PENGECAMAN ENTITI NAMA TEKS AL-QURAN TERJEMAHAN BAHASA MELAYU MENGGUNAKAN PENDEKATAN BERASASKAN CONDITIONAL RANDOMS FIELD (CRF)

Lily Haryanti Bt. Az.Muzni, Dr. Saidah Saad

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

P106508@siswa.ukm.edu.my, saidah@ukm.edu.my

ABSTRAK

Setiap anutan dan agama masing-masing pasti mempunyai kitab sebagai rujukan dan panduan untuk para penganutnya. Al-Quran merupakan kitab suci bagi umat Islam di mana ianya merupakan mukjizat teragung yang diturunkan kepada Nabi Muhammad S.A.W.; yang ditulis, dimushafkan dan diriwayatkan dengan mutawir dan membacanya adalah ibadah. Al-Quran juga adalah kitab suci yang wajib dipelajari oleh seluruh umat Islam di dunia ini. Perkembangan teknologi digital pastinya akan mempengaruhi disiplin ilmu dan salah satunya adalah sains agama. Al-Quran merupakan teks klasik yang mempunyai ciri-ciri yang unik dan berbeza dari teks biasa. Pengecaman Entiti Nama telah digunakan secara meluas dan memainkan peranan penting dalam bidang yang berkaitan Pemprosesan Bahasa Tabii. Kajian ini bertujuan untuk menambah entiti nama bagi Domain Islam berdasarkan kajian lepas dan indeks Al-Quran dan menganotasikannya. Selain itu, kajian ini turut mencari bagaimana cara melatih dan mengemaskini model PEN bagi mendapatkan nilai parameter optimum bagi Domain Islam. Pengecaman Entiti Nama adalah melalui pendekatan berasaskan Conditional Random Field (CRF) dan diaplikasikan melalui perpustakaan SpaCy melalui perlabelan secara jujukan. SpaCy menggabungkan pendekatan berasaskan CRF dengan vektor perkataan yang telah dilatih dan ciri-ciri lingustik lain seperti penandaan golongan kata untuk meningkatkan prestasi PEN. Ciri-ciri ini, bersama dengan lapiran CRF, menyumbang kepada keupayaan pengecaman entiti nama dengan lebih tepat dan cekap. Dataset yang digunakan adalah surah Al-Anbiyaa, 122 ayat. Kajian ini melibatkan penyediaan data, pra-pemprosesan data, penganotasian entiti nama, melatih model PEN dan pengujian yang dirangkumkan dalam rekabentuk metodologi kajian. Hasil rumusan kajian menunjukkan bahawa objektif kajian ini telah berjaya dicapai dalam skop yang ditentukan apabila hasil kajian serta pengujian terhadap model PEN yang telah dilatih dalam Domain Islam menunjukkan keputusan yang positif. Kekangan kajian dibentang bagi membolehkan para penyelidik menambahbaik kajian ini dari semasa ke semasa. Cadangan kajian bagi penyelidikan pada masa akan datang telah dihuraikan agar kajian ini menjadi bermanfaat dan lebih diperluaskan bagi menjadi panduan kepada pra penyelidik yang lain seterusnya kepada umat Islam.

Keyword: Pengecaman Entiti Nama, Al-Quran terjemahan Bahasa Melayu, pembelajaran mesin, Conditional Randoms Field, CRF

I. PENGENALAN

Dalam era digital ini, semuanya sedang menuju ke arah pendigitalan; dengan kemunculan pelbagai platform media sosial seperti Twitter, Instagram, TikTok dan sebagainya dengan kapasiti data dalam jumlah yang besar dan pelbagai sama ada dalam bentuk teks, imej, audio mahupun video, dijana dan dikongsi secara berterusan dan berleluasa tanpa had.

Merujuk kepada International Data Corporation (IDC), menjelang 2025, jumlah data yang dicipta, dikumpul, disalin dan digunakan secara global diramalkan meningkat dengan pesatnya daripada 64.2 Zettabait pada tahun 2020 kepada lebih 180 Zettabait. Sebahagian besar perkembangan data yang tidak berkesudahan ini adalah terdiri daripada data yang tidak berstruktur menyumbang kepada 95% daripada data global pada tahun 2020 (D. Reinsel et al., 2018). Kebanyakan ciri-ciri data tidak berstruktur in sering ditemui dalam bentuk atau format yang tiada piawaian, dipenuhi dengan teks yang berat dan ianya daripada sumber yang pelbagai (K. Adnan et al., 2019).

Pemprosesan Bahasa Tabii (PBT) merupakan satu bidang yang sangat penting pada masa kini. Ianya merupakan salah satu cabang Kecerdasan Buatan (Artificial Inteligence) yang bertumpu pada pengolahan bahasa tabii. Bahasa tabii adalah bahasa manusia yang perlu difahami oleh komputer sebagai medium komunikasi antara mesin dan manusia di mana ianya hanya difahami oleh manusia. Data-data tidak berstruktur ini jika dianalisis berkemungkinan mengandungi maklumat-maklumat penting dan berharga dalam menyelesaikan isu-isu bagi pelbagai domain seperti maklumat perubatan (M. Abulaish et al., 2019), geologi (A. Rauch et al., 2019), pembinaan (M. Marzouk & M. Enaba, 2019), semakan dan cadangan pengguna (M. Siering et al., 2018) serta pengurusan tender (K. Rajbabu et al., 2018). Walau bagaimanapun, peningkatan ketersediaan dokumen teks tidak berstruktur dalam format elektronik ini telah membawa kepada perkembangan penyelidikan dalam bidang Pengekstrakan Maklumat (PM).

Pengestrakan maklumat merupakan proses bagi mendapatkan konsep penting dalam mewakili kandungan teks dari dokumen yang tidak berstruktur. Teknik pengekstrakan yang digunakan dalam pengecaman entiti nama (PEN) juga melibatkan proses lain seperti pengenalan entiti, penggabungan entiti dan pengekstrakan entiti. Pengecaman Entiti Nama (PEN) merupakan salah satu dari cabang penyelidikan Pengesktrakan Maklumat (PM) yang amat penting bagi mendapatkan senarai kata kunci yang relevan bagi sesuatu dokumen. Melalui PEN, pengestrakan maklumat dapat dilakukan dengan mengenalpasti entiti nama (nama orang,

lokasi, organisasi), nilai masa (tarikh, masa, tempoh) dan nilai nombor (wang, peratus, berangka, kardinal) dalam koleksi teks atau dokumen.

Al-Quran mengandungi kira-kira 77,000 perkataan yang disusun dalam bentuk bab-bab kecil sebanyak 114 bab yang dikenali sebagai surah. Setiap surah tersebut pula mengandungi ayat-ayat. Al-Quran mengandungi pelbagai maklumat dan maksud yang luas dan ianya dalam keadaan tidak tersusun dan berselerakan namun berkait antara satu sama lain secara konsep (Eric S Atwell et al., 2009). Entiti di dalam Al-Quran boleh merujuk kepada namaorang, golongan atau kaum, peristiwa dan sebagainya. Mencari entiti secara manual adalah sukar dan memerlukan masa yang lama.

Salah satu penyelesaian kepada masalah ini adalah dengan mengenalpasti dan mengekstrak entiti nama yang terdapat dalam teks Al-Quran terjemahan Bahasa Melayu. Di samping itu, dengan kajian ini juga dapat menambah entiti nama bagi Domain Islam berdasarkan kajian-kajian lepas dan indeks Al-Quran serta menganotasikan entiti nama tersebut.

Penyelesaian lain yang boleh diutarakan dalam kajian ini adalah dengan mencari bagaimana cara melatih dan mengemaskini model bagi mendapatkan tetapan (*setting*) parameter yang terbaik untuk model PEN bagi Domain Islam. Kajian ini akan membangunkan, melatih dan mengemaskini model NER yang disesuaikan untuk mengenalpasti dan mengklasifikasikan entiti nama bagi Domain Islam seperti nama nabi, malaikat, kitab suci, dan istilah utama yang unik kepada ajaran Islam, dalam teks Al-Quran terjemahan Bahasa Melayu. Ini boleh membantu dalam menyediakan maklumat dan meningkatkan pemahaman semasa membaca dan belajar teks Al-Quran terjemahan tersebut. Kajian ini akan membuktikan sejauh mana keberkesanan menggunakan pendekatan pembelajaran mesin dalam mengekstrak entiti nama hasil daripada model PEN yang telah dilatih dan dikemaskini mengikut tetapan parameter yang terbaik bagi Domain Islam dalam teks Al-Quran terjemahan Bahasa Melayu.

