PENGECAMAN ENTITI NAMA MELALUI PENDEKATAN BERASASKAN PETUA UNTUK BAHASA MELAYU

Sweetnie Wong Chi Lam
Nazlia Omar

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Pengekstrakan maklumat merupakan satu proses penting untuk mendapatkan maklumat yang berharga dan penting dari dokumen atau sumber yang tidak berstruktur. Pada era globalisasi ini, pengekstrakan maklumat telah digunakan secara meluas untuk memperoleh konsep yang bermakna dari dokumen yang tidak berstruktur seperti twit, artikel, blog dan seumpamanya yang tidak dapat difahami oleh komputer. Antara satu cara yang sering digunakan bagi pengekstrakan maklumat adalah pengecaman entiti nama. Pengecaman entiti nama adalah proses untuk mengenal pasti dan mengelaskan entiti nama di dalam sesuatu artikel atau teks. Antara kategori yang selalu dikenal pasti adalah nama individu, lokasi, organisasi, tarikh dan masa. Terdapat beberapa alatan pengecaman entiti nama bahasa Melayu telah dibangunkan. Namun begitu, alatan sedia ada tersebut masih mempunyai ralat dan kekurangan. Terdapat banyak perkataan bagi entiti nama bahasa Melayu yang masih sukar difahami dan diekstrak oleh alatan tersebut. Jadi, kajian ini bertujuan untuk menambahbaikkan alatan pengecaman entiti nama bahasa Melayu melalui kaedah pendekatan berasaskan petua yang berfungsi mengenal pasti dan menandakan nama entiti perkataan Bahasa Melayu. Kajian ini dilakukan dengan mewujudkan korpus bahasa Melayu yang mengandungi entiti nama. Seterusnya, korpus akan dianalisis dengan peraturan yang dibangunkan bagi pengecaman entiti nama seperti individu, organisasi, masa, tarikh dan sebagainya. Kajian ini menggunakan bahasa pengaturcaraan *Python*. Persekitaran pembangunan bersepadu (IDE) yang digunakan dalam kajian ini adalah *Jupyter Notebook*.

1 PENGENALAN

Pemprosesan Bahasa Tabii (PBT) atau Natural Language Processing (NLP) merupakan alatan penting yang digunakan oleh komputer untuk menganalisis dan memahami bahasa manusia (Cynthia 2000). Salah satu bidang yang terlibat dalam teknologi PBT adalah pengekstrakan maklumat (PE) atau information extraction (IE). Pengestrakan maklumat merupakan satu cabang dari PBT dalam kecerdasan buatan yang mempunyai pelbagai aplikasi, termasuk menjawab soalan dan populasi asas pengetahuan (Thien, 2018). PE adalah penting dalam mengekstrakan maklumat yang berharga daripada teks yang tidak berstruktur (Gaizauskas & Wilks, 1998).

Entiti nama adalah maklumat berstruktur yang merujuk kepada nama yang ditetapkan seperti individu, lokasi dan organisasi (Bremer et al, 2006). Pengecaman Entiti Nama (PEN) atau Named Entity Recognition (NER) merupakan bidang yang penting dalam membentuk bidang PBT (Farid et al., 2016). Ia juga adalah salah satu teknik yang popular bagi PE untuk menganalisis dan menandakan entiti nama mengikut kategori (Diego,Menno,Daniel. 2008). Terdapat tiga teknik bagi PEN iaitu pendekatan berasaskan petua, pendekatan berasaskan mesin dan pendekataan hibrid. Tujuan utama PEN adalah untuk mengklasifikasikan teks mengikut entiti nama yang ditetapkan (Hussain, 2020). Istilah entiti nama dicadangkan semasa Persidangan MUC-6 (Alfred et.al.,2014). Merujuk kepada MUC-6, entiti nama mempunyai tiga kategori yang domain iaitu nama (*ENAMEX*), masa (*TIMEX*) dan numerik (*NUMEX*) (Sundheim, 1995, Chinchor, Marsh, 1998). *ENAMEX* berfungsi mengelaskan teks mengikut nama individu, lokasi dan organisasi. *TIMEX* berfungsi untuk mengelaskan teks mengikut masa dan tarikh. *NUMEX* berfungsi untuk mengelaskan peratusan dan nilai daripada sumber teks (Sherief et al., 2012). Entiti nama akan digunakan berdasarkan keperluan domain kajian. Pada masa kini, terdapat banyak kajian PEN telah dilaksanakan dengan beberapa pendekatan seperti pendekatan berasaskan petua, pembelajaran berasaskan mesin dan pendekatan berasaskan hybrid (Inés, 2020). Pembangunan PEN telah mewujudkan pelbagai alatan PEN seperti *FreeLing*, *NCHLT Tagger*, *Name Tage*, *OpenNLP Name Finder* dan sebagainya.

2 PENYATAAN MASALAH

Walaupun terdapat banyak alatan dan teknik berkaitan dengan PEN telah diwujudkan pada masa kini, namun alatan PEN untuk bahasa Melayu masih adalah terhad pada era globalisasi ini. Dengan ini, penyelidikan untuk menghasilkan alatan dan teknik yang lebih sempurna untuk PEN bahasa Melayu adalah diperlukan untuk membantu pengekstrakan maklumat dalam masa singkat. Sebagai contoh, PEN dapat mengimbas keseluruhan artikel dan mengenal pasti entiti nama seperti individu, organisasi, dan tempat. Dengan ini, artikel dapat dikategorikan secara automatik dalam hierarki yang ditentukan dan kandungannya juga mudah ditemui (Hussain, 2020).

Pada masa kini, terdapat beberapa kajian PEN untuk bahasa Melayu telah dijalankan. Namun, kod dan peraturan bagi klasifikasi PEN bahasa Melayu masih tidak mencukupi dan tidak dapat mengecam entiti nama secara menyeluruh (Nadia, 2019). Beberapa kajian PEN bahasa Melayu yang dijalankan telah menunjukkan kekurangan sumber kamus dalam PEN. Kajian PEN Rayner et al. (2013) yang mendapatkan ukuran-f sebanyak 89.47% telah menyimpulkan bahawa kekurangan sumber kamus merupakan salah satu faktor yang mengakibatkan ketidaktepatan hasil dapatan kajian. Seterusnya, kajian PEN bahasa Melayu dengan menggunakan pendekatan yang berasaskan petua kurang diberikan perhatian. Dengan ini, model prototaip berkaitan dengan PEN bahasa Melayu menggunakan pendekatan berasaskan petua perlu dibangunkan.

