Pendekatan Teknik Pengecaman Entiti Nama Bagi Capaian Berita Jenayah Bahasa Melayu

Saidah Saad
saidah@ukm.edu.my
Fakulti Teknologi dan Sains Komputer,
Universiti Kebangsaan Malaysia

Mohamed Kamil Mansor
kamilnizam@gmail.com
Fakulti Teknologi dan Sains Komputer,
Universiti Kebangsaan Malaysia

ABSTRAK

Pengekstrakan maklumat merupakan satu proses bagi mendapatkan konsep penting dalam mewakili kandungan teks dari dokumen yang tidak berstruktur. Pada masa kini, terdapat banyak dokumen yang tidak berstruktur seperti teks berita, artikel blog, forum, tweet serta mikro blog dari rangkaian sosial. Dokumen-dokumen ini amat sukar untuk difahami oleh komputer. Oleh itu, kajian berkaitan pengekstrakan maklumat menjadi sangat penting bagi mengatasi permasalah ini. Salah satu teknik pengekstrakan yang banyak digunakan ialah pengecaman entiti nama. Kajian ini dijalankan bagi mengimplementasikan teknik pengecaman entiti nama dari sumber dokumen berita jenayah bahasa Melayu. Objektif utama kajian ini adalah untuk membangunkan sistem prototaip model pengekstrakan maklumat berita jenayah dalam bahasa Melayu dengan menggunakan teknik pengecaman entiti nama melalui pendekatan berasaskan peraturan. Kajian ini dilakukan dengan mewujudkan korpus berita jenayah dalam bahasa Melayu yang diperolehi dari sumber arkib berita BERNAMA. Korpus ini kemudiannya diteliti secara manual oleh pakar bahasa bagi mengecam entiti nama seperti individu, organisasi, lokasi, tarikh, masa, kewangan, peratusan, jenayah dan senjata. Dalam masa yang sama, sistem prototaip dibangunkan serta diuji dengan korpus yang sama dan hasil dari pengujian ini dibandingkan dengan keputusan pakar. Secara keseluruhannya, ujian sistem prototaip ini menunjukkan hasil yang baik dengan nilai dapatan bagi *recall* sebanyak 78.67%, manakala bagi *precision* ialah sebanyak 71.11% dan *F-measure* sebanyak 74.7%. Hasil dari kajian ini diharap dapat menyumbang kepada pengetahuan mengenai keberkesanan teknik pengecaman entiti nama bagi berita jenayah bahasa Melayu dan seterusnya dapat membantu para penyelidik, polis, peguam serta pihak berkuasa yang terlibat dalam bidang jenayah menyelesaikan jenayah dengan lebih cepat dan berkesan.

Kata kunci: pengekstrakan maklumat; pengecaman entiti nama; Bahasa Melayu, berita jenayah, pendekatan berasaskan peraturan.

Named Entity Recognition Approach for Malay Crime News Retrieval

ABSTRACT

Information extraction is a process of obtaining an important concept in representing the textual content of unstructured documents. At present, there are a lot of unstructured documents such as news, articles, blogs, forums, tweets and micro-blogs of social networks. These documents are very difficult to be understood by the computer. Therefore, studies on the extraction of information is very important to overcome this problem. One extraction

technique that is widely used is the entity name recognition. This research aims to implement the entity name recognition techniques of crime news source document in Malay language. The main objective of this study is to develop a prototype system model information extraction crime news in the Malay language using name entity recognition through a rule-based approach. This assessment is done by creating a corpus of crime news in the Malay language which is derived from the archival source; BERNAMA news. The corpus is then examined manually by linguists to identify individual entities such as name, organization, location, date, time, financial, percentage, crime and weapons. At the same time, a prototype system was developed and tested with the same corpus and the results of these tests were compared with the results of an expert. Overall, these tests showed good results with the findings for the recall at 78.67%, while precision is at 71.11% and for F-measure at 74.7%. The results of this study are expected to contribute knowledge regarding the effectiveness of the entity's name recognition techniques for crime news Malay language. This could further assist investigators, police, lawyers and authorities involved in the field of crime in solving crimes more quickly and effectively.

Keywords: information extraction; named entity recognition; malay language, crime news, rule-based approach

PENGENALAN

Pemprosesan Bahasa Tabii (NLP) merupakan satu bidang yang sangat penting pada masa kini. Ianya bermula pada penghujung tahun 1940-an ketika mesin penterjemah mula diperkenalkan dalam perang dunia kedua bagi menyahkod maklumat pihak musuh. Kemudian pada era 1980-an, penyelidikan dalam bidang ini mula berkembang dan menjadi sangat penting sejajar dengan perkembangan teknologi komputer (Alfred et al., 2014). Terdapat beberapa cabang yang mengaplikasikan penggunaan teknologi NLP ini termasuklah Capaian Maklumat (IR) (Masnizah et al.,2018), Pengekstrakan Maklumat (IE) (Shahrul Azman et al., 2018) dan Soalan-Jawapan (QA) (Liddy 2001).

Pengekstrakan Maklumat (IE) merupakan satu proses bagi mengekstrak maklumat dari dokumen yang tidak berstruktur. Menurut takrifan Kamus Dewan Edisi Ketiga, dokumen membawa maksud sesuatu yang bertulis atau bercetak seperti berita yang digunakan sebagai rekod atau bukti (Kamus Dewan Edisi Ketiga 2002). Terdapat tiga jenis dokumen yang digunakan sebagai data input iaitu dokumen berstruktur seperti pangkalan data, dokumen separa struktur seperti fail XML dan dokumen tidak berstruktur seperti dokumen teks berita (Alfred et al., 2014).

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

LATAR BELAKANG KAJIAN

Pengecaman Entiti Nama (NER) merupakan salah satu dari cabang penyelidikan Pengekstrakan Maklumat (IE) yang amat penting. Pengekstrakan maklumat ini bertujuan untuk mendapatkan senarai kata kunci yang relevan bagi sesuatu dokumen. Melalui NER, pengekstrakan maklumat dapat dilakukan dengan mengenalpasti kata nama serta

mengklasifikasikan mengikut kategori individu, organisasi, lokasi, nilai kewangan, nilai peratusan dan tarikh atau masa. Sebagai contoh nama bagi individu ialah Datuk Seri Najib, manakala lokasi ialah Putrajaya.

Terdapat dua badan penyelidikan yang aktif dalam bidang ini iaitu Message Understanding Conference (MUC) yang dipelopor dan dibiaya oleh DARPA dan program Automatic Content Extraction (ACE) (Cunningham, 2006; Darwich, 2014). Kedua-dua badan ini telah menganjurkan persidangan dan menggariskan panduan bagi proses pengekstrakan maklumat ini. NER boleh dibangunkan melalui tiga teknik iaitu pendekatan berasaskan peraturan, pendekatan berasaskan statistik serta gabungan kedua-dua pendekatan ini (Alfred et al., 2014). Pendekatan berasaskan peraturan memerlukan penglibatan kepakaran manusia untuk mengenalpasti dan mengekstrak entiti, sama ada dari segi tatabahasa, sintaksis atau gabungan dengan senarai kata kunci (Sari, Hassan & Zamin, 2010). Manakala pendekatan berasaskan statistik pula melibatkan penggunaan teknik pembelajaran mesin seperti *neural network, decision tree* dan lain-lain (Mohammed & Omar, 2012).