Melalui kajian ini, ianya dapat membantu dalam mengenalpasti entiti nama yang pastinya merupakan konsep penting yang dibincangkan dalam teks Al-Quran terjemahan Bahasa Melayu. Melalui PEN ini, ianya akan dapat membantu dalam mengenalpasti dan mengekstrak entiti nama dan seterusnya memudahkan dalam proses capaian maklumat. PEN juga dapat membantu dalam mengenalpasti konsep-konsep utama sesuatu domain serta

membekalkan maklumat penting untuk kefahaman yang jitu tentang sesuatu maksud yang terkandung dalam sesebuah teks dokumen.

Laporan ini mengandungi lima (5) seksyen. Seksyen I menerangkan latarbelakang kajian ini yang merangkumi permasalahan kajian dalam pengecaman entiti nama teks Al-Quran terjemahan Bahasa Melayu. Seksyen II pula membincangkan mengenai pengekstrakan maklumat, pengecaman entiti nama dan pendekatan-pendekatan yang ada. Seksyen III membincangkan metodologi kajian yang digunakan untuk mencapai objektif dan skop kajian. Penerangan mengenai rekabentuk kajian dan teknik-teknik yang digunakan akan dibincangkan pada seksyen ini. bab perbincangan iaitu pengenalan, kajian literasi, metodologi kajian, analisis kajian dan kesimpulan. Setiap bab menerangkan secara terperinci terhadap kajian yang dilaksanakan. Seksyen IV akan membentangkan analisis bagi keputusan kajian. Akhir sekali, seksyen V merumuskan secara keseluruhan kajian, kekangan dan cadangan di masa hadapan.

II. KAJIAN LITERASI

A. Pengekstrakan Maklumat (PM)

Pengestrakan Maklumat merupakan satu bidang yang melakukan proses pengekstrakan maklumat daripada data digital dan mencari kekerapan kelas objek tertentu dan hubungan antara objek. Sebagai contoh mengekstrak alamat dari halaman web, seperti jalan, bandar, daerah, negeri dan poskod. Contoh lain; mengekstrak maklumat kejadian ribut daripada laporan cuaca, data seperti suhu, kelajuan angin dan hujan.

Di antara kepentingan PM yang dikenalpasti adalah membantu enjin pencarian dokumen daripada halaman web. Teknik pengekstrakan diperlukan dalam mencari maklumat yang tepat daripada satu atau lebih dokumen web. Selain itu, pengekstrakan maklumat diperlukan dalam proses pemindahan data daripada sistem asal ke sistem yang baru. Situasi ini sering berlaku apabila pengguna bertukar sistem komputer. Data daripada sistem asal akan diekstrak dan diubah format yang sesuai dengan sistem yang baru.

B. Pengecaman Entiti Nama (PEN)

Pengecaman Entiti Nama (PEN) telah digunakan secara meluas dan memainkan peranan penting khususnya dalam bidang berkaitan dengan PBT. PEN juga merupakan salah satu teknik

pengekstrakan maklumat yang membantu mengekstrak dan mengenalpasti maklumat yang diingini. Maklumat ini kemudiannya disimpan ke dalam pangkalan data bagi memudahkan carian maklumat dilakukan. Melalui PEN, pengesktrakan maklumat dapat dilakukan dengan mengenalpasti entiti nama dan mengkelaskannya mengikut kategori yang ditentukan seperti individu, lokasi, organisasi, tarikh, masa, nilai kewangan dan nilai peratusan.

i) Pengestrakan Ciri

Entiti Nama (EN) mempunyai ciri-ciri tersendiri dalam bahasa, entiti nama adalah merujuk kepada sesuatu entiti atau konsep tertentu dan biasanya tidak disenaraikan di dalam leksikon. Fungsi PEN adalah mencari dan mengklasifikasikan entiti nama di dalam sesuatu teks kepada kategori seperti nama orang, organisasi, lokasi, ungkapan masa, kuantiti dan sebagainya. Ia boleh juga dinyatakan PEN dilihat sebagai proses dua fasa (*two-phases process*); a) mengenalpasti sempadan entiti b)mengklasifikasi entiti ke dalam kategori yang betul. Sebagai contoh, Jika perkataan Muhammad ada di dalam sesebuah ayat, adalah penting untuk mengenalpasti permulaan dan akhiran entiti nama tersebut di dalam ayat itu. Maka entiti tersebut akan diklasifikasikan ke dalam kategori yang betul iaitu kategori nama orang.

Terdapat pelbagai ciri telah dipertimbangkan dalam kajian yang berkaitan dengan PEN, di antaranya adalah ortografik, N-gram, penandaan golongan kata (*POS tag*), perkataan pemberat (*word weight*), tokenisasi dan sebagainya.

ii) Anotasi Entiti Nama

Bahasa mesin tidak dapat memahami konteks frasa, makna setiap kata, ayat atau frasa, situasi atau percakapan tertentu mahupun makna holistik bagi sesuatu pernyataan. Konsep seperti humor dan elemen abstrak lainnya tidak dapat ditafsirkan menjadikan pelabelan data teks menjadi lebih sukar. Atas sebab itu anotasi teks mempunyai beberapa peringkat mengikut kategori anotasi iaitu anotasi semantik, anotasi maksud, kategori teks dan anotasi entity.

iii) Teks Tanpa Label dan Berlabel

Perlabelan data adalah penting untuk pemprosesan bahasa tabii dan pembelajaran mesin di mana ianya membolehkan model memahami dan mentafsir data dengan lebih baik. Dengan menggunakan pelbagai jenis anotasi data dan menggunakan alatan dan platform yang betul, dapat melatih dan menambah baik model pembelajaran mesin dengan lebih berkesan dan mencapai hasil yang lebih baik.

iv) Gazetir

Gazetir adalah senarai entiti nama yang ditakrifkan dan ianya mengandungi senarai nama khusus untuk jenis kelas entiti nama tertentu. Sebagai contoh; Malaysia berada dalam kelas gazetir lokasi manakala Tun Mahathir berada di dalam kelas nama orang. Gazetir dibangunkan mengikut keperluan setiap kelas entiti nama (Nadia & Omar, 2019). Gazetir juga dikenali sebagai senarai putih atau kamus (K. Shaalan & H. Raza, 2008).

C. Pendekatan Pengecaman Entiti Nama (PEN)

Matlamat utama PEN adalah untuk mengklasifikasikan entiti nama seperti individu, lokasi, organisasi da sebagainya (Hadi, 2011). Terdapat tiga pendekatan dalam sistem PEN iaitu pendekatan berasaskan peraturan (*rule-based approach*), pendekatan pendekatan berasaskan pembelajaran mesin (*machine learning approach*) dan pendekatan berasaskan hibrid iaitu melibatkan gabungan pendekatan peraturan, statistikal dan pembelajaran mesin. Namun hanya dua pendekatan yang popular dalam pelaksanaan pengecaman entiti nama iaitu pendekatan berasaskan peraturan dan pendekatan berasaskan pembelajaran.

i) Pendekatan Berasaskan Peraturan

Pendekatan berasaskan peraturan merupakan kaedah yang bergantung sama ada peraturan berasaskan heuristik (*heuristic*) atau ungkapan pemalar (*regular expression*) untuk mengklasifikasikan entiti nama. Ia juga boleh dipengaruhi oleh ontologi luaran (Rindflesch et al., 2000), lingustik (Proux et al., 1998) dan juga konteks (Fukada et al., 1998; Humphreys et al., 2000). Selain itu, pendekatan berasaskan peraturan juga turut digunakan bersama-sama sumber bahasa seperti kamus, senarai gazetir atau senarai penanda di samping memerlukan penglibatan kepakaran manusia untuk mengenalpasti dan

mengekstrak entti nama tersebut, sama ada dari segi tatabahasa, sintaksis atau gabungan dengan senarai kata kunci.

ii) Pendekatan Berasaskan Mesin

Bagi pendekatan pembelajaran mesin adalah teknik pembelajaran corak klasifikasi daripada korpus yang besar. Pendekatan ini memerlukan sejumlah bilangan korpus yang besar untuk dijadikan data pembelajaran. Bagi membolehkan pengecaman entiti nama berfungsi, sistem perlu mempelajari ciri-cirinya dengan mengenalpasti corak daripada korpus yang besar dan mendapatkan hasil keputusan berdasarkan data tersebut.

iii) Pendekatan Hibrid

Pendekatan hibrid menggabungkan lebih daripada satu pendekatan PEN, termasuklah kombinasi pendekatan seperti pendekatan berasaskan peraturan dan pendekatan berasaskan pembelajaran mesin bagi mengoptimumkan prestasi keseluruhan kajian tersebut (C. Kiefer et al., 2019).