3 OBJEKTIF KAJIAN

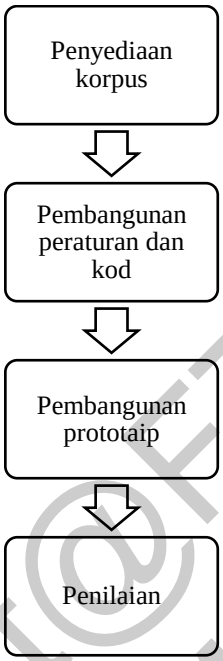
Matlamat utama kajian ini adalah:

- Membangunkan peraturan bagi PEN bahasa Melayu yang berasaskan petua.
- Menghasilkan prototaip yang dapat mengenal pasti jenis entiti nama bahasa Melayu dengan menggunakan peraturan dan kod yang dibangunkan.

4 METOD KAJIAN

Metodologi dalam kajian ini mengandungi 4 fasa iaitu fasa penyediaan korpus, fasa pembangunan peraturan dan kod, fasa pembangunan prototaip bagi klasifikasi entiti

nama dan akhirnya fasa penilaian. Bahagian ini juga akan menunjukkan dan menghuraikan proses bagi setiap fasa secara terperinci. Rajah 1 menunjukkan carta alir proses PEN yang terlibat dalam kajian ini.



Rajah 1 Carta alir proses PEN

4.1 Fasa Penyediaan Korpus

Fasa penyediaan korpus merupakan aktiviti yang utama dalam melakukan sesuatu kajian. Dalam fasa ini, pencarian maklumat yang berkaitan dengan bidang yang ingin dikaji akan dilaksanakan untuk mendapat gambaran awal mengenai tajuk kajian. Seterusnya, rujukan ke atas tesis-tesis dilakukan untuk mengenal pasti kaedah penyelidikan dan dijadikan sebagai panduan dalam menjalankan kajian ini. Perbincangan bersama penyelia juga dilakukan untuk mendapatkan pandangan dan cadangan dalam memperbaiki proses penyelidikan kajian ini. Penyelidikan kajian ini melibatkan sumber teks atau artikel dalam PEN bahasa Melayu dengan menggunakan peraturan dan kod serta sistem prototaip yang dibangunkan. Oleh itu, pengumpulan sumber data daripada laman sesawang yang mempunyai morfologi bahasa Melayu yang sempurna seperti Berita Harian dan Malaysiakini adalah penting dalam fasa ini. Laman sesawang tersebut mempunyai pelbagai kandungan domain seperti sukan, politik,

semasa dan bencana yang sesuai untuk pelaksanaan PEN bagi kajian ini. Rajah 2 menunjukkan contoh sumber berita diambil dari Berita Harian.

Kerjasama, kepakaran Australia signifikan untuk ASEAN - PM

Oleh Sophia Ahmad dan Luqman Arif Abdul Karim - November 14, 2020 @ 1:17pm

bhnews@bh.com.my

KUALA LUMPUR: Kerjasama serta kepakaran Australia dalam mendepani pandemik COVID-19, khususnya dari aspek penyelidikan dan pembangunan, adalah sangat signifikan untuk ASEAN, kata Tan Sri Muhyiddin Yassin.

Perdana Menteri berkata, ASEAN amat mengalu-alukan segala perkongsian pengetahuan dan teknologi inovatif serta kerjasama aktif dari aspek kesihatan awam.

Beliau berkata, kolaborasi mampan itu juga perlu melihat elemen penting lain, termasuk pembangunan vaksin dan perubatan yang selamat, berkesan serta berpatutan, demi faedah pelbagai lapisan masyarakat.

Selain itu, katanya, tumpuan tidak harus terhad kepada usaha menangani pandemik COVID-19 saja, sebaliknya turut diperluaskan kepada kecemasan atau penyakit lain yang mungkin berlaku pada masa depan.

Beliau berkata demikian dalam ucapan intervensi pada Sidang Kemuncak Dwitahunan ASEAN-Australia Kedua yang diadakan secara maya, hari ini.

Rajah 1 Contoh berita teks

4.2 Fasa Pembangunan Peraturan Dan Kod

Peraturan dan kod bagi prototaip PEN bahasa Melayu akan dibangunkan dalam fasa ini. Gazetir digunakan untuk membina senarai kamus bagi PEN. Senarai gazetir akan disediakan mengikut keperluan setiap entiti nama seperti nama untuk individu, nama untuk lokasi dan nama untuk organisasi.

Sebelum membangunkan peraturan bagi PEN, sumber teks yang digunakan dalam kajian ini akan dibahagikan kepada bahagian perkataan yang kecil iaitu token. Proses ini dikenali sebagai tokenisasi (Christopher D, 2008). Seterusnya, gazetir digunakan untuk menghasilkan senarai gazetir yang dapat membantu dalam proses PEN.