Pengecaman dan pengklasifikasian NER turut bergantung pada pengkhususan domain dan juga bahasa tertentu. Sebagai contoh, domain jenayah memerlukan kata kunci khusus bagi mengenalpasti jenis jenayah seperti merompak, membunuh dan sebagainya. Begitu juga dengan bahasa, setiap bahasa mempunyai sifat dan tatabahasa yang berbeza untuk melakukan pengecaman. Sebagai contoh, kata nama seperti kata nama khas bagi bahasa Inggeris adalah mudah dikenalpasti dengan pengenalan huruf besar di awal perkataan. Namun berbeza dengan bahasa Arab yang mana ianya tidak mempunyai huruf besar (Asharef, 2012). Begitu juga dengan bahasa Melayu yang mempunyai sifat dan tatabahasa yang berbeza dengan bahasa Inggeris.

Penyelidikan NER turut dilakukan dalam pelbagai domain seperti domain jenayah, domain pendidikan, umum dan seumpamanya. Untuk domain jenayah, pegawai penguatkuasa undang-undang dan pegawai penyelidik memerlukan maklumat penting berkaitan dengan sesuatu kes jenayah. Maklumat-maklumat ini tersimpan di dalam dokumen yang tidak berstruktur sama ada dari dokumen teks berita mahupun laporan polis yang dibuat oleh pengadu. Capaian yang dilakukan secara manual terhadap dokumen-dokumen ini akan mengambil masa yang lama dan tidak praktikal dilakukan bagi dokumen yang banyak. Pegawai penyelidik sepatutnya memerlukan capaian yang pantas agar kes yang dikendali dapat diselesaikan dengan cepat dan berkesan. Oleh itu, penyelidikan terhadap permasalahan ini telah dilakukan dan beberapa model sistem telah dibangunkan dan terbukti berkesan dalam membantu para penyiasat untuk mendapatkan maklumat dari dokumen jenayah dengan lebih cepat dan berkesan (Chao et al., 2002; Hao et al., 2008; Alruily et al., 2009; Darwich, 2014).

Penyelidikan NER dalam domain jenayah turut dilakukan dalam bahasa selain dari bahasa Inggeris. Sebagai contoh, aplikasi Arabic Named Entity Recognition (NERA) dibina khusus bagi domain jenayah dalam bahasa Arab (Asharef, 2012). Selain itu, penyelidikan NER turut giat dilakukan ke atas bahasa Indonesia, Melayu dan Iban (Alfred et al., 2013; Budi, Bressan & Wahyudi, 2005; Fong, Ranaivo-Malançon & Wee, 2011, Shahrul Azman et al., 2018). Namun, buat masa ini masih belum ada lagi penyelidikan yang dibina khusus bagi mengenalpasti jenis NER dalam domain jenayah bahasa Melayu.

Berdasarkan kajian lepas, perangkaan statistik menunjukkan kadar jenayah di Malaysia didapati meningkat setiap tahun (Tang, 2009; Amin et al., 2014; Faizah, 2015; Ishak & Bani, 2017). Walaupun terdapat pengurangan dari segi laporan yang dibuat pada kebelakangan ini, namun kes-kes yang diterima masih tinggi dan kes yang tertunggak masih banyak. Pegawai penyelidik dan para penyiasat memerlukan bahan rujukan untuk dijadikan kajian bagi menyelesaikan kes-kes yang kian bertambah. Oleh itu, pengurusan data bagi membolehkan pemprosesan yang lebih cepat dan berkesan perlu dilaksanakan.

Pengecaman Entiti nama (NER) merupakan salah satu teknik pengekstrakan maklumat yang dapat membantu mengekstrak serta mengenalpasti maklumat yang diingini. Maklumat ini kemudiannya disimpan ke dalam pangkalan data bagi memudahkan carian maklumat dilakukan. Matlamat utama NER adalah untuk mengklasifikasikan entiti nama seperti individu, organisasi, lokasi dan sebagainya (Hadi, 2011). Pada kebiasaannya, teknik NER dipengaruhi oleh domain yang dikaji. Setiap domain memerlukan pengecaman yang tersendiri bagi memperoleh keputusan yang baik. Begitu juga dengan penggunaan bahasa, setiap bahasa mempunyai sifat dan tatabahasa yang unik dan tersendiri (Alfred et al., 2014). Ini kerana NER kebiasaannya melibatkan proses pengecaman perkataan seperti morfologi, part of speech (POS) dan klasifikasi berasaskan tesaurus serta penggunaan senarai kata kunci atau kamus perkataan.

Daripada penyelidikan lepas didapati kajian melibatkan teks domain jenayah merangkumi pelbagai bahasa seperti bahasa Inggeris, Arab dan Hindu (Alfred et al., 2014). Namun, kajian NER bagi domain jenayah di dalam bahasa Melayu masih belum dilakukan kerana penyelidikan dalam bidang Pengekstrakan Maklumat (IE) yang menggunakan dokumen bahasa Melayu masih baru. Manakala dalam bidang analisis jenayah pula, masih belum terdapat lagi korpus yang dibina dalam bahasa Melayu. Walaupun berita jenayah turut ditulis dalam bahasa Inggeris, namun maklumat seperti laporan polis masih menggunakan bahasa Melayu. Ini menjadi motivasi untuk membangunkan model prototaip sistem pengecaman entiti nama bagi dokumen jenayah bahasa Melayu.

Sehubungan itu, kajian ini bertujuan membangunkan satu teknik pengekstrakan maklumat NER bagi mengenalpasti dan mengekstrak entiti nama melalui teknik pendekatan berasaskan peraturan. Pengecaman dibuat ke atas dokumen teks berita jenayah bahasa Melayu yang diperolehi dari sumber berita di Malaysia melalui arkib BERNAMA.

KAJIAN LITERATUR

PENGECAMAN ENTITI NAMA (NER)

Pengecaman Entiti nama (NER) telah digunakan secara meluas dan memainkan peranan penting khususnya dalam bidang berkaitan NLP. Entiti ini boleh dikenalpasti sama ada melalui kata nama khas, kata hubung serta kata sendi. Sebagai contoh, bagi kata nama khas individu diklasifikasikan sebagai kelas individu, manakala nama tempat atau lokasi diklasifikasikan dalam kelas lokasi (Hadi, 2011).

Dalam persidangan Message Understanding Conference (MUC-6) yang telah dianjurkan oleh Defense Advanced Research Projects Agency (DARPA) pada tahun 1995, para penyelidik telah menggariskan tiga kategori entiti (Grishman & Sundheim, 1996) iaitu:

- *ENAMEX*: Mengandungi kelas *PERSON* bagi nama individu, kelas *LOCATION* bagi lokasi dan kelas *ORGANIZATION* bagi organisasi.
- *TIMEX*: Mengandungi kelas *DATE* bagi tarikh atau hari dan kelas *TIME* untuk masa.
- *NUMEX*: Mengandungi kelas *PERCENTAGE* bagi nilai peratusan dan kelas *MONETARY* bagi nilai kewangan.

Terdapat dua pendekatan yang popular untuk melaksanakan pengecaman entiti nama (NER) iaitu melalui pendekatan berasaskan peraturan (rule-based approach) dan pendekatan statistik (statistical approach). Pendekatan berasaskan peraturan merupakan satu kaedah yang bergantung sama ada pada peraturan heuristik (heuristic) atau ekspresi peraturan (regular expression) untuk mengklasifikasikan nama. Ia juga boleh bergantung pada 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 turut digunakan bersama sumber bahasa seperti kamus, senarai gazetteers atau senarai penanda bagi mengklasifikasikan entiti nama tersebut. Kajian-kajian yang dibentangkan dalam persidangan MUC, didapati pendekatan berasaskan peraturan menghasilkan keputusan yang lebih baik berbanding pendekatan yang lain (Eikvil, 1999).

Manakala pendekatan statistik atau juga dikenali sebagai pendekatan pembelajaran mesin (machine learning approach) merupakan teknik pembelajaran corak klasifikasi daripada korpus yang besar. Bagi membolehkan pengecaman entiti nama berfungsi, sistem perlu mempelajari dengan mengenalpasti corak daripada korpus yang besar dan membina keputusan berdasarkan data tersebut.