Pendekatan hibrid juga mempunyai kelebihan dan keburukannya. Kebanyakan kajian yang menggunakan pendekatan hibrid ini, ianya akan meningkatkan prestasi dari segi ketepatan keputusan (K. Shaalan et al., 2014). Ianya mempunyai beberapa modul berdasarkan domain kajian. Tetapi beberapa kajian yang telah dilakukan banyak menggabungkan pendekatan pembelajaran mesin dan pendekatan peraturan berasaskan corak untuk mengekstrak beberapa entiti nama seperti individu, lokasi dan organisasi

D. Korpus Bahasa Melayu dan Domain Islam

Korpus memainkan peranan penting dalam penyelidikan PBT yang berasaskan dokumen teks dan kebiasaannya, korpus tersebut adalah dokumen teks tidak berstruktur. Penyelidikan PEN semakin berkembang dan kebanyak jumlah korpus yang tersedia di dalam Bahasa Inggeris menyebabkan ramai para penyelidik menjalankan penyelidikan PEN dalam bahasa tersebut. Manakala korpus Bahasa Melayu masih terhad berbanding dengan sumber Bahasa Inggeris (Sazali, 2016).

Kajian PEN boleh dilaksanakan dalam pelbagai domain sebagai contoh domain pendidikan, jenayah, dokumen teks terjemahan Al-Quran atau Hadis, umum dan seumpamanya. Di dalam agama Islam, terdapat dua sumber rujukan utama iaitu Al-Quran dan hadis. Al-Quran ditulis dalam Bahasa Arab dan kemudiannya diterjemahkan kepada bahasa-bahasa lain seperti Bahasa Inggeris, Bahasa Indonesia dan juga Bahasa Melayu untuk kemudahan memahami maksud ayat-ayat yang dibaca. Setiap surah atau bab mengandungi naratif yang berbeza dan kadangkala tiada perkaitan antara surah-surah atau bab-bab tersebut. Di dalam setiap bab, terdapat juga banyak entiti di dalam setiap ayat, yang boleh menyukarkan seseorang untuk mencari entiti tertentu. Entiti dalam Al-Quran boleh merujuk kepada nama orang atau sesuatu kumpulan.

Maka penyelidikan dan pembangunan PEN diperlukan dalam mengenalpasti entiti dan mengklasifikasikannya bagi mendapatkan maklumat dan pemahaman sesuatu bab tersebut dengan lebih bermakna.

E. Kajian Terdahulu

Kajian *Malay Named Entity Recognition Based on Rule-Based Approach* oleh Alfred et al (2014) mencadangkan algoritma pengecaman entiti nama berasaskan peraturan untuk artikel Bahasa Melayu. Pengecaman Entiti Nama dalam Bahasa Melayu yang dicadangkan ini direkabentuk berdasarkan kepada ciri-ciri Penandaan Golongan Kata (*POS tagging*) bagi Bahasa Melayu dan ciri-ciri kontekstual yang telah dihasilkan untuk menangani pengecaman entiti nama bsgi artikel-artikel dalam Bahasa Melayu. Berdasarkan kepada keputusan *POS tagging* tersebut, entiti nama yang betul akan dikenalpasti atau dikesan di mana berkemungkinan sesuai untuk dianotasikan. Di samping itu, terdapat juga beberapa simbol dan kata hubung yang akan dipertimbangkan dalam proses untuk mengenalpasti entiti-entiti nama bagi artikel Bahasa Melayu. Terdapat juga beberapa kamus kata yang dibina secara manual untuk digunakan bagi pengecaman entiti nama seperti individu, lokasi dan organisasi. Hasil keputusan kajian menunjukkan nilai dapatan sebanyak 89.74% bagi ukuran-f. Penulis menyatakan bahawa algoritma pengecaman entiti nama dalam Bahasa Melayu boleh ditambahbaik lagi dengan mewujudkan lebih banyak kamus kata yang lebih lengkap dan memperhalusikan lagi peraturan yang digunakan untuk mengenalpasti sistem entiti nama bagi Bahasa Melayu yang betul.

Merujuk kepada kajian Saad & Mansor (2018), penulis menjalankan kajian dengan membangunkan sistem prototaip model pengestrakan maklumat dokumen berita jenayah dalam Bahasa Melayu dengan menggunakan teknik pengecaman entiti nama melalui pendekatan berasaskan peraturan. Kajian ini dilakukan dengan menggunakan korpus berita jenayah dalam Bahasa Melayu yang diperolehi dari sumber arkib berita BERNAMA. Hasil keputusan yang diperolehi daripada sistem prototaip yang dibangunkan menunjukkan hasil yang baik iaitu dengan nilai dapatan adalah sebanyak 78.6%, manakala bagi nilai kejituan adalah sebanyak 71.11% dan ukuran-f pula sebanyak 74.7%.

Walaupun kajian NER Bahasa Melayu sudah ada dilaksanakan, tetapi terdapat beberapa peraturan yang dihasilkan masih tidak mencukupi dan tidak menyeluruh dalam mengecam entiti nama, terutamanya bagi domain-domain tertentu (Nadia & Omar, 2019). Penulis menyatakan isu yang mencabar dalam PEN Bahasa Melayu adalah rujukan silang antara satu entiti nama dengan entiti nama yang lainnya, pencampuran entiti nama dan pengulangan entiti nama. Penulis mencadangkan peraturan baru bagi mengatasi isu dalam PEN Bahasa Melayu dengan menyediakan korpus fail teks berita Bahasa Melayu atas talian, pembangunan gazetir, pembangunan peraturan dan penilaian. Kajian ini memberi fokus kepada pengecaman entiti nama yang melibatkan sembilan entiti nama iaitu nama individu, lokasi, organisasi, jawatan, tarikh, masa, kewangan, ukuran dan peratusan. Secara keseluruhannya, hasil pengujian menunjukkan nilai dapatan sebanyak sebanyak 90.23%, nilai kejituan sebanyak 92.13% dan ukuran-f sebanyak 91.05%.

Pendekatan berasaskan peraturan juga digunakan dalam kajian Maha et al., (2018), untuk memadankan nama-nama Allah dalam domain tertentu iaitu Al-Quran, sistem cadangan tersebut dibina daripada koleksi peraturan dengan menggunakan ungkapan pemalar (*regular expression*). Peraturan yang dicadangkan akan dipadankan dengan semua nama Allah di dalam Al-Quran. Terdapat banyak ciri telah dipertimbangkan dalam kajian ini seperti ortografik, N-gram dan imbuhan. Sistem yang dicadangkan akan menyelesaikan masalah penampilan nama-nama Allah dalam Al-Quran sebagai sub jujukan dalam perkataan lain serta menganggap diakritik (Al-Tashkeel) dalam perkataan Arab. Keefisyenan persamaan nama diambilkira untuk mengurangkan kerumitan asas peraturan. Keputusan pengujian menunjukkan bahawa peraturan berdasarkan ungkapan biasa adalah teknik yang berkesan, berkuasa dan cekap dalam menyesuaikan nama sasaran dengan koleksi peraturan yang sedikit.

Kajian oleh Asmai, Salleh, Basiron dan Ahmad (2018), menyatakan bahawa model pengecaman entiti nama dalam Bahasa Melayu boleh dipertingkatkan dengan menggunakan gabungan teknik Algoritma *Fuzzy c-means* dan *K-Nearest Neighbours* bagi domain analisis jenayah. Hasil keputusan menunjukkan bahawa gabungan kedua-dua teknik tersebut boleh meningkatkan prestasi ketepatan pada pengecaman entiti data jenayah dalam Bahasa Melayu. Nilai kejituan secara keseluruhannya menunjukkan sebanyak 95.24% semasa pengklasifikasi k-NN. Keputusan nilai kejituan tersebut boleh dijadikan ukuran yang boleh menjadi perspektif keseluruhan proses penilaian dapat dipertingkatkan lagi dengan meningkatkan set data latihan untuk hasil yang lebih baik pada eksperimen akan datang.