Contoh senarai kata kunci bagi Individu:

1. Sophia Ahmad
2. Papagamo
3. Hashim

Contoh senarai kata kunci bagi Lokasi:

1. Kuala Lumpur
2. Malaysia
3. Jalan Sultan Ismail

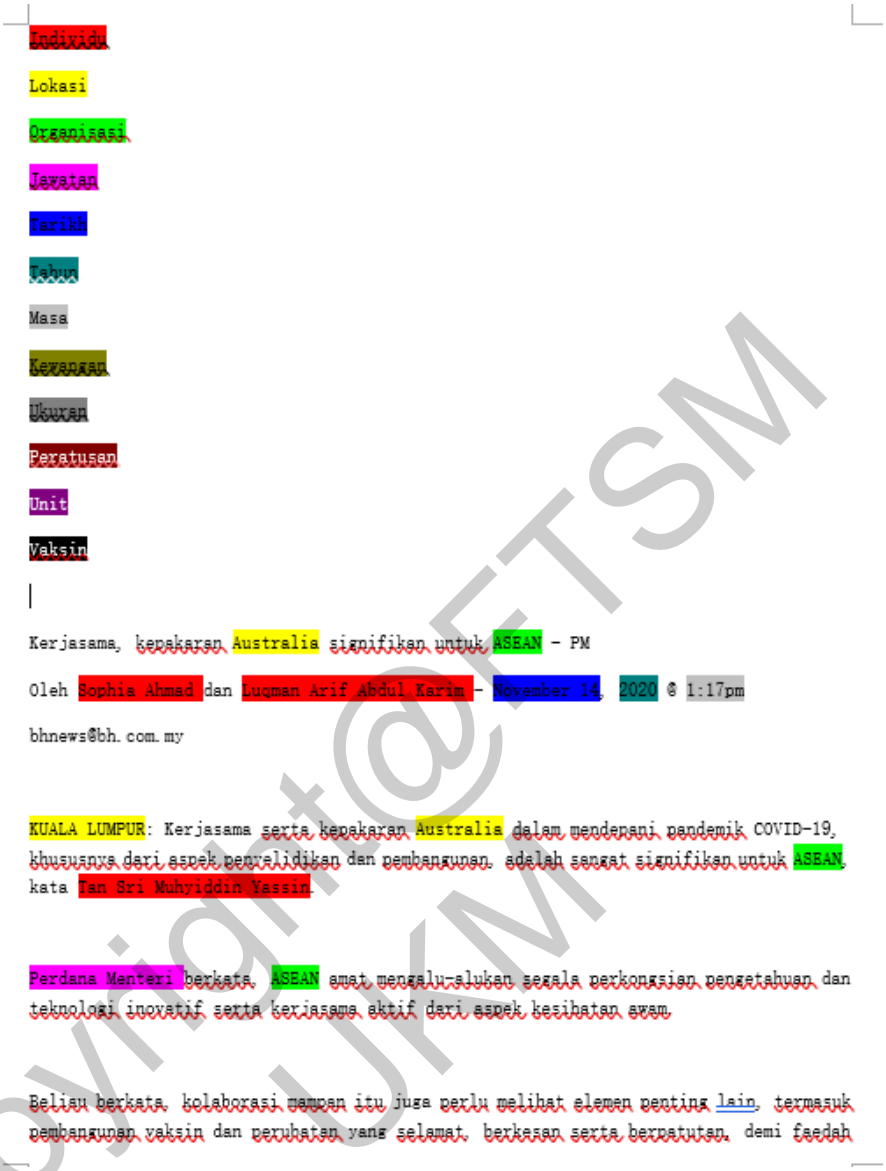
Contoh senarai kata kunci bagi Organisasi:

1. Jabatan Imigresen Malaysia
2. Kementerian Kesihatan Malaysia
3. Tanjung Evolusi Sdn. Bhd

Seterusnya, peraturan- peraturan bagi PEN akan dibangunkan dalam fasa ini. Sistem prototaip yang dibangunkan dalam kajian ini akan menggunakan peraturan dan gazetir pada masa yang sama dalam PEN bahasa Melayu. Kajian ini melibatkan dua belas entiti iaitu nama individu, lokasi, organisasi, jawatan, tarikh, tahun, masa, kewangan, ukuran, peratusan, vaksin dan unit.

Contoh peraturan yang dibangunkan bagi individu:

Jika perkataan pertama (W) mengandungi perkataan gelaran dari senarai gazetir (Nor, Nur, Muhamed, …) dan perkataan seterusnya mempunyai perkataan huruf besar (Wi +1), perkataan tersebut merupakan entiti nama individu. PEN bagi entiti individu akan selesai sekiranya Wi+1 mempunyai titik, koma atau perkataan daripada gazetir. Rajah 3 menunjukkan contoh penandaan entiti yang menggunakan kepakaran manusia. Jadual 1 menunjukkan contoh peraturan entiti nama yang digunakan dalam kajian ini.



Rajah 2 Penandaan entiti nama dalam artikel dengan kepakaran manusia

Jadual 1 Contoh peraturan entiti nama

Jenis Entiti	Penerangan	Peraturan	Contoh
Individu	Entiti yang menerangkan nama orang yang dimula dengan gelaran/pangkat.	Jika perkataan pertama (W) mengandungi perkataan dari gazetir contohnya Sophia dan perkataan seterusnya (Wi+1) mempunyai perkataan huruf besar, jadi kedua- dua perkataaan tersebut adalah entiti nama individu. Sekiranya Wi+1 mempunyai tanda baca atau perkataan yang tidak terlibat dalam senarai gazetir, maka pengecaman untuk entiti nama individu selesai.	**[Sophia INDIVIDU]** **[Ahmad INDIVIDU]** telah menyiapkan …
Lokasi	Entiti yang menerangkan nama tempat.	Jika perkataan pertama (W) mengandungi perkataan dari gazetir contohnya Kuala dan perkataan seterusnya (Wi+1) mempunyai perkataan huruf besar, jadi kedua- dua perkataaan tersebut adalah entiti nama lokasi. Sekiranya Wi+1 mempunyai tanda baca atau perkataan yang tidak terlibat dalam senarai gazetir, maka pengecaman untuk entiti nama lokasi selesai.	**[KUALA LOKASI] [LUMPUR LOKASI**]: Kerjasama serta…
Organisasi	Entiti yang menerangkan nama organisasi, syarikat, perusahaan, pejabat, hospital, polis dan lain-lain.	Jika perkataan pertama (W) mengandungi perkataan dari gazetir contohnya Kementerian dan perkataan seterusnya (Wi+1) mempunyai perkataan huruf besar, jadi kedua- dua perkataaan tersebut adalah entiti nama organisasi. Sekiranya Wi+1 mempunyai tanda baca atau perkataan yang tidak terlibat dalam senarai gazetir, maka pengecaman untuk entiti nama organisasi selesai.	…di [**Kementerian ORGANISASI] [Kerja ORGANISASI] [Raya ORGANISASI**].