Terdapat tiga kategori teknik bagi pendekatan statistik iaitu pengawasan (supervised), semi-pengawasan (semi-supervised) dan tanpa pengawasan (unsupervised). Contoh bagi teknik pengawasan ialah seperti penggunaan teknik Decision Tree (Sekine, 1998), dan teknik Hidden Markov Model (Bikel et al., 1997). Contoh bagi teknik semi-pengawasan ialah dengan menggunakan teknik *bootstrapping* yang memuatkan sedikit pengawasan pada permulaan prosedur pembelajaran. Manakala contoh bagi teknik tanpa pengawasan adalah seperti bergantung pada corak leksikal serta pengiraan statistik ke atas korpus yang besar. Jadual berikut memaparkan perbandingan antara pendekatan berasaskan peraturan dan statistik serta kelebihan dan kekurangannya.

DOMAIN JENAYAH

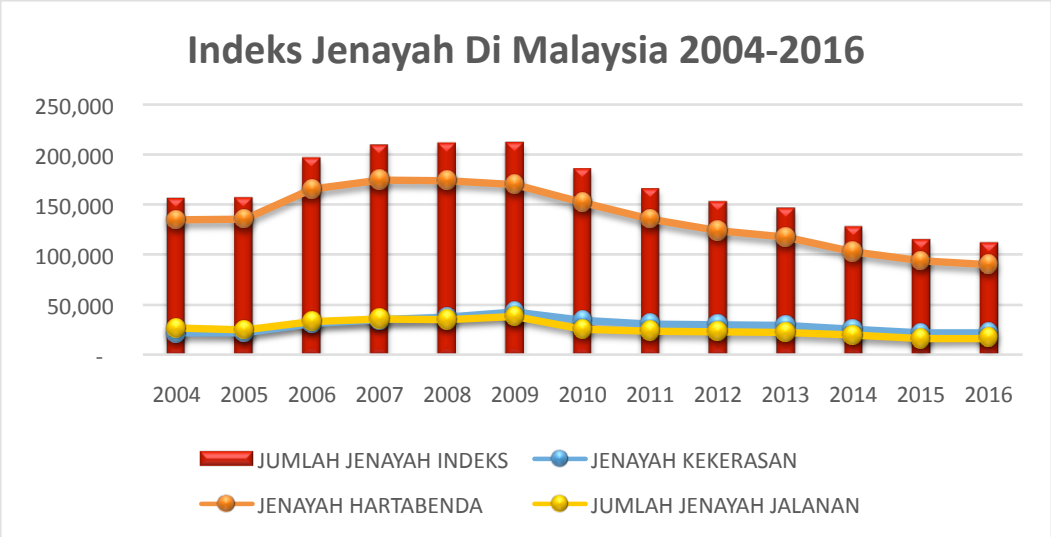
Menurut Kamus Dewan Edisi Ketiga, jenayah didefinasikan sebagai satu perbuatan jahat yang salah di sisi undang-undang seperti mencuri, merompak, membunuh dan lain-lain (Kamus Dewan Edisi Ketiga, 2002). Terdapat dua kategori jenayah iaitu jenayah kekerasan dan jenayah harta benda. Jadual 1 di bawah menyenaraikan jenis jenayah mengikut kategori tersebut.

JADUAL 1. Senarai jenayah kekerasan dan jenayah harta benda

Jenayah Kekerasan	Jenayah Harta Benda
Bunuh	Curi
Rogol	Curi lori/van
Samun berkumpulan bersenjata api	Curi motokar
Samun berkumpulan tanpa bersenjata api	Curi motosikal
Samun bersenjata api	Curi ragut
Samun tanpa bersenjata api	Pecah rumah siang
Mencederakan	Pecah rumah malam

Sumber: Portal Jabatan Perangkaan Malaysia

Berdasarkan statistik terkini yang dikeluarkan oleh Jabatan Perangkaan Malaysia, statistik indeks jenayah di Malaysia didapati telah menurun kepada angka 112,000 kes pada penghujung tahun 2016. Walau bagaimanapun, jumlah keseluruhan indeks jenayah di Malaysia masih dianggap tinggi berdasarkan terdapat banyak kes-kes yang masih tertangguh dan juga kes-kes yang tidak dapat diselesaikan (Darwich, 2014; Tang, 2009; Ohashi, 2004). Rajah 1 memaparkan graf indeks jenayah mengikut pecahan jenis jenayah kekerasan dan jenayah harta benda.



Sumber: Jabatan Perangkaan Malaysia dan PDRM, Bukit Aman

RAJAH 1. Indeks jenayah di Malaysia mengikut jenis jenayah (2004-2016)

KAJIAN LEPAS

Kajian yang melibatkan penggunaan teknik pengecaman entiti nama dalam domain jenayah telah banyak dilakukan oleh para penyelidik diseluruh dunia. Kajian ini penting bagi memastikan capaian dapat dilakukan dengan lebih cepat dan tepat. Memandangkan penyiasatan kes jenayah yang dilakukan secara manual agak rumit, ia akan menjadi lebih mudah jika maklumat yang berkaitan seperti individu, lokasi, masa, kenderaan, senjata dan lain-lain entiti dapat diekstrak terlebih dahulu sama ada dari laporan polis atau dokumen berita jenayah. Jadual 2 menunjukkan perbandingan kajian-kajian lepas dalam domain jenayah serta pengecaman entiti nama dalam bahasa Melayu.

JADUAL 2. Jadual perbandingan kajian-kajian lepas

Penyelidik (Tahun)	Bahan Penyelidikan	Domain Penyelidikan	Kaedah Penyelidikan
Chao (2002)	Dokumen bahasa Inggeris	Jenayah	Pengecaman entiti nama dengan menggabungkan teknik pencarian leksikal serta pendekatan berasaskan peraturan dan rangkaian neural.
Hao (2008)	Dokumen bahasa Inggeris	Jenayah	Pengecaman entiti nama dengan menggunakan teknik pencarian leksikal serta pendekatan berasaskan peraturan.
Alruily et al. (2009)	Dokumen bahasa Inggeris dan Arab	Jenayah	Pengecaman entiti nama (CTRS) dengan menggabungkan teknik pengecaman berasaskan peraturan dan *gazetteer*.
Safwati (2011)	Dokumen bahasa Inggeris	Umum	Pengecaman entiti nama dengan menggunakan alatan OpenCalais.
Darwich (2014)	Dokumen bahasa Inggeris	Jenayah	Pengecaman entiti nama dan *co-reference* bagi mengenalpasti kewarganegaraan bagi suspek ataupun mangsa.
Mona Asharef (2012)	Dokumen bahasa Arab	Jenayah	Pengecaman entiti nama dengan menggunakan pendekatan berasaskan peraturan bagi maklumat mofrologi serta senarai jenayah dan kata kunci.
Naji (2012)	Dokumen bahasa Arab	Jenayah	Pengecaman entiti nama dengan menggunakan pendekatan rangkaian neural dan koleksi korpus bahasa Arab.

Indra Budi et al. (2005)	Dokumen bahasa Indonesia	Umum	Pengecaman entiti nama (InNER) dengan menggunakan pendekatan berasaskan peraturan terhadap kontekstual, morfologi dan tanda pengelasan perkataan (POS).
Yunita Sari et al. (2010)	Dokumen bahasa Melayu	Umum	Pengecaman entiti nama dengan menggabungkan teknik pendekatan berasaskan peraturan dan statistik semi-pengawasan melalui alatan *Link Grammer* dan *Stanford POS* serta algoritma *Self-Training*.
Soo-Foo Yong et al. (2011)	Dokumen bahasa Iban	Umum	Pengecaman entiti nama (NERSIL) dengan menggunakanpendekatan berasaskan peraturan melalui alatan GATE dan modul ANNIE bagi mengekstrak entiti dan menjana peraturan.
Rayner Alfred et al. (2013)	Dokumen bahasa Melayu	Umum	Pengecaman entiti nama melalui pendekatan berasaskan peraturan berdasarkan kontekstual dan tanda pengelasan perkataan (POS) serta gabungan senarai kamus perkataan *gazetteer*.