Jadual **Error! No text of specified style in document.**.1 Perbandingan kajian lepas dalam PEN - Islamik

Penyelidik	**Domain**	**Bahasa**	**Entiti Yang Diestrak**	**Kaedah Penyelidikan**	**Hasil**
Maha et al. (2018)	Al-Quran	Arab	Nama-nama Allah	Koleksi peraturan dibina dengan menggunakan ekspresi peraturan (*regular expression*) kemudian dipadankan dengan semua nama Allah di dalam Al-Quran. Ciri-ciri lain yang dipertimbangkan: ortografik, N-gram dan imbuhan (*affixes*)	Pendekatan berasaskan peraturan lebih efektif dan efisyen

bersambung…

sambungan…

Azalia et al. (2019)	Terjemahan Hadis	Indonesia	Nama Perawi	Menggunakan teknik pengklasifikasi *Naive Baiyes*. Ciri terlibat*: title case, POS tag* dan unigram. Ukuran-F1 digunakan dalam kajian ini untuk mengukur features dan prestasi entiti nama.	*F1-Score*= 82.63%
Maulana et al. (2019)	Terjemahan Al-Quran	Inggeris	Orang	Menggunakan teknik SVM. Dataset diperolehi dari laman web tanzil.net yang mengandungi 19,473 token dan 720 entiti.	*F-Measure*= 75%

Jadual **Error! No text of specified style in document.**.2 Perbandingan kajian lepas dalam PEN - Umum

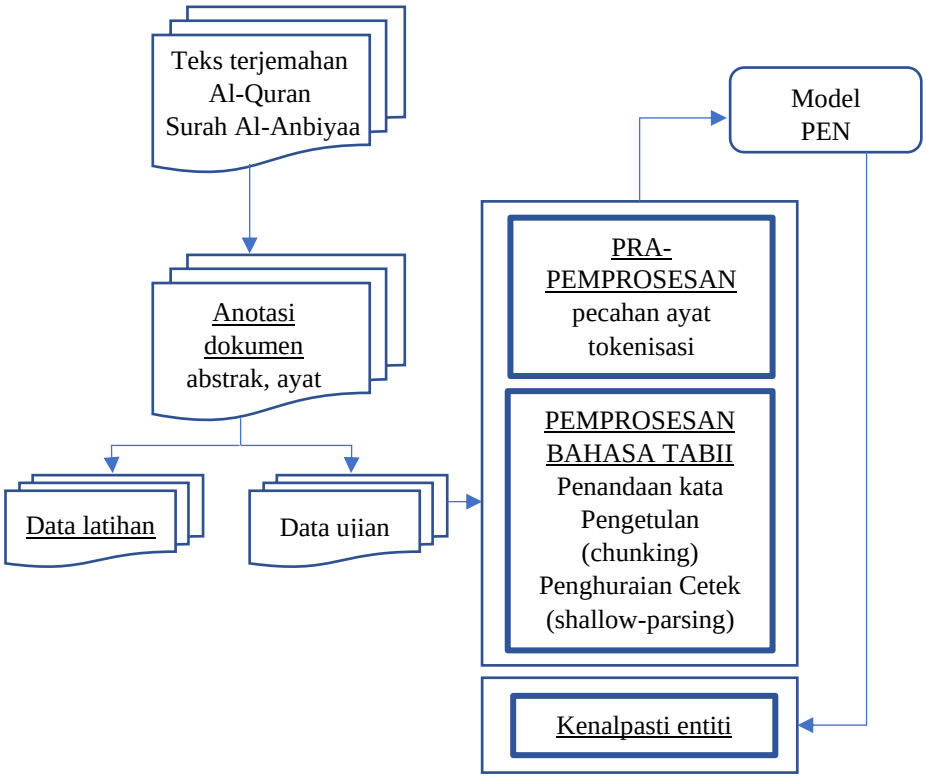
Penyelidik	Domain	Bahasa	Entiti Yang Diestrak	Kaedah Penyelidikan	Hasil
Alfred et al. (2014)	Artikel Bahasa Melayu	Melayu	Orang, Organisasi dan Lokasi	RPOS (berasaskan petua)	*F-Measure* = 89.47%
Sulaiman et al. (2017)	Artikel Bahasa Melayu	Melayu	Orang, Organisasi dan Golongan Kata	Merujuk kepada piawai Stanford dan Illnois	Kejituan Stanford = 36.55% Kejituan Illnois = 35.64%
Halid et al. (2017)	Artikel Berita Harian atas talian	Melayu	30 Petua GK termasuk petua imbuhan dan 16 petua hubungan kata	Golongan Kata(berasaskan peraturan)	Nilai Kejituan=93.06%
Ulanganathan et al. (2017)	Artikel Berita dari BERNAMA dan media sosial (Twitter, blog dan wikis)	Melayu	Orang, Organisasi, Lokasi dan Sarana	Tiga langkah dilaksanakan dalam PEN: 1) Pra-pemprosesan pemecahan perkataan pada data latihan. 2) Anotasi manual bagi entiti individu, lokasi , organisasi dan sarana. 3) Perkataan diproses menggunakan Linear-Chain CRF. 4) Hasil keputusan PEN dibandingkan di antara sistem Mi-NER, PEN berasaskan peraturan dan sistem sedia ada; Semantria.	Mi-NER= 89.87% Berasaskan peraturan =78.95% Semantria= 52.74%
Salleh et al. (2018)	Berita Jenayah daripada Laman PDRM	Melayu	Orang, Organisasi, Lokasi, Tarikh dan Jenis Jenayah	Menggunakan gabungan teknik *Fuzzy c-means* dan *K-Nearest Neighbors.*	Nilai Kejituan= 95.24%
Saad et al. (2018)	Dokumen Berita Jenayah	Melayu	Orang, Organisasi, Lokasi, Tarikh, Masa, Kewangan,	Menggunakan kontekstual, penandaan golongan kata, saringan kata kunci dan saringan jenis morfologi.	*Precision*= 71.11% *Recall*= 78.67% *F-Measure* = 74.7%

bersambung…

...sambungan

			Ukuran, Peratusan, Jenayah dan Senjata		
Nadia et al. (2019)	Artikel Berita	Melayu	Orang, Organisasi, Lokasi, Tarikh, Masa, Kewangan, Ukuran, Peratusan	PEN berasaskan peraturan yang dimulakan: 1) Menyediakan korpus fail teks berita atas talian Bahasa Melayu 2) Pembangunan gazetir dan pembangunan peraturan.	*Precision*= 90.23% *Recall*= 92.13% *F-Measure* = 91.05%
Sugiarta et al. (201)	Dokumen Teks Bali	Bali	Lokasi	Pendekatan PEN berasaskan peraturan	*Precision*= 93% *Recall*= 93% *F-Measure*= 92%

III. METODOLODI KAJIAN



Rajah 3.1 menggambarkan rekabentuk metodologi kajian yang merangkumi fasa-fasa yang terlibat beserta penerangan seperti berikut:

A. Fasa Pertama, Penyediaan Data

Fasa penyediaan data melibatkan pengenalpastian dataset dan pemilihan skop domain bagi kajian ini. Dataset yang dipilih adalah daripada ayat terjemahan Al-Quran dalam Bahasa Melayu yang boleh didapati di https://www.surah.my/21; Surah An-Anbiya; yang mengandungi 112 ayat.

B. Fasa Kedua, Pra-Pemprosesan Data

Bagi menyediakan dataset tersebut, dokumen teks Islamik perlu melalui beberapa pendekatan pra-pemprosesan sebelum proses seterusnya dilaksanakan. Pra-pemprosesan adalah teknik untuk menyiapkan data agar lebih teratur untuk ke proses yang selanjutnya dalam mengekstraksi dokumen teks. Dengan kata lain, proses pembersihan untuk mendapatkan data bersih daripada ‘data mentah’. Ini adalah untuk menghindari kecelaruan ketika menganalisis data tersebut.

i) Pembuangan Tanda Baca

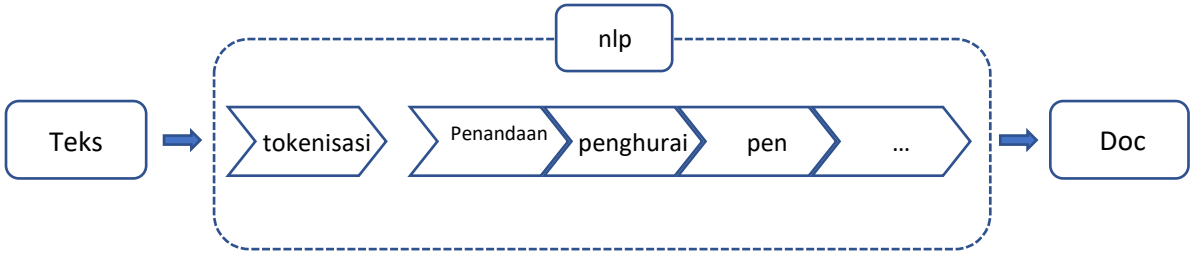
Tanda baca ditaktifkan sebagai satu set tanda yang digunakan untuk menjadikan sesuatu ayat atau perenggan dapat dinyatakan maksudnya dengan tepat dan bermakna. Tanda baca menentukan di mana kita perlu berhenti atau memberikan pemahaman terhadap kata-kata bagi pembaca tersebut. Sebagai contoh tanda baca dalam sesebuah dokumen teks adalah seperti “?”, “!”, “,”, “*” dan banyak lagi tanda baca.

ii) Pentokenisasi

Salah satu tugas paling asas dalam Pemprosesan Bahasa Tabii (PBT) ialah tokenisasi atau dengan kata lain memsegmentasikan atau memecahkan teks-teks dalam sesuatu ayat atau perenggan atau dokumen menjadi token individu (perkataan dan tanda baca).