Jawatan	Entiti yang menerangkan jawatan dengan diikuti nama organisasi atau tempat.	Jika perkataan pertama (W) mengandungi perkataan dari gazetir contohnya Perdana dan perkataan seterusnya (Wi+1) mempunyai perkataan dari gazetir juga, jadi kedua- dua perkataaan tersebut adalah entiti nama jawatan. Sekiranya Wi+1 mempunyai tanda baca atau perkataan yang tidak terlibat dalam senarai gazetir, maka pengecaman untuk entiti nama jawatan selesai.	… meninggalkan kediaman **[Perdana JAWATAN] Menteri JAWATAN]**, Tan Sri…
Tarikh	Entiti yang menerangkan hari, bulan, tanggal dan gabungannya.	Jika perkataan pertama (W) mengandungi perkataan dari gazetir contohnya November dan perkataan seterusnya (Wi+1) dan sebelumnya (Wi-1) merupakan nombor, jadi kedua- dua perkataaan tersebut adalah entiti nama tarikh. Sekiranya Wi+1 atau Wi-1 adalah perkataan atau symbol, maka pengecaman untuk entiti nama tarikh selesai.	…Arif Abdul Karim – **[November 14 tarikh]** …
Tahun	Entiti yang menerangkan tahun.	Jika perkataan pertama (W) adalah nombor dan bermula dengan 19 dan 20, perkataan tersebut adalah entiti nama tahun. Sekiranya perkataan seterusnya (Wi+1) mempunyai kata yang sama dari gazetir, maka kedua-dua perkataan adalah entiti nama tahun.	Pada tahun [**1990 TAHUN**], terdapat beberapa…
Masa	Entiti yang menerangkan waktu kejadian.	Jika perkataan pertama (W) merupakan nombor dan perkataan akhir adalah am atau pm, maka perkataan tersebut adalah entiti nama masa. Sekiranya perkataan pertama (W) adalah perkataan dari gazetir dan perkataan sebelumnya (Wi-1) adalah nombor, Maka kedua- dua perkataaan	… jam **[7.00 MASA] [pagi MASA]** hari ini.

		tersebut adalah entiti nama masa.	
Kewangan	Entiti yang diawali atau diakhiri dengan tanda mata wang.	Jika perkataan pertama (W) bermula dengan "RM", maka perkataan tersebut adalah entiti nama kewangan. Sekiranya perkataan seterusnya (Wi+1) mengandungi perkataan dari gazetir contohnya ringgit atau juta, maka kedua- dua perkataaan tersebut adalah entiti nama kewangan.	… bernilai **[RM10000.00 KEWANGAN]**.
Ukuran	Entiti yang menerangkan ukuran seperti berat, tinggi, lebar dan seumpamanya.	Jika perkataan pertama (W) adalah nombor dan perkataan seterusnya (Wi+1) mengandungi perkataan dari gazetir contohnya litre, maka kedua- dua perkataaan tersebut adalah entiti nama ukuran.	… sebanyak [**10 UKURAN**][**meter UKURAN**] bagi …
Peratusan	Entiti yang diakhiri dengan peritus (%) atau perkataan peratus.	Jika perkataan pertama (W) adalah nombor dan perkataan seterusnya (Wi+1) mengandungi perkataan dari gazetir contohnya peratus atau %, maka kedua- dua perkataaan tersebut adalah entiti nama peratusan.	… telah meningkat sebanyak **[20% PERATUSAN]**.
Unit	Entiti yang menerangkan unit dengan melibatkan dua atau dua ke atas entiti ukuran seperti kelajuan (ukuran km dengan ukuran s).	Jika perkataan pertama (W) merupakan nombor dan perkataan seterusnya (Wi+1) megandungi perkataan atau symbol daripada gazetir seperti $L^{+1}T^{-1}$, jadi kedua- dua perkataaan tersebut adalah entiti nama unit.	… pelajuan adalah sebanyak **[180km/s UNIT]**.
Vaksin	Entiti yang menerangkan nama vaksin.	Jika perkatan pertama (W) adalah sama dengan perkataan dari gazetir dan perkataan seterusnya (Wi+1) adalah berhuruf besar, maka perkataan	… **[vaksin COVID-19 VAKSIN]** kini mengandungi…

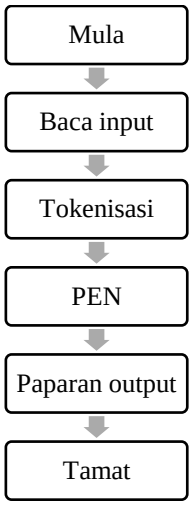
	tersebut adalah entiti vaksin.

Peraturan untuk entiti nama tahun dan unit telah diwujudkan dalam kajian ini. Hal ini kerana berdasarkan penyelidikan Ulfa (2019), entiti nama bagi tahun seperti "*1990an*" tidak dapat dianalisis dalam sistem prototaip tersebut. Seterusnya, entiti nama unit yang mengandungi dua jenis entiti ukuran seperti gabungan kilometer dengan saat tidak dapat dianalisis dalam kajian tersebut. Oleh itu, kajian ini bertujuan untuk menambah baik peraturan yang melibatkan entiti nama unit dan tahun dengan menambah peraturan baru.

4.3 Fasa Pembangunan Prototaip

Prototaip PEN bahasa Melayu yang berasaskan petua dibangunkan dengan menggunakan kod dan peraturan yang dicipta dari fasa pembangunan peraturan dan kod. Prototaip ini akan menandakan entiti nama kepada token atau perkataan bahasa Melayu merujuk kepada senarai gazetir

Sistem prototaip akan dibangunkan dalam fasa ini mengikut perancangan dan fungsi yang ditetapkan. Terdapat beberapa 4 fungsi yang terlibat dalam sistem prototaip ini iaitu fungsi baca input, fungsi tokenisasi, fungsi PEN dan fungsi paparan output. Rajah 4 merupakan carta alir proses sistem protaip.