Kajian berkaitan dengan teknik pengecaman entiti nama bagi dokumen bahasa Melayu masih baru jika dibandingkan dengan bahasa lain. Pada masa ini, aplikasi komersil bagi pengecaman entiti nama hanya terdapat dalam bahasa Inggeris. Ia menjadi keperluan bagi penyelidik mengkaji dan menghasilkan model yang sesuai dengan bahasa Melayu.

Selain itu, teknik pengecaman entiti nama ini amat penting jika digunakan bagi domain tertentu khususnya seperti domain jenayah yang memerlukan capaian yang pantas dan tepat (Croft, Metzler & Strohman, 2010). Ini kerana pencarian maklumat dari sumber yang pelbagai akan memakan masa yang lama. Melalui penggunaan pengecaman entiti nama ini, entiti-entiti tertentu dapat diekstrak terlebih dahulu dan disimpan ke dalam pangkalan data bagi tujuan carian pada masa akan datang.

Kajian yang dijalankan adalah tertumpu kepada pembangunan model pengecaman entiti nama berasaskan peraturan bagi dokumen jenayah bahasa Melayu. Pemilihan pendekatan berasaskan peraturan dibuat berbanding pendekatan berasaskan statistik kerana bilangan korpus yang dikumpulkan adalah tidak mencukupi untuk dijadikan data pembelajaran. Ini kerana pendekatan berasaskan statistik memerlukan sejumlah bilangan korpus yang besar untuk dijadikan data pembelajaran. Proses pengumpulan korpus ini memerlukan masa yang panjang dan tidak sesuai bagi pembangunan sistem prototaip yang singkat seperti ini (Alfred et al., 2013).

Di samping itu, berdasarkan kajian lepas didapati penggunaan penandaan golongan kata (POS) turut memberi kesan yang lebih rendah berbanding saringan morfologi (Budi et al., 2005). Begitu juga dengan pendekatan berasaskan statistik, hasil keputusan adalah sederhana jika dibandingkan dengan pendekatan berasaskan peraturan dan gazetteer (Alfred et al., 2013; Asharef, 2012; Budi et al., 2005; Esmaail, 2012). Oleh itu, pembangunan model pengecaman entiti nama bagi dokumen jenayah bahasa Melayu dengan menggunakan pendekatan berasaskan peraturan perlu dibangunkan.

JUSTIFIKASI DAN SUMBER DATA

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

Kajian yang akan dijalankan adalah tertumpu kepada pembangunan model pengecaman entiti nama bagi dokumen jenayah bahasa Melayu. Berdasarkan kajian lepas, didapati penggunaan tanda pengklasifikasian perkataan (POS) turut memberi kesan yang lebih rendah berbanding saringan morfologi (Budi et al., 2005). Begitu juga dengan pendekatan berasaskan statistik, hasil keputusan adalah sederhana jika dibandingkan dengan pendekatan berasaskan peraturan dan *gazetteer* (Alfred et al., 2013; Asharef, 2012; Budi et al., 2005; Esmaail, 2012). Oleh itu, pembangunan model pengecaman entiti nama bagi dokumen jenayah bahasa Melayu dengan menggunakan pendekatan berasaskan peraturan perlu dibangunkan.

Seperti yang dinyatakan, sumber data bagi kajian ini diperolehi dari laman web arkib berita BERNAMA, kerana BERNAMA menyediakan arkib berita bahasa Melayu yang lengkap termasuk berita dari surat khabar lain seperti Berita Harian dan Utusan Malaysia.

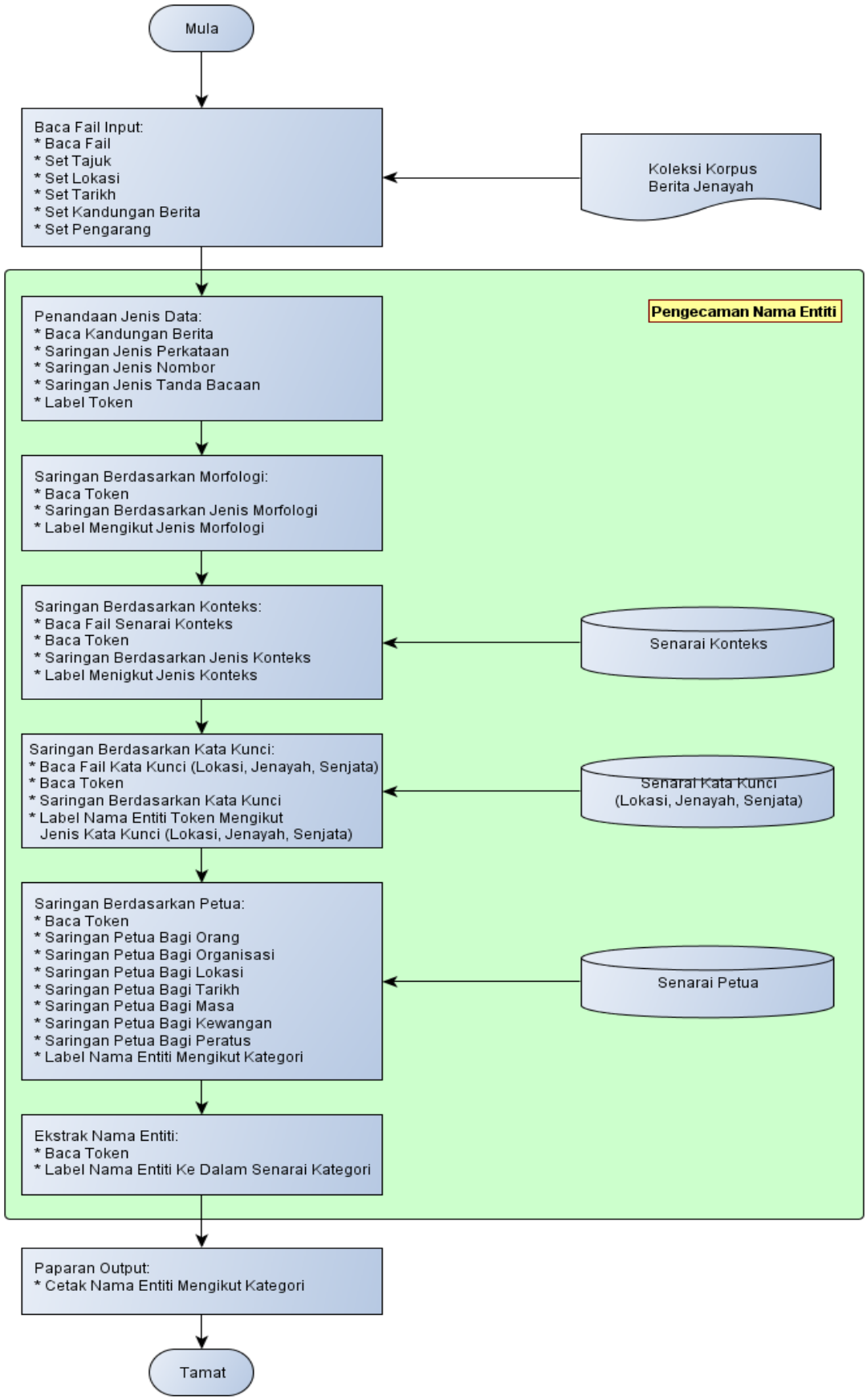
METODOLOGI KAJIAN

Methodologi yang digunakan dalam penyelidikan ini melibatkan lapan fasa (rujuk Rajah 2) iaitu fungsi baca fail input, fungsi penandaan jenis data, fungsi saringan berdasarkan jenis morfologi, fungsi saringan berdasarkan jenis konteks, fungsi saringan berdasarkan senarai kata kunci (lokasi, jenayah dan senjata), fungsi saringan berdasarkan senarai peraturan, fungsi pengekstrakan nama entiti dan penilaian mengikut kategori.