Tokenisasi adalah salah satu bahagian utama dalam kebanyakan *pipeline* Pemprosesan Bahasa Tabii. Ini adalah kerana ia membolehkan kita menjadikan sesuatu perkataan dan tanda baca tertentu sebagai unit diskret.

Proses tokenisasi ini dimudahkan dengan menggunakan perpustakaan dari SpaCy dengan melakukan tokenisasi dengan objek PBT dan kelas *Doc*. Untuk menandakan segmentasi teks dengan SpaCy, kita boleh menghantar rentetan teks tersebut ke objek *Doc* yang kemudiannya akan diproses ke peringkat seterusnya di dalam *pipeline.*



Rajah 3.1 Pemprosesan *pipeline*

C. Fasa Ketiga, Penganotasian Entiti Nama

Anotasi entiti nama ini dilakukan pada dataset mentah dalam bentuk fails teks. Kaedah anotasi ini dilakukan dengan bantuan alatan (*tools*) anotasi entiti nama yang boleh didapati di laman https://tecoholic.github.io/ner-annotator/. Kategori yang dijalankan proses penandaan anotasi ialah Kitab Suci, Nabi, Nama Tuhan, Golongan, Peristiwa, Malaikat, Agama, Tempat, Syurga, Neraka dan Jin. Setelah selesai mengkategorikan entiti-entiti tersebut, kategori entiti tersebut akan disimpan dalam format fail *.json.

D. Fasa Keempat, Pembangunan Model Latihan PEN

SpaCy ialah perpustakaan pemprosesan bahasa semula jadi yang ditulis dalam Python dan Cython, dan ia serasi dengan CPython 64-bit 2.7 / 3.5+ dan boleh beroperasi pada Unix/Linux, macOS/OS X dan Windows. SpaCy juga merupakan perpustakaan yang popular yang menggunakan kedua-dua pembelajaran mesin (machine learning) dan teknik pembelajaran mendalam (deep learning) untuk melaksanakan pelbagai tugas PBT. Ia direka bentuk untuk menjadi cekap, pantas dan mesra pengguna. Fungsi teras Spacy termasuk tokenisasi, pengetegan sebahagian daripada pertuturan, pengecaman entiti bernama, penghuraian kebergantungan dan banyak lagi. Beberapa komponen SpaCy, seperti pengecaman entiti nama (PEN) dan klasifikasi teks, memanfaatkan penggunaan teknik pembelajaran mendalam. Model ini dilatih pada data berlabel menggunakan rangkaian saraf (neural network) yang membolehkan ianya mempelajari corak kompleks dan meningkatkan ketepatan bagi pelbagai tugas PBT.

i) Model SpaCy

Sistem PEN Spacy mengandungi strategi pembenaman perkataan (word embedding) menggunakan ciri-ciri sub perkataan dan pembenaman "*Bloom*" dan *deep convolution neural network* bersama *residual* CNNs. Sistem ini direkabentuk untuk memberikan keseimbangan kecekapan, ketepatan dan kebolehsuaian yang baik. Di antara senibina model yang digunakan untuk melatih model pengecaman entiti nama ini adalah seperti berikut model tokenisasi (*tok2vec*), model pengecaman entity nama (*ner*), pembenaman perkataan (*word embedding*) dan pengoptimum (*optimizer*).

ii) Komponen *Pipeline*

SpaCy mempunyai pelbagai komponen pipeline yang telah sedia ada . Di antaranya adalah penanda golongan kata (*POS tagging*), kebergantungan penghurai (*dependency parser*), pengecaman entiti nama (*ner*) dan sebagainya. Bagi kajian ini, komponen *pipeline* yang digunakan adalah komponen *pipeline* pengecaman entiti nama. Komponen ini akan menambah entiti yang dikenalpati menggunakan arahan *doc.ents*. Ia juga menetapkan jenis atribut entiti pada token yang menunjukkan samada token tersebut adalah sebahagian daripada entiti atau pun tidak.

iii) Fail Konfigurasi

SpaCy menggunakan fail konfigurasi, biasanya dipanggil *config.cfg*, sebagai sumber untuk semua tetapan. Fail konfigurasi akan mentakrifkan cara untuk memulakan objek PBT, komponen *pipeline* yang ingin digunakan dan cara bagaimana pelaksanaan model harus dikonfigurasikan. Ia juga termasuk semua tetapan untuk proses latihan dan cara bagaimana untuk memuatkan data, termasuk hiperparameter. Sebagai contoh, [*components.ner.model*] berfungsi untuk mentakrifkan tetapan untuk pelaksanaan model pengecaman entiti nama. Selain itu, beberapa hiperparameter telah dikenalpasti bagi mengoptimumkan proses dan kecekapan bagi mendapatkan nilai ketepatan yang lebih tinggi. Pengoptimuman kecekapan merujuk kepada inferens lebih cepat, model yang lebih kecil, penggunaan memori yang lebih rendah. Antara hiperparameter latihan yang digunakan ialah ‘*learning rate*’, ‘*max_epoch*s’ dan ’*optimizer*’.

E. Fasa Kelima, Pengujian

Di dalam Pengecaman Entiti Nama, ketepatan adalah paling utama. Namun ketepatan sahaja bukanlah aras ukuran kepada sesuatu kualiti bagi sesebuah model. Kita perlu menilai model daripada pelbagai aspek. Dalam kajian ini, metriks penilaian yang akan digunakan adalah '*precision*' (P), '*recall*' (R), '*accuracy*' (A) dan '*F1-scores*'. Keempat-empat metriks penilaian ini berkait rapat dengan empat nilai dalam '*confusion matrix*' seperti Jadual 3.2 bawah; TP (*True Positive*), TN (*True Negative*), FP (*False Positive*) dan FN (*False Negative*).

Jadual **Error! No text of specified style in document.**.3 Jadual *confusion* matrix

Nilai Sebenar	Nilai Jangkaan	
	Positif	Negatif
Positif	*True Positive* (TP)	*True Negative* (TN)
Negatif	*False Positive* (FP)	*False Negative* (FN)

Keempat-empat nilai ini akan mempengaruhi hasil pengiraan keempat-empat metriks penilaian iaitu nilai kejituan, nilai dapatan semula, nilai pengukuran-F1 dan nilai ketepatan.

i) Kejituan (*Precision*)

$$Precision = \frac{TP}{TP + FP}$$

ii) Dapatan Semula (*Recall*)

$$Recall = \frac{TP}{TP + FN}$$

iii) Pengukuran-F1 (*F1-Score*)

$$Recall = \frac{TP}{TP + FN}$$

iv) Ketepatan (*Accuracy*)

$$Recall = \frac{TP}{TP + FN}$$

IV. ANALISIS KEPUTUSAN MELATIH MODEL PEN

Bagi membangunkan model latihan, pengkaji memilih SpaCy sebagai perpustakaan untuk melatih model bagi pengecaman entiti nama bagi dokumen teks Al-Quran terjemahan Bahasa Melayu. Model dalam pembelajaran mesin adalah merupakan '*output*' algoritma pembelajaran mesin yang dilaksanakan pada data.

Seterusnya, proses mengkonfigurasi fail model latihan dengan mengubahsuai parameter iaitu jumlah *epochs*, *batch size* dan *learning rate*. Contoh fail konfigurasi bagi model latihan tersebut yang diberikan nama *config.cfg* dapat dilihat pada Rajah 4.1. Model ini akan dilatih dengan nilai parameter yang berbeza dan akan dicatat di akhir proses pelatihan model tersebut. Parameter yang menghasilkan nilai *loss function* terendah dan kadar ketepatan yang tinggi semasa model latihan dilatih akan digunakan dalam pengujian model.

```
path = ${paths.dev}
# Whether to train on sequences with 'gold standard' sentence boundaries
# and tokens. If you set this to true, take care to ensure your run-time
# data is passed in sentence-by-sentence via some prior preprocessing.
gold_preproc = false
# Limitations on training document length
max_length = 0
# Limitation on number of training examples
limit = 0
# Optional callback for data augmentation
augmenter = null

[training]
dev_corpus = "corpora.dev"
train_corpus = "corpora.train"
dropout = 0.1
accumulate_gradient = 1
# Controls early-stopping, i.e., the number of steps to continue without
# improvement before stopping. 0 disables early stopping.
patience = 1600
max_epochs = 300
# Maximum number of update steps to train for. 0 means an unlimited number of steps.
max_steps = 20000
eval_frequency = 200
# Control how scores are printed and checkpoints are evaluated.
score_weights = {}
# Names of pipeline components that shouldn't be updated during training
frozen_components = []
# Names of pipeline components that should set annotations during training
annotating_components = []
# Optional callback before nlp object is saved to disk after training
before_to_disk = null
# Optional callback that is invoked at the start of each training step
before_update = null
```