Rajah 3 Carta alir proses PEN sistem prototaip

Sistem prototaip bermula dengan membaca teks atau artikel yang diambil daripada laman sesawang. Seterusnya, prototaip akan melaksanakan tokenisasi ke atas teks untuk mendapatkan bahagian perkataan yang kecil atau dikenali sebagai token. Hal ini kerana saringan token adalah penting untuk proses PEN nanti. Bagi proses PEN, entiti nama akan dianalisis mengikut susunan yang teratur iaitu individu dan seterusnya lokasi, organisasi, jawatan, tarikh, tahun, masa, kewangan, ukuran, peratusan dan akhirnya unit. Selepas proses PEN, prototaip akan mengeluarkan output, iaitu perkataan yang dikategori mengikut entiti nama tertentu. Proses PEN prototaip akan tamat sekiranya semua perkataan dari teks telah dianalisis.

4.4 Fasa Penilaian

Dalam fasa ini, pengujian sistem prototaip akan dijalankan untuk memastikan setiap unit prototaip berfungsi dengan baik. Sumber korpus yang disediakan akan dijadikan sebagai input bagi sistem prototaip untuk menilai keberkesanan sebagaimana yang digariskan dalam objektif kajian. Seterusnya, hasil pengujian akan direkodkan bagi menilai kejituan kajian dengan formula yang tertentu. Penilaian kajian dilaksanakan dengan formula yang dicadangkan di dalam MUC (Budi, 2005) iaitu kejituan (*precision*), dapatan (*recall*) dan Ukuran -F (*F-measure*). Formula sesuai untuk pelaksanaan kajian ini kerana dapat menperoleh keputusan penilaian yang tepat dengan melibatkan entiti nama yang telah dikategori kepda tiga jenis iaitu tepat, separa tepat dan tidak tepat (Alfred, 2014).

a) Dapatan: bilangan PEN yang tepat oleh sistem.
b) Separa tepat: bilangan PEN separa tepat oleh sistem.
c) Jumlah keseluruhan PEN oleh manual: bilangan PEN yang dilakukan secara manual oleh pakar bahasa.
d) Jumlah keseluruhan PEN oleh sistem: bilangan keseluruhan PEN yang dilakukan oleh sistem termasuk pengecaman yang tepat, separa tepat dan tidak tepat

Formula penilaian kajian:

$$Kejituan = \frac{\text{tepat } + (0.5 * \text{separa tepat})}{\text{jumlah keseluruhan PEN oleh sistem}'} \quad \ldots(1)$$

$$Dapatan = \frac{\text{tepat } + (0.5 * \text{separa tepat})}{\text{jumlah keseluruhan PEN oleh manual}'} \quad \ldots(2)$$

$$Ukuran - F = \frac{\text{kejituan} * \text{dapatan}}{0.5 * (\text{kejituan} + \text{dapatan})} \quad \ldots(3)$$

5 HASIL KAJIAN

Pengujian keberkesanan sistem PEN terhadap korpus latihan dan korpus ujian akan dilakukan. Kajian ini melibatkan 220 teks berita dan akan dibahagikan kepada 70 peratus sebagai korpus latihan dan 30 peratus sebagai korpus ujian. Hasil pengujian iaitu kejituan, dapatan dan ukuran-f akan diperolehi melalui perbandingan antara dapatan yang diperoleh melalui sistem PEN dengan dapatan yang ditetapkan oleh pakar. Pengujian ini melibatkan 3 fasa iaitu penilaian korpus latihan, penilaian korpus ujian dan perbandingan dengan kajian terdahulu.

5.1 Korpus Latihan

Terdapat 190 teks berita telah dikumpul dari laman sesawang Berita Harian dan Malaysiakini untuk dijadikan sebagai korpus latihan dalam kajian ini. Korpus latihan tersebut melibatkan 1048 entiti individu, 1551 entiti lokasi, 1785 entiti organisasi, 372 entiti jawatan, 448 entiti tarikh, 359 entiti tahun, 271 entiti masa, 170 entiti kewangan, 132 entiti peratusan, 111 entiti ukuran, 49 entiti unit dan 71 entiti vaksin. Hasil pengiraan bagi setiap entiti nama dari teks korpus latihan akan direkod ke dalam aplikasi *Microsoft Excel* untuk membantu pengiraan nilai dapatan, kejituan dan ukuran-f.

Jadual 2 menunjukkan satu contoh penandaan entiti korpus latihan dengan pengecaman manual dan pengecaman sistem. Rajah 5 menunjukkan sebahagian hasil pengiraan entiti korpus latihan. Berdasarkan jadual 2, perkataan M dari jadual merupakan jumlah entiti yang dianotasi oleh pakar. Perkataan T merupakan jumlah entiti yang tepat dengan keputusan yang dianotasi oleh pakar. Perkataan S merupakan

jumlah entiti yang separa tepat dengan keputusan yang dianotasi secara manual. Perkataan X merupakan jumlah entiti yang salah ditandakan atau diabaikan oleh sistem.

Jadual 2 Contoh PEN artikel korpus latihan

Artikel	**Entiti Pakar**	**M**	**Entiti Sistem**	**T**	**S**	**X**
Individu	-	0	-	0	0	0
Lokasi	Pelaburan Klang Utara, Malaysia	2	Pelaburan Klang Utara, Malaysia	2	0	0
Organisasi	Jabatan Perkhidmatan Kuarantin dan Pemariksa, Maqis	2	Jabatan Perkhidmatan Kuarantin dan Pemariksa, Maqis	2	0	0
Jawatan	Pengarah	1	Pengarah	1	0	0
Tarikh	Khamis, 3 Apr	2	Khamis, 3 Apr	2	0	0
Tahun	2021, 2011	2	2021, 2011	2	0	0
Masa	5:02pm, 11 pagi	2	5:02pm, 11 pagi	2	0	0
Kewangan	RM3.23 juta, RM3 juta, RM100,000	3	RM3.23 juta, RM3 juta, RM100,000	3	0	0
Peratusan	-	0	-	0	0	0
Ukuran	173,074 kilogram	1	173,074 kilogram	1	0	0
Unit	-	0	-	0	0	0
Vaksin	-	0	-	0	0	0