Pada fasa pertama, proses ini bertindak dengan membaca kandungan teks tidak berstruktur dari koleksi korpus berita jenayah yang telah dikumpulkan. Data diperolehi dari teks berita jenayah, laman web arkib berita BERNAMA. Sebanyak 150 teks berita jenayah telah dipilih secara rawak dan dikumpulkan bermula dari tarikh 1/5/2014 hingga 30/5/2014 untuk dijadikan korpus berita jenayah bagi pengujian ini. Pemilihan secara rawak dilakukan berdasarkan katakunci pada teks berita yang menggambarkan perbuatan jenayah yang dilakukan.

Seterusnya sistem melakukan pengecaman dan mengekstrak kandungan entiti nama seperti nama individu, organisasi, lokasi, tarikh, masa, kewangan, peratusan, jenayah dan senjata. Kandungan yang telah diekstrak dipaparkan sebagai hasil dari pengecaman sistem ini.

eISSN: 2550-2131
ISSN: 1675-8021



RAJAH 2. Carta alir terperinci sistem prototaip Pengecaman Nama Entiti

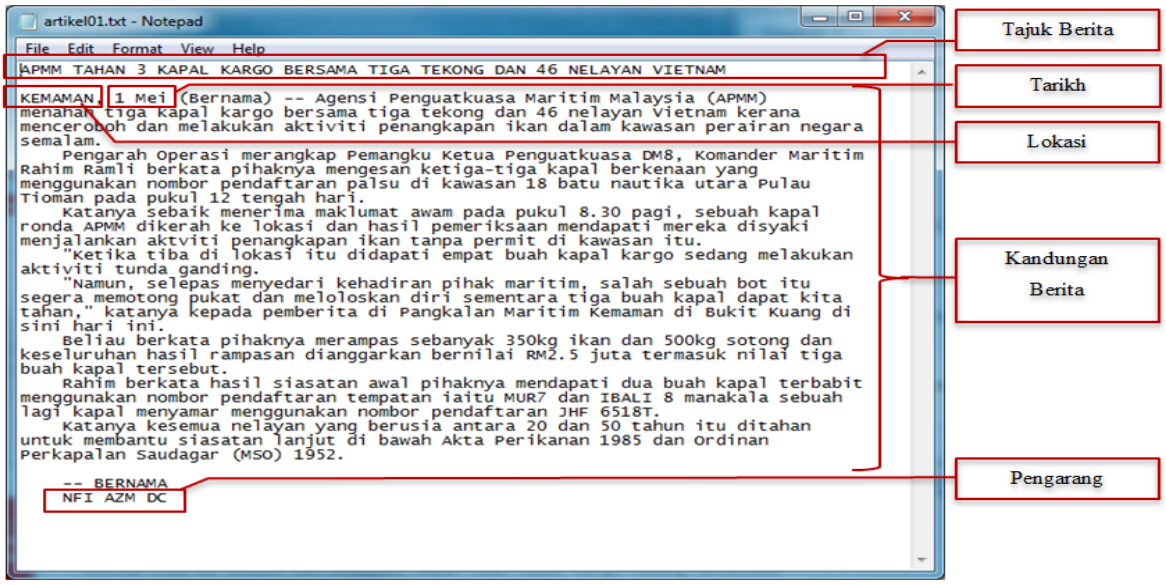
Sebagai contoh, kandungan teks berita jenayah pada Rajah 3 akan menghasilkan output seperti berikut:

Individu: *Pengarah Operasi; Komander Maritim Rahim Ramli*
Lokasi: *Pulau Tioman*

"Pengarah Operasi merangkap Pemangku Ketua Penguatkuasa DM8, Komander Maritim Rahim Ramli berkata pihaknya mengesan ketiga-tiga kapal berkenaan yang menggunakan nombor pendaftaran palsu di kawasan 18 batu nautika utara Pulau Tioman."

RAJAH 3. Contoh kandungan teks berita jenayah

Seterusnya, kandungan teks berita ini dibahagikan kepada 5 bahagian (rujuk Rajah 4) iaitu, tajuk berita, lokasi, tarikh berita dikeluarkan, pengarang dan kandungan berita. Bahagian-bahagian ini perlu diekstrak terlebih dahulu bagi tujuan mengklasifikasikan berita serta mendapatkan kandungan berita sebenar untuk diproses pada bahagian seterusnya.



RAJAH 4. Contoh teks berita jenayah beserta bahagian yang diekstrak

Proses penandaan jenis data pula adalah permulaan kepada proses pengecaman nama entiti. Fungsi penandaan jenis data ini adalah untuk mengklasifikasikan setiap kandungan teks yang telah disaring menjadi token mengikut jenis sama ada perkataan, nombor ataupun tanda bacaan. Token-token ini dilabelkan sebagai '*WORD*' bagi perkataan, '*NUM*' bagi nombor dan '*OPUNC*' bagi tanda bacaan. Memandangkan tanda bacaan juga mempunyai maksud yang tersendiri, oleh itu tanda bacaan turut diklasifikasikan kepada beberapa jenis sebagaimana yang disenaraikan pada Jadual 3 di bawah.

JADUAL 3. Senarai tanda bacaan dan jenis token

Tanda Bacaan	Jenis Token	Penerangan
.	OPUNC	Titik, noktah
,	OPUNC	Koma
:	OPUNC	Titik bertindih
;	OPUNC	Koma bertitik
$	MPUNC	Simbol matawang dolar
%	PPUNC	Simbol peratus

‘	SQPUNC	Simbol petikan
“	DQPUNC	Simbol petikan berganda
(	SPPUNC	Buka tanda kurungan (*Parentheses*)
)	EPPUNC	Tutup tanda kurungan (*Parentheses*)
[	SBPUNC	Buka tanda kurungan (*Bracket*)
]	EBPUNC	Tutup tanda kurungan (*Bracket*)
{	SCPUNC	Buka tanda kurungan (*Braces*)
}	ECPUNC	Tutup tanda kurungan (*Braces*)
<	SDPUNC	Buka tanda kurungan (*Chevrons/Diamond*)
>	EDPUNC	Tutup tanda kurungan (*Chevrons/Diamond*)

Setelah proses penandaan jenis data dilakukan, token-token ini kemudiannya disaring mengikut jenis morfologi. Berdasarkan kajian lepas, (Budi et al., 2005) telah menggariskan 11 jenis morfologi bagi perkataan. Morfologi ini kemudiannya ditambah baik dengan penambahan *MonetaryForm* dan *PercentageForm* bagi tujuan kajian ini.

Seterusnya, saringan ini mengklasifikasikan setiap token dengan melabelkan mengikut jenis kategori entiti nama sebagaimana yang disenaraikan pada Jadual 4. Setiap token boleh mempunyai lebih dari satu jenis kategori entiti nama bergantung kepada sifat perkataan tersebut. Sebagai contoh, morfologi bagi perkataan ‘*Ahmad’* boleh diklasifikasikan sebagai *TitleCase*, *MixedCase* dan juga *CapStart*.