Rajah 4.2 Fail konfigurasi *config.cfg*

Hasil melatih model PEN diperolehi melalui dengan mengubah jumlah *epochs* dan *learning rate*. Berikut merupakan hasil latihan model PEN:

a. Epochs = 30 dan learning rate = 0.01

```
============================= Training pipeline =============================
ℹ Pipeline: ['tok2vec', 'ner']
ℹ Initial learn rate: 0.01
E    #       LOSS TOK2VEC  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
---  ------  ------------  --------  ------  ------  ------  ------
  0       0          0.00     14.29    0.00    0.00    0.00    0.00
  1     200      23061.46    773.78    0.00    0.00    0.00    0.00
  3     400     112112.54    854.69    6.35    7.23    5.66    0.06
  5     600     187297.70    732.66   56.60   84.91   42.45    0.57
  7     800     110994.32    403.01   28.57   27.93   29.25    0.29
  8    1000     820831.52    592.00   70.59   67.83   73.58    0.71
 10    1200     228502.44    360.76   52.44   49.58   55.66    0.52
 12    1400     114844.06    389.54   47.31   55.00   41.51    0.47
 14    1600     524875.26    657.65   51.46   53.00   50.00    0.51
 16    1800    1154465.11    550.86   66.36   64.04   68.87    0.66
 17    2000     462402.46    479.48   42.53   40.87   44.34    0.43
 19    2200     111856.91    324.14   75.80   73.45   78.30    0.76
 21    2400     362065.61    287.37   79.40   84.95   74.53    0.79
 23    2600     146946.81    235.67   64.22   62.50   66.04    0.64
 25    2800    1267793.89    519.39   68.34   73.12   64.15    0.68
 26    3000     182343.23    272.06   73.08   74.51   71.70    0.73
 28    3200      93785.25    181.15   58.60   57.80   59.43    0.59
✓ Saved pipeline to output directory
model-last
```

Rajah 4.3 *Epochs* 30 dan *learning rate* 0.01

Pada permulaan *epochs*, dapat dilihat pada Rajah 4.2, nilai bagi kolum *LOSS TOK2VEC* dan *LOSS NER* kedua-duanya adalah 0 yang menunjukkan model tersebut belum memulakan latihan. Setelah itu, kemudiannya nilai untuk kedua kolum tersebut mula menunjukkan nilai kekurangan dan ini menunjukkan model sedang dilatih dan meningkatkan prestasi secara perlahan.

Untuk permulaan, kolum metrik penilaian (*ENTS_F*, *ENTS_P*, *ENTS_R*) untuk proses pengecaman entiti nama menunjukkan prestasi model agak rendah pada permulaan *epochs* (0-7) tetapi ianya bertambah baik secara beransur-ansur apabila latihan *epochs* yang seterusnya (8-28). Walaupun *ENTS_P* dan *ENTS_R* menunjukkan nilai turun dan naik, tetapi secara amnya ianya bertambah baik dari semasa ke semasa di mana model menjadi tepat dan dapat mencari lebih banyak entiti nama.

Bagi latihan model ini, *epochs* 21 yang mempunyai nilai ENTS_F (ukuran-F1) tertinggi iaitu 79%, menunjukkan bahawa model tersebut menunjukkan prestasi yang agak baik dalam mengenalpasti entiti nama serta mengekalkan keseimbangan antara nlai kejituan dan nilai dapatan semula.

b. Epochs = 30 dan learning rate = 0.15

Nilai pada kolum *LOSS TOK2VEC* dan *LOSS NER* secara amnya akan berkurangan apabila model belajar daripada data. Berdasarkan Rajah 4.3 tersebut, didapati pada *epochs* 23 nilai terbaik *ENTS_F* adalah 73%. Pada *epochs* 23 menunjukkan keseimbangan yang baik antara nilai kejituan dan nilai dapatan semula untuk pengecaman entiti nama (Rajah 4.3).

Rajah 4.4 *Epochs* 30 dan *learning rate* 0.015

```
============================= Training pipeline =============================
ℹ Pipeline: ['tok2vec', 'ner']
ℹ Initial learn rate: 0.015
E    #       LOSS TOK2VEC  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
---  ------  ------------  --------  ------  ------  ------  ------
  0       0          0.00     14.29    0.00    0.00    0.00    0.00
  1     200     609331.26   1765.53   36.87   36.04   37.74    0.37
  3     400    2208651.34   1401.05   15.30   18.18   13.21    0.15
  5     600     965718.31    755.04   26.73   26.13   27.36    0.27
  7     800     339278.66    564.86   20.00   13.17   41.51    0.20
  8    1000   11409020.60   1045.73   38.15   49.25   31.13    0.38
 10    1200    1259160.26    483.44   58.37   59.22   57.55    0.58
 12    1400    1991999.68    603.04   68.49   66.37   70.75    0.68
 14    1600     378590.78    375.29   48.15   47.27   49.06    0.48
 16    1800    5557320.25    820.57   19.05   60.00   11.32    0.19
 17    2000    5211792.48    664.27   50.00   48.25   51.89    0.50
 19    2200     386921.67    340.34   42.15   40.17   44.34    0.42
 21    2400    1784349.01    362.97   55.72   58.95   52.83    0.56
 23    2600   13195190.64    516.35   72.64   76.84   68.87    0.73
 25    2800    1560848.73    406.11   53.13   59.30   48.11    0.53
 26    3000    1994891.85    542.62   53.42   78.18   40.57    0.53
 28    3200    5425061.69    497.25   40.72   39.13   42.45    0.41
✓ Saved pipeline to output directory
model-last
```

c. Epochs = 30 dan learning rate = 0.02

Setelah itu, kali ini diubah pula kadar pembelajaran (*learning rate*) kepada 0.02. Berdasarkan Rajah 4.4 tersebut, didapati pada *epochs* 23, nilai terbaik *ENTS_F* adalah 56.25%. Walaupun *ENTS_P* dan *ENTS_R* menunjukkan nilai yang turun naik sebelum itu, tetapi secara amnya ia bertambah baik dari semasa ke semasa di mana model menjadi tepat dan dapat mencari lebih banyak entiti.

```
============================= Training pipeline =============================
ℹ Pipeline: ['tok2vec', 'ner']
ℹ Initial learn rate: 0.02
E    #       LOSS TOK2VEC  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
---  ------  ------------  --------  ------  ------  ------  ------
  0       0          0.00     14.29    0.00    0.00    0.00    0.00
  1     200     419617.59   1049.15    8.33    9.30    7.55    0.08
  3     400    1913434.99   1423.15   33.52   41.10   28.30    0.34
  5     600    1965767.59    866.68   29.39   25.90   33.96    0.29
  7     800    5268675.82   1028.61    0.00    0.00    0.00    0.00
  8    1000    3252863.11    794.05   17.19   16.52   17.92    0.17
 10    1200    4254218.91    860.62   39.47   65.22   28.30    0.39
 12    1400   11124439.18    838.54   33.93   32.20   35.85    0.34
 14    1600   11909464.76   1032.19   46.91   67.86   35.85    0.47
 16    1800    5073778.04    765.60   36.36   39.13   33.96    0.36
 17    2000   14167184.39    885.37   13.08   12.96   13.21    0.13
 19    2200    4791494.49    729.67   28.71   30.21   27.36    0.29
 21    2400   21232567.42   1071.26   36.36   33.60   39.62    0.36
 23    2600   13756040.25    569.05   56.25   83.33   42.45    0.56
 25    2800    3232215.32    517.00   44.81   40.00   50.94    0.45
 26    3000    8502021.09    635.98   26.67   26.92   26.42    0.27
 28    3200   12822822.52    750.94   16.50   17.00   16.04    0.17
✓ Saved pipeline to output directory
model-last
```