Individu				Lokasi				Organisasi				Jawatan				Tarikh				Tahun			
M	T	S	X	M	T	S	X	M	T	S	X	M	T	S	X	M	T	S	X	M	T	S	X
0	0	0	0	5	5	0	0	6	6	0	0	0	0	0	0	5	5	0	0	5	5	0	0
4	4	0	0	14	14	0	0	3	3	0	0	0	0	0	0	2	2	0	0	2	2	0	0
2	2	0	0	2	2	0	0	4	4	0	0	1	1	0	0	1	1	0	0	1	1	0	0
6	6	0	0	10	10	0	0	22	22	0	6	10	10	0	0	1	1	0	0	1	1	0	0
1	1	0	0	2	2	0	1	23	23	0	13	2	2	0	0	6	6	0	0	3	3	0	0
7	7	0	0	2	2	1	5	73	73	0	0	7	7	0	0	1	1	0	0	2	2	0	0
28	28	1	1	3	3	0	0	6	6	0	3	0	0	0	0	7	7	0	0	2	2	0	0
18	18	0	0	1	1	0	0	2	2	0	0	0	0	0	0	3	3	0	0	1	1	0	0
1	1	0	0	3	1	0	2	13	13	0	2	1	1	0	0	7	7	0	0	4	4	0	0
0	0	0	0	22	22	0	0	9	9	0	0	0	0	0	0	1	1	0	0	1	1	0	0
10	10	0	0	1	1	0	0	3	3	0	0	2	2	0	0	2	2	0	0	1	1	0	0
4	4	0	0	8	8	0	0	6	6	0	0	6	6	0	0	3	3	0	0	1	1	0	0
11	11	0	0	8	8	0	0	1	1	0	0	0	0	0	0	4	4	0	1	2	2	0	0
4	4	0	0	4	4	0	0	1	1	0	0	3	3	0	0	1	1	0	0	1	1	0	0
3	3	0	0	3	3	0	0	10	10	0	0	1	1	0	0	2	2	0	0	2	2	0	0
5	5	0	0	2	2	0	0	16	16	0	0	2	2	0	0	2	2	0	0	2	2	0	0
11	11	0	0	16	16	0	0	6	6	0	1	3	3	0	0	7	7	0	0	1	1	0	0

Rajah 5 Sebahagian hasil pengiraan entiti korpus latihan

Secara keseluruhannya, kebanyakan entiti dari korpus latihan mendapat keputusan tepat dan separa tepat. Penilaian kejituan, dapatan dan ukuran-f bagi keseluruhan korpus latihan adalah 95.44%, 98.58% dan 96.92%. Hal ini kerana peraturan dan kod yang dibina adalah bersesuaian dengan entiti korpus latihan. Walaupun keputusan penilaian korpus latihan adalah tinggi, namun keputusan penilaian tersebut tidak dapat dijadikan sebagai rujukan kerana peraturan dan kod adalah dibina

mengikut perkataan yang terdapat dalam korpus latihan. Dengan korpus ujian perlu digunakan untuk mengaji keberkesanan sebenar sistem dan peraturan yang dibangukan.

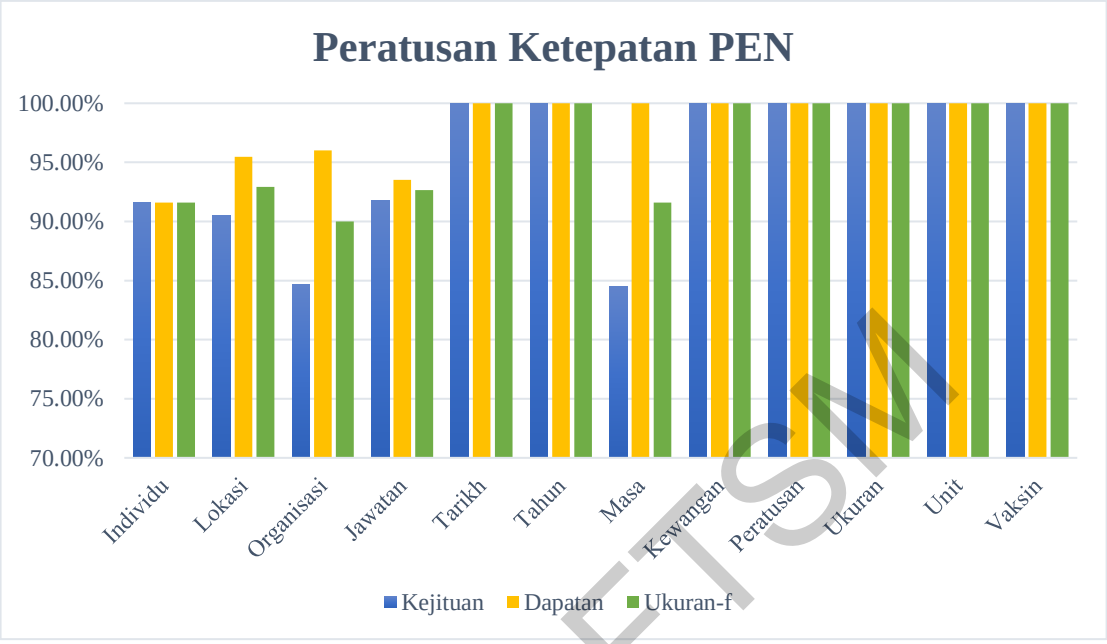
5.2 Korpus Ujian

Korpus ujian melibatkan 30 teks berita yang dikumpul dari laman senawang Berita Harian dan Malaysiakini. Korpus ujian melibatkan 137 entiti individu, 220 entiti lokasi, 150 entiti organisasi, 51 entiti jawatan, 66 entiti tarikh, 49 entiti tahun, 60 entiti masa, 50 entiti kewangan, 24 entiti peratusan, 43 entiti ukuran, 13 entiti unit dan 25 entiti vaksin. Cara penilaian korpus ujian adalah sama dengan cara penilaian korpus latihan. Keputusan korpus ujian akan dinilai dengan formula kejituan, dapatan dan ukuran-f.