JADUAL 4. Senarai jenis-jenis morfologi

Jenis	Penerangan	Contoh
TitleCase	Bermula dengan huruf besar dan diikuti huruf kecil	Ahmad, Malaysia
UpperCase	Kesemua huruf besar	APMM
LowerCase	Kesemua huruf kecil	menangkap
MixedCase	Campuran huruf besar dan huruf kecil	ESSZone
CapStart	Bermula dengan huruf besar	Ahmad, ESSZone
CharDigit	Huruf dan nombor	JPA3
Digit	Nombor sahaja	2014
DigitSlash	Nombor dan tanda pangkas (/)	20/9
Numeric	Nombor dengan tanda titik atau koma	1,956.25
NumStr	Nombor dalam perkataan	satu, dua, puluh
Roman	Nombor Roman	I, IV, XIX
DateForm	Nombor dalam format tarikh	15/8/2014
TimeForm	Nombor dalam format masa	11:46, 13.55
MonetaryForm	Nombor dalam format kewangan	$17.55, RM9.95
PercentageForm	Nombor dalam format peratus	85%

Proses seterusnya adalah saringan berdasarkan jenis simbol yang digunakan bagi menggambarkan entiti nama pada suatu konteks perkataan atau frasa. Perkataan-perkataan yang mempunyai maksud bagi mengenalpasti entiti nama dimasukkan ke dalam senarai konteks ini. Kemudian, perkataan-perkataan didalam senarai ini diklasifikasikan mengikut jenis konteks sebagaimana yang disenaraikan pada Jadual 5 di bawah.

JADUAL 5. Senarai konteks bagi mengambarkan entiti nama

Jenis	Penerangan	Contoh kategori konteks
PPRE	Kata awalan individu	Raja, Tengku, Nik, Syed
PMID	Kata tengah individu	bin, binti, a/l, a/p, a/k, anak
PTIT	Pangkat/Gelaran individu	haji, datuk, dato, tan sri, tun
OPOS	Pangkat dalam organisasi	tuan, ketua, pengarah, doktor, menteri, senator, tuai

OPRE	Kata awalan organisasi	agensi, lembaga, persatuan, syarikat, pertubuhan, kelab, hospital, sekolah
OSUF	Kata akhiran organisasi	sendirian berhad, sdn bhd
OCON	Kontekstual lain organisasi	muktamar
LPRE	Kata awalan lokasi	teluk, terusan, negeri, kawasan
POLP	Kata sendi kebiasaannya diikuti oleh nama individu	oleh, untuk, kepada, daripada, pada, kata
POLS	Kata akhiran kebiasaannya dimulai oleh individu	berkata
LOPP	Kata sendi kebiasaannya diikuti oleh nama lokasi	di, ke, dari, dalam
DAY	Hari	ahad, isnin, selasa, jumaat, sabtu
DSUF	Kata akhiran hari	haribulan
MONTH	Bulan	januari, september, jan, dis
TPRE	Kata awalan masa	pukul, sejak, sampai, jam
TSUF	Kata akhiran masa	pagi, malam, AM, PM, minit, saat
MPRE	Kata awalan kewangan	RM, Rp
MSUF	Kata akhiran kewangan	dinar, ringgit, sen, yen, baht, rupiah
CSUF	Kata akhiran peratusan	peratus

Berdasarkan kajian lepas, Fong et al. (2011) telah menyenaraikan 11 jenis konteks bagi penggunaan bahasa Iban, manakala Budi et al. (2005) menyenaraikan 15 jenis konteks bagi penggunaan bahasa Indonesia. Setelah dianalisi, didapati penggunaan kategori konteks pada kedua-dua bahasa ini hampir sama, sebahagian dari jenis kategori konteks ini diadaptasi dengan beberapa penambahan jenis konteks lagi bagi kesesuaian penggunaan dalam bahasa Melayu.

Penambahan ini dibuat berdasarkan rujukan dari Kamus Dewan, buku tatabahasa dan buku tesaurus bahasa Melayu. Melalui buku Tatabahasa Dewan Edisi Ketiga (Nik Safiah Karim et al. 2013), kata sendi yang digunakan dapat mengenalpasti nama individu, organisasi, lokasi dan masa. Sebagai contoh, kata sendi seperti ‘di’, ‘ke’ dan ‘dari’ merupakan kata sendi yang kebiasaannya digunakan dengan nama tempat manakala ‘daripada’ dan ‘pada’ pula kebiasannya digunakan selepas kata nama atau frasa nama (Rujuk Jadual 6).

JADUAL 6. Penggunaan Kata Sendi Dari dan Daripada

Kata Sendi	Kenyataan	Maksud
dari	Kata sendi nama 'dari' digunakan di hadapan kata nama atau frasa nama.	Ia menyatakan arah, tempat, masa atau waktu(ATM).
daripada	Kata sendi nama 'daripada' digunakan di hadapan kata nama atau frasa nama.	Ia menyatakan punca bagi manusia atau institusi, haiwan, benda dan unsur abstrak.

Selain itu, nama panggilan, gelaran atau pangkat seperti ‘Haji’, ‘Tun’ dan ‘Datuk’ turut disenaraikan dan dilabelkan sebagai gelaran atau pangkat kepada nama seseorang. Perkataan yang berhubung kait dengan tempat seperti ‘teluk’, ‘terusan’ dan ‘negeri’ pula dilabelkan sebagai kata awalan bagi lokasi.

Setelah itu, proses saringan berdasarkan senarai kata kunci iaitu lokasi, jenayah dan senjata dilakukan. Token yang telah dikenalpasti dilabelkan sebagai entiti nama mengikut kategori lokasi, jenayah atau senjata. Bagi kajian ini, kata kunci lokasi dipilih dari senarai nama negara, negeri dan daerah di Malaysia untuk dimasukkan ke dalam senarai. Ini kerana nama negara, negeri dan daerah merupakan kata nama khas dan ianya amat sesuai untuk dijadikan sebagai kata kunci.

Senarai jenis jenayah, senjata dan negara diperolehi dari kajian terdahulu yang dilakukan oleh (Darwich, 2014) manakala senarai negeri dan daerah di Malaysia diperolehi dari portal rasmi Jabatan Perangkaan Malaysia.

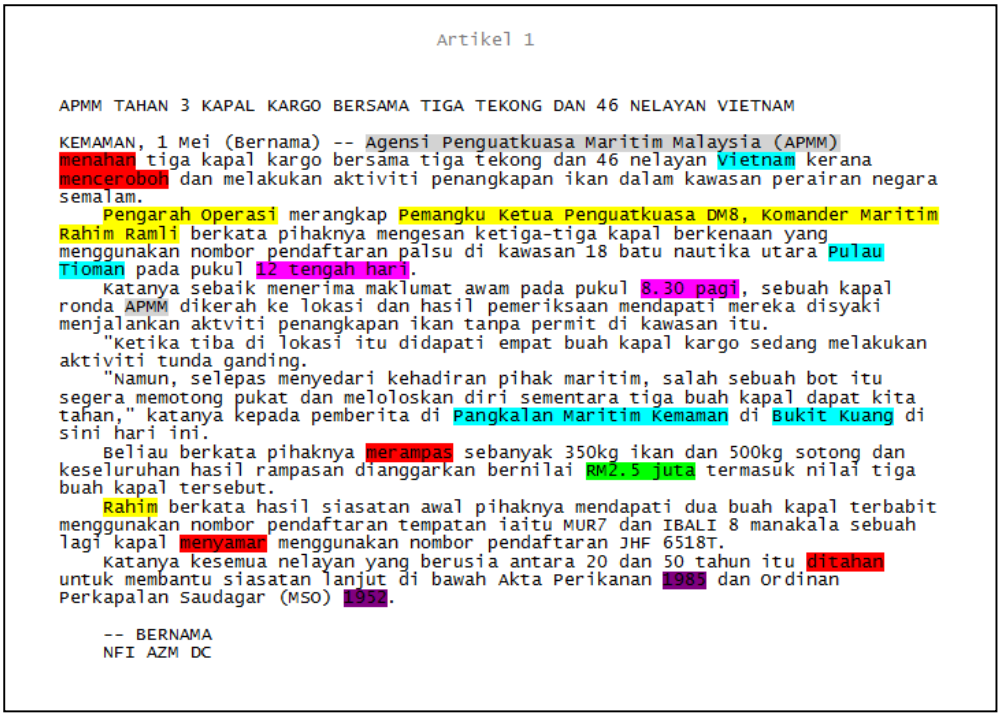
Saringan peraturan pula dibuat ke atas setiap kategori entiti nama bagi mengenalpasti token yang memenuhi kriteria entiti nama tersebut. Terdapat tujuh kategori entiti nama bagi kajian ini iaitu orang, organisasi, lokasi, tarikh, masa, kewangan dan peratus. Setelah token ini dikenalpasti, token ini kemudiannya dilabelkan mengikut kategori entiti nama tersebut. Penerangan lanjut berkenaan peraturan bagi setiap entiti nama dihuraikan pada bahagian berikut.