Rajah 4.5 *Epochs* 30 dan *learning rate* 0.02

d. Epochs = 30 dan learning rate = 0.001

Pada kadar pembelajaran (*learning rate*) iaitu 0.001 pula menunjukkan nilai terbaik *ENTS_F* adalah sebanyak 100% pada *epochs* 17 dan *epochs* 19 (Rajah 4.5).

```
============================= Training pipeline =============================
ℹ Pipeline: ['tok2vec', 'ner']
ℹ Initial learn rate: 0.001
E    #       LOSS TOK2VEC  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
---  ------  ------------  --------  ------  ------  ------  ------
  0       0          0.00     14.29    0.00    0.00    0.00    0.00
  1     200       1311.31    594.59   65.70   67.33   64.15    0.66
  3     400         31.47    118.42   72.25   81.18   65.09    0.72
  5     600        127.28    101.29   80.75   80.37   81.13    0.81
  7     800         33.89     84.55   87.08   88.35   85.85    0.87
  8    1000         36.08     49.96   91.51   91.51   91.51    0.92
 10    1200        892.94     79.32   91.35   93.14   89.62    0.91
 12    1400         65.24     69.25   93.40   93.40   93.40    0.93
 14    1600         49.86     29.36   98.56  100.00   97.17    0.99
 16    1800         96.73     25.00   98.56  100.00   97.17    0.99
 17    2000         60.92     16.33  100.00  100.00  100.00    1.00
 19    2200          9.50      7.05  100.00  100.00  100.00    1.00
 21    2400         71.45     24.31   98.10   99.04   97.17    0.98
 23    2600       5536.58     50.07   99.53   99.07  100.00    1.00
 25    2800         16.56      6.88   99.53   99.07  100.00    1.00
 26    3000        181.04     62.76   97.61   99.03   96.23    0.98
 28    3200        117.23     29.56   99.06   99.06   99.06    0.99
✓ Saved pipeline to output directory
model-last
```

Rajah 4.6 *Epochs* 30 dan *learning rate* 0.001

Berdasarkan pada Jadual 4.1, capaian terbaik yang dapat diperolehi adalah dengan merendahkan kadar pembelajaran pada kadar 0.001 bagi mencapai pengukuran-F1 yang terbaik iaitu 100% pada *epochs* 17 dan 19. Ianya menunjukkan bahawa model tersebut menunjukkan prestasi yang sangat baik dalam mengenalpasti entiti nama serta mengekalkan keseimbangan antara nilai kejituan dan nilai dapatan semula dengan kadar *loss function* yang sangat rendah berbanding dengan kadar pembelajaran yang lebih tinggi. Apabila dilakukan pengujian menggunakan model tersebut, dapat dilihat perkataan 'Allah', 'Yang Maha Pemurah' dan 'Maha Mengasihani' dilabelkan pada entiti nama NAMA TUHAN yang betul (Rajah 4.7).

Jadual 4.4 Penilaian berdasarkan *learning rate* dan *epochs* bagi Model PEN

E	**#**	***Learning Rate* = 0.01**					***Learning Rate* = 0.015**				
		F1	**P**	**R**	**Score**	**Loss**	**F1**	**P**	**R**	**Score**	**Loss**
1	200	0	0	0	0	773.78	36.87	36.04	37.74	0.37	1765.53
3	400	6.35	7.23	5.66	0.06	854.69	15.30	18.18	13.21	0.15	1401.05
5	600	56.60	84.91	42.45	0.57	732.66	26.73	26.13	27.36	0.27	755.04
7	800	28.57	27.93	29.25	0.29	403.01	20.00	13.17	41.51	0.20	564.86
8	1000	70.59	67.83	73.58	0.71	592.00	38.15	49.25	31.13	0.38	1045.73
10	1200	52.44	49.58	55.66	0.52	360.76	58.37	59.22	57.55	0.58	483.44
12	1400	47.31	55.00	41.51	0.47	389.54	68.49	66.37	70.75	0.68	603.04
14	1600	51.46	53.00	50.00	0.51	657.65	48.15	47.27	49.06	0.48	375.29
16	1800	66.36	64.04	68.87	0.66	550.86	19.05	60.00	11.32	0.19	820.57
17	2000	42.53	40.87	44.34	0.43	479.48	50.00	48.25	51.89	0.50	664.27
19	2200	75.80	73.45	78.30	0.76	324.14	42.15	0.17	44.34	0.42	340.34
21	2400	79.40	84.95	74.53	0.79	287.37	55.72	58.95	52.83	0.56	362.97
23	2600	64.22	62.50	66.04	0.64	235.67	72.64	76.84	68.87	0.73	516.35
25	2800	68.34	73.12	64.15	0.68	519.39	53.13	59.30	48.11	0.53	406.11
26	3000	73.08	74.51	71.70	0.73	272.06	53.42	78.18	40.57	0.53	542.62
28	3200	58.60	57.80	59.43	0.59	181.15	40.72	39.13	42.45	0.41	497.25

Bersambung…

…sambungan

E	#	*Learning Rate* = 0.02					*Learning Rate* = 0.001				
		F1	P	R	Score	Loss	F1	P	R	Score	Loss
1	200	8.33	9.30	7.55	0.08	1049.15	65.70	67.33	64.15	0.66	594.59
3	400	33.52	41.10	28.30	0.34	1423.15	72.25	81.18	65.09	0.72	118.41
5	600	29.39	25.90	33.96	0.29	866.68	80.75	80.37	81.13	0.81	101.29
7	800	0	0	0	0	1028.61	87.08	88.35	85.85	0.87	84.55
8	1000	17.19	16.52	17.92	0.17	794.05	91.51	91.51	91.51	0.92	49.96
10	1200	39.47	65.22	28.30	0.39	860.62	91.35	93.14	89.62	0.91	79.32
12	1400	33.93	32.90	35.85	0.34	838.54	93.40	93.40	93.40	0.93	68.25
14	1600	46.91	67.86	35.85	0.47	1032.19	98.56	100	97.17	0.99	29.36
16	1800	36.36	39.13	33.96	0.36	765.60	98.56	100	97.17	0.99	25.00
17	2000	13.08	12.96	13.21	0.13	885.37	100	100	100	1.00	16.33
19	2200	28.71	30.21	27.36	0.29	729.67	100	100	100	1.00	7.05
21	2400	36.36	33.60	39.62	0.36	1071.26	98.10	99.04	97.17	0.98	24.31
23	2600	56.25	83.33	42.45	0.56	569.05	99.53	99.07	100	1.00	50.07
25	2800	44.81	40.00	50.94	0.45	517.00	99.53	99.07	100	1.00	6.88
26	3000	26.67	26.92	26.42	0.27	635.98	97.61	99.03	96.23	0.98	62.76
28	3200	16.50	17.00	16.04	0.17	750.94	99.06	99.06	99.06	0.99	29.56

Rajah 4.7 Hasil Pengujian Model PEN pada *learning rate* 0.001

Manakala pada Rajah 4.8 pula menunjukkan hasil pengujian ke atas model PEN menggunakan *learning rate* 0.01. Pada rajah tersebut menunjukkan perkataan ‘Maha Mengasihani’ sahaja dilabelkan dengan entiti nama NAMA TUHAN yang betul berbanding dua lagi perkataan iaitu ‘Allah’ dan ‘Yang Maha Pemurah’.

```
#learning rate = 0.01
filetest = open('/content/drive/MyDrive/FYP NER/alkahfi.txt', 'r')
text_test = filetest.read()
filetest.close() # Ensure that you always close the connection to avoid memory leaks
#print(text)

nlp = spacy.load("/content/model-best")
doc = nlp(text_test)

colors = {'NAMA TUHAN': 'lightgreen', 'KITAB SUCI': 'blue', 'NABI': 'red', 'GOLONGAN': 'orange'
          , 'PERISTIWA': 'yellow', 'TEMPAT': 'cyan', 'NERAKA': 'purple', 'SYURGA': 'brown', 'AGAMA': 'pink'
          , 'JIN': 'lime', 'MALAIKAT': 'lightblue'}

options = {'ents': ['NAMA TUHAN', 'KITAB SUCI', 'NABI', 'GOLONGAN', 'PERISTIWA', 'TEMPAT', 'NERAKA', 'SYURGA', 'AGAMA', 'JIN', 'MALAIKAT'], 'colors':colors}

#Visualizing Named Entities
spacy.displacy.render(doc, style="ent", jupyter=True, options=options) # display in Jupyter
```

Dengan nama Allah, Yang Maha Pemurah, lagi Maha Mengasihani NAMA TUHAN .

Rajah 4.8 Hasil Pengujian Model PEN pada *learning rate* 0.01

Pemilihan *learning rate* boleh memberi kesan ketara kepada proses latihan dan prestasi model yang dilatih dengan perpustakaan seperti SpaCy. Apabila menggunakan kadar pembelajaran yang rendah, model boleh memperhalusi parameternya dengan lebih beransur-ansur, membolehkannya menyesuaikan kepada minima setempat yang lebih baik bagi fungsi kehilangan. Kadar pembelajaran yang lebih rendah cenderung membawa kepada penumpuan yang lebih stabil semasa latihan dan juga pengemaskinian yang lebih kecil kepada parameter model. Ini bermakna bahawa parameter model dilaraskan dalam langkah yang lebih kecil, yang boleh menghalang *overshooting* dan *oscillations*, terutamanya dalam kes di mana landskap kerugian adalah kompleks. Dan seterusnya boleh membawa kepada generalisasi yang lebih baik dan ukuran-F1 yang lebih tinggi.