Pengujian korpus ujian akan dilakukan dengan peraturan dan kod yang dibina dalam kajian ini untuk mengukur prestasi sistem prototaip. Nilai pengujian setiap entiti telah direkodkan semasa fasa pengujian. Jadual 3 menunjukkan keputusan ringkas yang terperinci bagi pengujian setiap entiti korpus ujian. Rajah 6 menunjukkan carta bar keputusan pengujian ketepatan PEN sistem.

Jadual 3 Keputusan pengujian ringkas bagi korpus ujian

Jenis Entiti	**M**	**T**	**S**	**X**	**Kejituan**	**Dapatan**	**Ukuran-F**
Individu	137	125	1	11	91.60%	91.60%	91.60%
Lokasi	120	208	4	20	90.52%	95.46%	92.92%
Organisasi	150	137	14	19	84.70%	96.00%	90.00%
Jawatan	51	41	7	6	82.41%	87.26%	84.76%
Tarikh	66	66	0	0	100%	100%	100%
Tahun	49	49	0	0	100%	100%	100%
Masa	60	60	0	11	84.51%	100%	91.60%
Kewangan	50	50	0	0	100%	100%	100%
Peratusan	24	24	0	0	100%	100%	100%
Ukuran	43	43	0	0	100%	100%	100%
Unit	13	13	0	0	100%	100%	100%
Vaksin	25	25	0	0	100%	100%	100%



Rajah 6 Carta bar keputusan pengujian sistem PEN

Secara keseluruhannya, jumlah nilai kejituan, dapatan dan ukuran-f bagi korpus ujian adalah sebanyak 94.48%, 97.53% dan 95.91%. Hal ini kerana peraturan yang dibangunkan dalam kajian ini dapat mengelakkan kesalahan penandaan entiti nama secara berkesan. Peraturan spesifik yang dibina juga membantu sistem dalam mengecam perkataan yang tidak dapat dianalisis dengan peraturan umum. Selain itu, pernambahan perkataan dalam sumber gazetir juga mengurangkan kegagalan sistem dalam proses PEN.

5.3 Perbandingan Kajian Lepas

Perbandingan nilai ketepatan dengan kajian Ulfa Nadia (2019) telah dilaksana. Jadual 4 telah menunjukkan hasil keputusan ringkas kajian lepas. Jadual 5 menunjukkan perbandingan keputusan korpus ujian kajian lepas antara dua kajian.

Jadual 4 Hasil keputusan ringkas kajian lepas (UlfaUlfa Nadia, 2019)

Jenis Entiti	Kejituan	Dapatan	Ukuran-F
Individu	88%	88%	88%
Lokasi	93%	93%	93%
Organisasi	97%	93%	95%
Jawatan	93%	92%	93%
Tarikh	97%	97%	97%

Tahun	100%	100%	100%
Kewangan	100%	100%	100%
Peratusan	100%	100%	100%
Ukuran	92%	92%	92%

Jadual 5 Perbandingan keputusan antara dua kajian

Jenis Entiti	**Kejituan**	**Dapatan**	**Ukuran-F**
Kajian lepas (Ulfa Nadia, 2019)	92.13%	90.23%	91.05%
Kajian ini	92.96%,	95.75%	94.27%.

Merujuk kepada jadual di atas, nilai dapatan, kejituan dan ukuran-f adalah lebih tinggi dalam kajian ini berbanding dengan kajian lepas. Hal ini kerana kajian ini melibatkan perkataaan yang lebih banyak dalam sumber gazetir. Selain itu, peraturan yang dibina juga membolehkan sistem mengecam perkataan yang tidak disediakan dalam sumber gazetir. Contohnya, penambahan peraturan untuk mengecam perkataan yang berhuruf besar di depan perkataan nama seperti "*Mohd*", "*Muhamad*", "*Muhamed*", "*Mohamed*". Dengan ini, kegagalan dan kesalahan penandaan entiti nama adalah lebih kurang jika berbanding dengan kajian lepas. Seterusnya, pengelakkan rujukan silang dengan mengubah suai peraturan penandaan entiti nama dalam kajian ini juga membantu dalam meningkatkan ketepatan sistem. Hal ini kerana kesalahan penandaan entiti nama dalam kajian lepas telah mengurangkan nilai dapatan dan kejituan seterusnya memgakibatkan kerendahan nilai ukuran-f.

6 KESIMPULAN

Pengecaman entiti nama adalah proses yang penting bagi pengekstrakan maklumat untuk mengenal pasti dan mengelaskan entiti nama di dalam sesuatu artikel atau teks. Individu dapat mengenal pasti sesuatu maklumat penting dari teks dalam masa yang singkat dengan bantuan entiti nama. Pada masa kini, alatan PEN bagi bahasa Melayu yang sedia ada masih mempunyai ralat yang perlu diperbaiki. Oleh itu, kajian ini bertujuan untuk menambahbaik alatan PEN bahasa Melayu yang berasaskan petua

dengan membangunkan peraturan dan kod yang baru serta menambah senarai kamus yang digunakan untuk analisis entiti nama. Pendekatan berasaskan petua digunakan dalam kajian ini kerana pendekatan ini tidak memerlukan sumber korpus anotasi bahasa Melayu yang banyak. Hal ini kerana penyediaan sumber korpus yang banyak memerlukan masa yang panjang dan ini akan menangguhkan perjalanan proses kajian (Rayner et al., 2013). Selain itu, pendekatan berasaskan petua adalah senang difaham dan boleh dibina mengikut domain yang ditentukan (Patrick, 2013). Penambahbaikan PEN juga mudah dilakukan kerana set kecil peraturan yang mudah dan kurang kompleks (Al-Olimat et al. 2017).

