CSV akan memenuhi bahagian objek dalam set *triple* yang merangkumi subjek, predikat dan objek. Contohnya, entiti nama untuk triple pertama yang diekstrak untuk topik kesihatan “Acupuncture” yang berlabel rawatan ialah “Acupuncture”. Entiti nama “Acupuncture” akan memenuhi bahagian objek iaitu "Acupuncture"^^xsd:string dalam *triple* yang merangkumi mlplusht:Acupuncturetreatment mlplusht:hasTreatment "Acupuncture"^^xsd:string.

Kesimpulannya, berdasarkan eksperimen yang dijalankan ke atas BioDischargeSummaryBERT dan graf pengetahuan data klinikal, model kajian semasa mempunyai skor F1 0.87 pada peringkat pelarasan halus manakala model mempunyai skor median F1 0.91 pada peringkat aplikasi.

Terdapat tiga cadangan yang akan membantu untuk menambah baik kerja masa hadapan. Cadangan pertama, untuk melakukan perbandingan selanjutnya menggunakan model Pengecaman Entiti Nama terkemuka yang lain seperti RoBERTa, model berasaskan spaCy atau GPT. Cadangan

kedua, untuk menambah baik lagi model dengan menggabungkan Bio_Discharge_Summary_BERT semasa dengan model berpotensi lain seperti LSTM atau GRU. Cadangan ketiga, untuk menguji lagi keberkesanan pada pangkalan data MedlinePlus lain seperti Ensiklopedia Perubatan dan Ujian Perubatan.

PENGHARGAAN

Pertama sekali, saya ingin merakamkan ucapan terima kasih kepada Dr Lailatul Qadri Binti Zakaria atas tunjuk ajar beliau yang banyak membantu saya menyiapkan kajian ini dengan jayanya. Selain itu, saya juga ingin berterima kasih kepada beliau atas kesabaran dan bimbingan yang berharga dalam penyelidikan ini. Terima kasih khas ditujukan kepada ibu saya dan semua orang yang menyokong saya.

RUJUKAN

Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T. & McDermott, M. 2019. Electronic sources: Publicly Available Clinical BERT Embeddings. *arXiv preprint arXiv:1904.03323* [12 February 2022].

Bozkurt, S., Gimenez, F., Burnside, E.S., Gulkesen, K.H. & Rubin, D.L. 2016. Using automatically extracted information from mammography reports for decision-support. *Journal of Biomedical Informatics* 62: 224–231.

Chalapathy, R., Borzeshi, E. Z., & Piccardi, M. 2016. Electronic sources: Bidirectional LSTM-CRF for clinical concept extraction. *arXiv preprint arXiv:1611.08373* [5 June 2022].

Chen, Q., Zhou, X., Wu, J. & Zhou, Y. 2021. Structuring electronic dental records through deep learning for a clinical decision support system. *Health Informatics Journal* 27: 1–18.

Das, D., Katyal, Y., Verma, J., Dubey, S., Singh, A.D., Agarwal, K., Bhaduri, S. & Ranjan, R.K. 2020. Information Retrieval and Extraction on COVID-19 Clinical Articles Using Graph Community Detection and Bio-BERT Embeddings. *Association for Computational Linguistics 2020 Workshop NLP-COVID Submission*, hlm: 1-9.

Harnoune, A., Rhanoui, M., Mikram, M., Yousfi, S., Elkaimbillah, Z. & El Asri, B. 2021. BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis. *Journal of Computer Methods and Programs in Biomedicine Update* 1: 1-13.

Jin, Q., Dhingra, B., Cohen, W. & Lu, X. 2019. Electronic sources: Probing Biomedical Embeddings from Language Models. *arXiv preprint arXiv:1904.02181* [12 February 2022].

Kim, T., Yun, Y. & Kim, N. 2021. Deep learning-based knowledge graph generation for covid-19. *Journal of Sustainability* 13(4): 1–19.

Kong, L., Winestock, C. & Bhatia, P. 2021. Electronic sources: Zero-shot Medical Entity Retrieval without Annotation: Learning From Rich Knowledge Graph Semantics. *arXiv preprint arXiv:2105.12682* [12 February 2022].