V. KESIMPULAN

Tujuan utama kajian ini adalah untuk mengenalpasti dan mengesktrakan serta menambah entiti nama bagi Domain Islam berdasarkan kajian lepas dan indeks Al-Quran dan menganotasikannya. Selain itu, kajian ini juga turut mencari bagaimana untuk mendapatkan nilai parameter optimum untuk Medan Rawak Bersyarat (CRF) menggunakan Perpustakaan SpaCy bagi bagi Domain Islam.

Kajian ini telah berjaya mencapai objektif yang ditetapkan iaitu dapat menambah entiti nama iaitu NAMA TUHAN, NABI, KITAB SUCI, AGAMA, MALAIKAT, SYURGA, NERAKA, JIN, PERISTIWA, TEMPAT dan GOLONGAN dan mengekstrak entiti nama tersebut melalui proses pengecaman entiti nama menggunakan perpustakaan SpaCy.

Melalui eksperimen yang dijalankan, pengecaman entiti nama dapat diekstrak pada nilai optimum iaitu pada *epochs* 17 dan 19 dan *learning rate* 0.001 dengan penggunaan *optimizer* Adam v1.

A. Kekangan Kajian

Terdapat beberapa kekangan yang dihadapi sepanjang kajian ini dijalankan. Antaranya ialah:

i) Proses dilakukan adalah hanya menggunakan pendekatan pembelajaran mesin yang terdapat pada perpustakaan SpaCy di mana pendekatan hibrid menggunakan pendekatan berasaskan peraturan mungkin dapat meningkatkan keupayaan model dalam mengekstrak entiti dengan lebih tepat. Pra-pemprosesan dengan melakukan penggantian ganti nama Allah boleh dilakukan dan peraturan yang merujuk kepada penerangan lanjut dalam mengkayakan maklumat data boleh dilakukan dengan merujuk kepada tanda kurungan pada ayat seperti ( ) dan [ ].

ii) Menambah data baharu pada set data latihan memerlukan kita untuk melatih semula keseluruhan model dan ia mungkin memakan masa yang agak lama disebabkan kerumitan fasa latihan algoritma yang tinggi.

B. Cadangan Kajian

Beberapa cadangan untuk penambahbaikan di masa hadapan terhadap kajian ini telah dikenalpasti. Antara cadangan – cadangan tersebut adalah:

i) Mendalami dan memperluaskan skop berkaitan domain Islam serta mendapatkan kerjasama dengan pakar yang lebih berkemahiran dalam pengajian Islam khususnya mendalami istilah-istilah dalam Al-Quran bagi mendapatkan hasil yang jitu.

ii) Mungkin perlu mempertingkatkan teknik atau pendekatan dengan menggabungkan dengan menggunakan Transformer seperti BERT sebagai contoh atau menggunakan pendekatan hibrid seperti menggabungkan algoritma HMM dan CRF.

3 Menguji model PEN dengan domain lain dalam Bahasa Melayu untuk melihat keberkesanan model PEN tersebut dapat mengenalpasti entiti nama menggunakan pendekatan CRF.

ACKNOWLEDGEMENT

Penulis ingin mengucapkan setinggi terima kasih kepada Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia kerana memberi peluang dalam menjalankan kajian ini dan sebagai memenuhi keperluan untuk memiliki Sarjana Sains Komputer (Kepintaran Buatan).

RUJUKAN

Abu Abdirohman Rosikhun Nur Muttaqin, Moch Arif Bijaksana & Donni Richasdy, (2020), Alat bantu anotasi pengenalan entitas manusia Al Quran fasilitas saran secara otomatis, e-Proceeding of Engienering, Vol.7, No. 2, Ogos 2020: 7647-7660.

Alsaaran, N., & Alrabiah, M. (2021). Classical Arabic named entity recognition using variant deep neural network architectures and BERT. IEEE Access, 9, 91537-91547.

Asmaa M. Aubid & Alok Mishra, (2020), A rule-based approach to embedding techniques for text document classfication, Applied Sciences, 2020, 10, 4009, Jun 2020: 1-22.

David Reinsel, John Gantz & John Rydning, (2018), The digitization of the world from edge to core, An IDC White Paper.

Kalsom A. Latiff et al., (2018), Pengekstrakan konsep dan hubungan bagi istilah Islam menggunakan pendekatan lexico sintaktik, FTSM, , PS-FTSM-2018-011: 1-15.

Kiran Adnan & Rehan Akbar, (2019), An analytical study of information extraction from unstructured abd multidemiensional big data, Journal of Big Data.

Lukman Fakih Lidimilah, (2017), Question answering terjemah Al Quran menggunakan named entity recognition, Jurnal Ilmiah Informatika, Vol.2, Disember 2017: 139-145.

Maha M. Hassan, Dhamyaa A. Al-Nasrawi, Redha J. Hassan & Noor T. Mahdi, (2018), Rule-based method of name entity recognition for matching Allah's finest names in Holy Quran, Journal of Engineering and Applied Sciences 13 (10) 2018: 3618-3623.

Maulana, M. A., Bijaksana, M. A., & Huda, A. F. (2019). Entity Recognition for Quran English Version with Supervised Learning Approach. Indonesia Journal on Computing (Indo-JC), 4(3), 77-86.

Mohanad Jasim Jaber & Saidah Saad, (2016), NER in english translation of hadith documents using classifiers combination, Journal of Theoretical and Applied Information Technology, Vo.84, No. 3, Februari 2016: 348-354.

Nursyuhada Razali & Saidah Saad, (2021), Pembangunan e-kamus Islam menggunakan ungkapan nalar berdasarkan terjemahan Al-Quran dalam makna teks Bahasa Inggeris, PTA-FTSM-2021-065.

Nur Ashikin Halid & Nazlia Omar, (2017), Malay part of speech tagging using rule-based approach, Asia-Pasific Journal of Information Technology and Multimedia Vol.6 No. 2, Disember 2017: 91-107.

Putri Maya Syakilla Hairol Akhma & Sabrina Tiun, (2022), Pengecaman entiti nama pada teks media sosial twitter Bahasa Melayu menggunakan kaedah BiLSTM-CRF, FTSM, UKM, PTA-FTSM-2022-089.

Ramzi Salahh & Lailatul Qadri Zakaria, (2017), Arabic rule-based named entity recognition systems progress and challenges, International Journal on Advanced Science Engineering and Information Technology, Jun 2017: 815-821.

Ridong Jiang, Rafael E. Banchs & Haizhou Li, (2016), Evaluating and combining named entity recognition systems, Proceedings of the Sixth Named Entity Workshop, joint with 54th ACL, Ogos 2016: 21-27.

S.Sulaiman, R.A. Wahid & F. Morsidi, (2018), Feature extraction using regular expression in detecting proper noun for malay news articles based on KNN algorithm, Journal of Fundamental and Applied Sciences 2016, 8(X), 1-4, Oktober 2017: 1-23.

Safwan Sufian Chang, Juhaida Abu Bakar & Norliza Katuk, (2021), Malay roman corpus annotation system, Multidisciplinary Applied Research and Innovation, Vol.2, No. 3, Disember 2021: 1-4.

Saidah Saad & Mohamed Kamil Mansor, (2018), Named entity recognition approach for malay crime news retrieval, GEMA Online Journal of Language Studies, Vo.18(4), November 2018: 216-235.

Salah, R. E., & Zakaria, L. Q. B. (2018, March). Building the classical Arabic named entity recognition corpus corpus (CANERCorpus), In 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP) (pp. 1-8). IEEE.

Shasha T et al., (2019), Annotation of named entity recognition for quranic text using rule-based method, Research Center for Arabic Language and Islamic Civilization, UKM, 2019: 40-43.

Siti Syakrah Sazali, Nurazzah Abdul Rahman & Zainab Abu Bakar, (2020), Characteristics of malay translated hadith corpus, Journal of King Saud University – Computer and Information Sciences.

Somnath Banerjee, Sudip Kumar, Paolo Rosso & Sivaji Bandyopadhyay, (2018), Named entity recognition on code-mixed cross-script social media content, Computacion y Sistemas, Vol.21, No. 4, Januari 2018: 681-692.

Sugiarta A & Sanjaya A, (2021), Location named entity recognition using rule-based approach for balinese texts, Jurnal Elektronik Ilmu Komputer Udayana, Vol.9, No. 3, Februari 2021: 435-442.

Ulfa Nadia & Nazlia Omar, (2019), Malay named entity recognition using rule-based approach, Asia-Pasific Journal of Information Technology and Multimedia Vol.8 No. 1, Jun 2019: 37-47.

Xiaorui Li & Saidah Saad, (2019), Named entity recognition on health informatic using machine learning approach, CAIT, 2019: 96-99.