Objektif kajian ini adalah membangunkan peraturan yang baru bagi alatan PEN bahasa Melayu dan menghasilkan prototaip yang dapat menganalisis jenis entiti nama bahasa Melayu dengan menggunakan peraturan dan kod yang dibangunkan. Kajian ini telah berjaya mencapai matlamat yang ditetapkan dengan menwujudkan peraturan dan entiti yang baru seperti entiti tahun, unit dan vaksin. Seterusnya, kajian ini juga berjaya dalam membangunkan alatan PEN bahasa Melayu yang menggunakan bahasa pengaturcaraan *Python* dalam menghasilkan peraturan dan kod kajian ini. Ketepatan kajian ini yang amat baik berbanding dengan kajian lepas iaitu sebanyak 94.48% bagi kejituan, 97.53% bagi dapatan dan 95.91% bagi ukuran-f telah menunjukkan keberkesanan peraturan yang dibina dalam kajian ini.

7 RUJUKAN

Adi Bronshtein. 2017. A Quick Introduction to the “Pandas” Python Library. https://towardsdatascience.com/a-quick-introduction-to-the-pandas-python-library-f1b678f34673. [16 December 2020].

Alireza Mansouri, Lilly Suraini Affendey, Ali Mamat. 2008. Named Entity Approaches. *IJCSNS International Journal of Computer Science and Network Security* 8(2), 339-344 http://paper.ijcsns.org/07_book/200802/20080246.pdf [14 Oktober 2020] [18 November 2020].

Annamma Abraham. 2013. Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *IJARAI International Journal of Advanced Research in Artificial Intelligence 2(2), 34-38*

https://www.researchgate.net/publication/273246843_Comparison_of_Supervised_and_Unsupervised_Learning_Algorithms_for_Pattern_Classification. [16 Oktober 2020].

Anthony, P., Alfred, R., Leong, L. C., On, C. K. & Anthony, P. 2014. Malay Named Entity Recognition Based on Rule-Based Approach (1). Doi:10.7763/IJMLC.2014. V4.428. [16 Oktober 2020].

Budi, I., Bressan, S., Wahyudi, G., Hasibuan, Z.A., Nazief, B.A.A. 2005. Named Entity Recognition for the Indonesian Language: Combining Contextual. Discovery Science. *LNCS (LNAI),* 3735: 57–69 https://link.springer.com/chapter/10.1007/11563983_7 [20 Oktober 2020] [15 November 2020] [4 May 2021].

Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze. 2008. Introduction to Information Retrieval. https://nlp.stanford.edu/IR-book/ [14 November 2020].

Diego Moila, Menno van Zaanen, Daniel Smith. 2008. AFNER-Named Entity Recognition. http://afner.sourceforge.net/what.html [18 Oktober 2020] [6 May 2021].

E. G. Bremer. 2006. Extracting Named Entities Using Support Vector Machines. KDLL, LNBI 3886, 91-103. https://link.springer.com/chapter/10.1007/11683568_8. [18 Oktober 2020].

Farid Morsidi, Sulaiman Sarkawi, Suliana Sulaiman, Siti Asma Mohammad, Rohaziah Abdul Wahid. 2015. Malay Named Entity Recognition: A Review. *Journal of ICT in Education (JICTIE) ISSN 2289-7844* 2(2015), 1-14 https://www.researchgate.net/publication/301515757_Malay_Named_Entity_Recognition_A_Review [16 Oktober 2020].

Husaain Mujtaba. 2020. What is Named Entity Recognition (NER) Application. https://www.mygreatlearning.com/blog/named-entity-recognition/ [20 Oktober 2020].

Inés Roldós. 2020. Named Entity Recognition: Concept Guide and Tools. https://monkeylearn.com/blog/named-entity-recognition/ [18 Oktober 2020].

Naji F. Mohammad. 2012. Arabic Named Entity Recognition using Artificial Neural Network. *Journal of Computer Science* 8(8): 1285-1293. https://core.ac.uk/download/pdf/25889816.pdf [15 Oktober 2020].

Rayner Alfred, Leow Chin Leong, Chin Kim On, Patricia Anthony. 2014. Malay Named Entity Recognition Based on Rule-Based Approach. *International Journal of Machine Learning and Computing*, 4(3), 300- 306 https://www.researchgate.net/publication/275644054_A_Rule-Based_Named-Entity_Recognition_for_Malay_Articles [14 Oktober 2020] [16 November 2020].

Rohini Srihani, Cheng Niu, Wei Li. 2002. A Hybrid Approach for Named Entity and Sub-Type Tagging. https://www.researchgate.net/publication/2538729_A_Hybrid_Approach_for_Named_Entity_and_Sub-Type_Tagging [16 Oktober 2020].

Saidah Saad. Mohamed Kamil Mansor. 2018. Pendekatan Teknik Pengecaman Entiti Nama Bagi Capaian Berita Jenayah Bahasa Melayu. *Journal of Language Studies Volume* 18(4): 216 -235. http://ejournal.ukm.my/gema/article/view/28999/8687 [15 November 2020].

Siti Azirah Asmai, Muhammad Sharilazlan Salleh, Halizah Basiron, Sabrina Ahmad. 2018. An Enhanced Malay Named Entity Recognition using Combination Approach for Crime Textual Data Analysis. *(IJACSA) International Journal of Advanced Computer Science and Applications,* 9(9): 474-483. https://thesai.org/Downloads/Volume9No9/Paper_60-An_Enhanced_Malay_Named_Entity_Recognition.pdf [21 Oktober 2020].

Sulaiman.S, R. Abdul Wahid, Sarkawi.S, Omar.N. 2017. Using Stanford NER and Illinois NER to Detect Malay. *International Journal of Computer Theory and Engineering* 9(2): 147-150. http://www.ijcte.org/vol9/1128-S014.pdf [18 Oktober 2020].

Ulfa Nadia. 2019. Malay Named Entity Recognition using Rule Based Approach. *Asia-Pacific Journal of Information Technology and Multimedia* 8(1), 37-47 http://journalarticle.ukm.my/14150/1/30965-104286-1-PB.pdf [15 Oktober 2020] [11 November 2020].