Setelah selesai proses penyaringan, proses terakhir adalah mengekstrak entiti nama yang telah dikenalpasti ke dalam senarai entiti nama dan melabelkannya mengikut kategori masing-masing. Pengkstrakan ini dilakukan bagi menggabungkan jujukan token sekiranya entiti nama tersebut dalam kategori yang sama. Sebagai contoh, empat token entiti nama yang berjujukan iaitu "Komander", "Maritim", "Rahim" dan "Ramli" telah dikenalpasti sebagai kategori entiti nama bagi individu. Melalui proses ini, kesemua token tersebut digabungkan menjadi satu entiti nama iaitu "Komander Maritim Rahim Ramli". Kesemua 4 token ini harus berada dalam keadaan berjujukan (sequence), untuk memastikan token ini merujuk kepada entiti yang sama.

Proses terakhir melibatkan penghasilan senarai entiti nama yang telah dikenalpasti mengikut kategori entiti nama iaitu individu, organisasi, lokasi, tarikh, masa, kewangan, peratus, jenayah dan senjata.

PENGECAMAN PAKAR BAHASA

Setelah korpus berita jenayah dikumpulkan, seterusnya korpus ini dicetak dan diteliti secara manual oleh pakar bahasa bagi mengecam entiti nama seperti individu, organisasi, lokasi, tarikh, masa, kewangan, peratusan, jenayah dan senjata. Teks berita ini kemudiannya ditanda mengikut warna tertentu bagi memudahkan pengekstrakan entiti nama dilakukan ke dalam jadual. Rajah 6 memaparkan contoh teks berita yang telah ditanda mengikut kategori klasifikasi entiti yang telah dinyatakan sebelum ini iaitu senarai jenis jenayah, senjata, negara, orang, organisasi, lokasi, tarikh, masa, kewangan dan peratus.



Artikel 1

APMM TAHAN 3 KAPAL KARGO BERSAMA TIGA TEKONG DAN 46 NELAYAN VIETNAM

KEMAMAN, 1 Mei (Bernama) -- Agensi Penguatkuasa Maritim Malaysia (APMM) menahan tiga kapal kargo bersama tiga tekong dan 46 nelayan Vietnam kerana mencerobohi dan melakukan aktiviti penangkapan ikan dalam kawasan perairan negara semalam.

Pengarah Operasi merangkap Pemangku Ketua Penguatkuasa DM8, Komander Maritim Rahim Ramli berkata pihaknya mengesan ketiga-tiga kapal berkenaan yang menggunakan nombor pendaftaran palsu di kawasan 18 batu nautika utara Pulau Tioman pada pukul 12 tengah hari.

Katanya sebaik menerima maklumat awam pada pukul 8.30 pagi, sebuah kapal ronda APMM dikerah ke lokasi dan hasil pemeriksaan mendapati mereka disyaki menjalankan aktviti penangkapan ikan tanpa permit di kawasan itu.

"Ketika tiba di lokasi itu didapati empat buah kapal kargo sedang melakukan aktiviti tunda ganding.

"Namun, selepas menyedari kehadiran pihak maritim, salah sebuah bot itu segera memotong pukat dan meloloskan diri sementara tiga buah kapal dapat kita tahan," katanya kepada pemberita di Pangkalan Maritim Kemaman di Bukit Kuang di sini hari ini.

Beliau berkata pihaknya merampas sebanyak 350kg ikan dan 500kg sotong dan keseluruhan hasil rampasan dianggarkan bernilai RM2.5 juta termasuk nilai tiga buah kapal tersebut.

Rahim berkata hasil siasatan awal pihaknya mendapati dua buah kapal terbabit menggunakan nombor pendaftaran tempatan iaitu MUR7 dan IBALI 8 manakala sebuah lagi kapal menyamar menggunakan nombor pendaftaran JHF 6518T.

Katanya kesemua nelayan yang berusia antara 20 dan 50 tahun itu ditahan untuk membantu siasatan lanjut di bawah Akta Perikanan 1985 dan Ordinan Perkapalan Saudagar (MSO) 1952.

-- BERNAMA
NFI AZM DC

RAJAH 6. Teks berita yang telah ditanda secara manual

Setelah penelitian dilakukan oleh pakar bahasa, entiti-entiti yang telah ditanda secara manual kemudiannya direkodkan ke dalam jadual pengujian mengikut kategori klasifikasi masing-masing bagi memudahkan perbandingan dilakukan dengan sistem prototaip. Rajah 7 di bawah memaparkan hasil yang diperolehi dari pengecaman oleh pakar bahasa ke atas fail teks artikel1.txt

Jadual Pengujian

Artikel1.txt	Entiti oleh Pakar Bahasa	Entiti oleh Sistem	T	S	X
Individu	Pengarah Operasi Pemangku Ketua Penguatkuasa DM8, Komander Maritim Rahim Ramli Rahim				
Organisasi	Agensi Penguatkuasa Maritim Malaysia (APMM) APMM				
Lokasi	Vietnam Pulau Tioman Pangkalan Maritim Kemaman Bukit Kuang				
Tarikh					
Masa	12 tengah hari 8.30 pagi				
Kewangan	RM2.5 juta				
Peratus					
Jenayah	menahan menceroboh tahan merampas menyamar ditahan				
Senjata					
Jumlah					

RAJAH 7. Entiti nama oleh pakar bahasa yang telah diekstrak ke dalam jadual

PENGUJIAN

Pada peringkat ini, perbandingan entiti nama antara pengecaman oleh pakar bahasa dan sistem dilakukan dan keputusan direkodkan pada bahagian kanan jadual. Terdapat tiga ruangan yang bertanda T, S dan X yang mewakili keputusan yang tepat, separa tepat dan tidak tepat. Keputusan yang tepat mempunyai entiti nama yang sama seperti pengujian yang dilakukan oleh pakar bahasa. Manakala keputusan separa tepat ialah entiti nama yang hanya mempunyai sebahagian dari perkataan yang dikenalpasti oleh pakar bahasa. Keputusan yang tidak tepat pula diberikan bagi entiti nama yang gagal dikenalpasti oleh sistem prototaip ataupun entiti nama yang salah pengecaman. Rajah 8 memaparkan jadual keputusan perbandingan pengujian yang dilakukan oleh pakar bahasa dan sistem prototaip ke atas senarai pada Rajah 7 seperti yang dibincangkan sebelum ini.