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. & Kang, J. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Journal of Bioinformatics* 36(4): 1234–1240.

Liu, H., Sun, Y., & Cao, S. 2021, March. Building a COVID-19 Literature Knowledge Graph Based on PubMed. *International Conference on Medical Imaging and Computer-Aided Diagnosis 2021*, hlm: 364-373.

Meng, Z., Liu, F., Clark, T.H., Shareghi, E. & Collier, N. 2021. Electronic sources: Mixture-of-Partitions: Infusing Large Biomedical Knowledge Graphs into BERT. *arXiv preprint arXiv:2109.04810* [12 February 2022].

Mondal, I. 2021. Electronic sources: BERTKG-DDI: Towards incorporating entity-specific knowledge graph information in predicting drug-drug interactions. *arXiv preprint arXiv:2012.11142* [12 February 2022].

Müller, L., Gangadharaiah, R., Klein, S.C., Perry, J., Bernstein, G., Nurkse, D., Wailes, D., Graham, R., El-Kareh, R., Mehta, S., Vinterbo, S.A. & Aronoff-Spencer, E. 2019. An open access medical knowledge base for community driven diagnostic decision support system development. *Journal of BMC Medical Informatics and Decision Making* 19: 1–7.

Naseem, U., Musial, K., Eklund, P. & Prasad, M. 2020. Biomedical Named-Entity Recognition by Hierarchically Fusing BioBERT Representations and Deep Contextual-Level Word-Embedding. *Proceedings of the International Joint Conference on Neural Networks 2020*, hlm: 1-8.

Prasad, V. K., Bharti, S., & Koganti, N. 2021. Entity-Based Knowledge Graph Information Retrieval for Biomedical Articles. *2nd International Conference on Communication and Intelligent Systems 2021*, hlm: 803-812.

Rodríguez-González, A., Costumero, R., Martinez-Romero, M., Wilkinson, M.D. & Menasalvas-Ruiz, E. 2018. Extracting Diagnostic Knowledge from MedLine Plus: A Comparison between MetaMap and cTAKES Approaches. *Journal of Current Bioinformatics* 13(6): 573–582.

Shortliffe, E. H. (Ed.). 2012. *Computer-based medical consultations: MYCIN*. Amsterdam: Elsevier.

Symeonidou, A., Sazonau, V. & Groth, P. 2019. Transfer learning for biomedical named entity recognition with BioBert. *Proceedings of the Posters and Demo Track of the 15th International Conference on Semantic Systems 2019*, hlm: 1–5.

Wada, S., Takeda, T., Manabe, S., Konishi, S., Kamohara, J. & Matsumura, Y. 2020. Electronic sources: Pre-training technique to localize medical BERT and enhance biomedical BERT. *arXiv preprint arXiv:2005.07202* [12 February 2022].

Wang, C., Liu, X. & Song, D. 2020. Electronic sources: Language Models are Open Knowledge Graphs. *arXiv preprint arXiv:2010.11967* [12 February 2022].

Yang, X., Wu, C., Nenadic, G., Wang, W. & Lu, K. 2021. Mining a stroke knowledge graph from literature. *Journal of BMC Bioinformatics* 22: 1–18.

Yu, X., Hu, W., Lu, S., Sun, X. & Yuan, Z. 2019. BioBERT based named entity recognition in electronic medical record. *10th International Conference on Information Technology in Medicine and Education 2019*, hlm: 49–52.

Zhang, Y., Zhang, Y., Qi, P., Manning, C. D. & Langlotz, C. P. 2020. Electronic sources: Biomedical and clinical English model packages for the Stanza Python NLP library. *arXiv preprint arXiv:2007.14640v1* [5 June 2022].

Zhu, X., Zhang, L., Du, J., & Xiao, Z. 2021. Full-Abstract Biomedical Relation Extraction with Keyword-Attentive Domain Knowledge Infusion. *Journal of Applied Sciences* 11: 1-8.

Zhu, Y., Li, L., Lu, H., Zhou, A., & Qin, X. 2020. Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions. *Journal of Biomedical Informatics* 106: 1-8.