Jadual Pengujian

Artikel1.txt	Entiti oleh Pakar Bahasa	Entiti oleh Sistem	T	S	X
Individu	Pengarah Operasi Pemangku Ketua Penguatkuasa DM8, Komander Maritim Rahim Ramli Rahim	Pengarah Operasi Pemangku Ketua Penguatkuasa Komander Maritim Rahim Ramli	1	2	1
Organisasi	Agensi Penguatkuasa Maritim Malaysia (APMM) APMM	Agensi Penguatkuasa Maritim Malaysia APMM APMM MSO	1	1	2
Lokasi	Vietnam Pulau Tioman Pangkalan Maritim Kemaman Bukit Kuang	Vietnam Pulau Tioman Pangkalan Maritim Kemaman Bukit Kuang	4	0	0
Tarikh			0	0	0
Masa	12 tengah hari 8.30 pagi	12 tengah hari 8.30 pagi	2	0	0
Kewangan	RM2.5 juta	RM2.5 juta	1	0	0
Peratus			0	0	0
Jenayah	menahan menceroboh tahan merampas menyamar ditahan	menahan menceroboh tahan merampas menyamar ditahan	6	0	0
Senjata			0	0	0
Jumlah			**15**	**3**	**3**

RAJAH 8. Jadual keputusan perbandingan pengujian bagi artikel1.txt

Langkah seterusnya adalah melakukan pengiraan terhadap jadual kejituan dengan menggunakan penilaian dapatan (recall), kejituan (precision) dan F-measure sebagaimana yang dicadangkan pada persidangan pemahaman mesej (MUC) (Budi et al. 2005; Douthat 1998). Berikut merupakan definasi bagi penilaian tersebut:

a) *Recall*: bilangan pengecaman entiti nama yang tepat oleh sistem.
b) *Partial*: bilangan pengecaman entiti nama separa tepat oleh sistem.
c) *Possible*: bilangan pengecaman entiti nama yang dilakukan secara manual oleh pakar bahasa.
d) *Actual*: bilangan keseluruhan pengecaman entiti nama yang dilakukan oleh sistem termasuk pengecaman yang tepat, separa tepat dan tidak tepat.

Berdasarkan definasi di atas, penilaian keberkesanan sistem prototaip dapat dirumuskan dengan menggunakan penilaian dapatan *recall*, *precision* dan *F-measure* seperti formula berikut:

$$Recall = \frac{Correct + 0.5 * Partial}{Possible}$$

$$Precision = \frac{Correct + 0.5 * Partial}{Actual}$$

$$F - Measure = \frac{Recall * Precision}{0.5 * (Recall + Precision)}$$

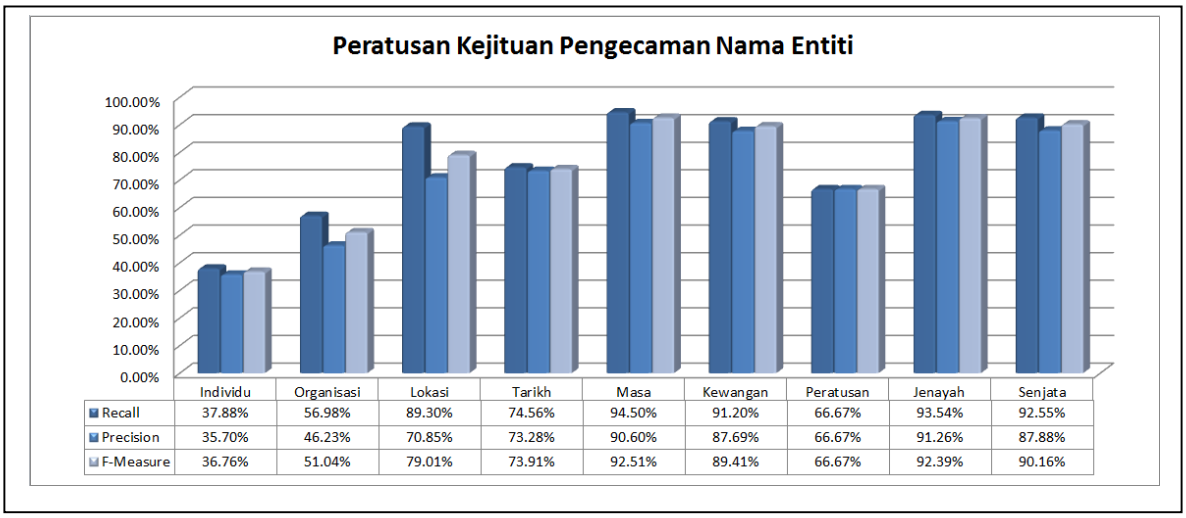
ANALISIS KEPUTUSAN

Berdasarkan keputusan pengujian yang dijalankan, nilai dapatan keseluruhan bagi sistem prototaip ini ialah 78.67% bagi recall, 71.11% bagi precision dan 74.7% bagi F-measure (Rujuk Jadual 6). Secara umumnya, sistem prototaip ini telah menghasilkan keputusan yang baik bagi pengecaman nama entiti. Namun, jika dilihat pada pecahan mengikut kategori klasifikasi nama entiti, terdapat entiti nama yang mempunyai keputusan yang rendah seperti entiti nama bagi individu. Pengecaman entiti nama bagi individu memperolehi nilai dapatan recall, precision dan F-measure adalah rendah iaitu 37.88%, 35.7% dan 36.76%. Ini kerana terdapat beberapa faktor yang menyumbang kepada kesilapan pengecaman. Sebagai contoh, kekurangan kata kunci pengenalan bagi individu seperti "Abu menjelaskan", "Ali memberitahu" dan sebagainya. Seterusnya, nama jawatan yang mempunyai rujukan silang dengan entiti nama lain seperti lokasi. Sebagai contoh perkataan "Ketua Polis Daerah Sentul" yang memisahkan "Daerah Sentul" dari entiti individu. Selain itu, terdapat juga nama yang sukar dikenalpasti melalui kata kunci yang disediakan.

JADUAL 6. Ringkasan keputusan bagi setiap kategori klasifikasi

Entiti	Tepat (T)	Separa Tepat (S)	Tidak Tepat (X)	Pakar (P)	Recall	Precision	F-Measure
Individu	119	109	258	458	37.88%	35.70%	36.76%
Organisasi	134	75	162	301	56.98%	46.23%	51.04%
Lokasi	537	44	208	626	89.30%	70.85%	79.01%
Tarikh	120	15	39	171	74.56%	73.28%	73.91%
Masa	196	3	19	209	94.50%	90.60%	92.51%
Kewangan	106	16	8	125	91.20%	87.69%	89.41%
Peratusan	2	0	1	3	66.67%	66.67%	66.67%
Jenayah	792	66	46	882	93.54%	91.26%	92.39%
Senjata	81	12	6	94	92.55%	87.88%	90.16%
Keseluruhan	**2087**	**340**	**747**	**2869**	78.67%	71.11%	74.70%

Rajah 9 di bawah memaparkan peratusan kejituan pengecaman entiti nama mengikut penilaian dapatan *recall, precision* dan *F-measure* bagi setiap kategori klasifikasi nama entiti.



	Individu	Organisasi	Lokasi	Tarikh	Masa	Kewangan	Peratusan	Jenayah	Senjata
Recall	37.88%	56.98%	89.30%	74.56%	94.50%	91.20%	66.67%	93.54%	92.55%
Precision	35.70%	46.23%	70.85%	73.28%	90.60%	87.69%	66.67%	91.26%	87.88%
F-Measure	36.76%	51.04%	79.01%	73.91%	92.51%	89.41%	66.67%	92.39%	90.16%

RAJAH 9. Peratusan kejituan pengecaman nama entiti

Bagi entiti organisasi, nilai dapatan bagi recall, precision dan F-measure yang diperolehi adalah sederhana iaitu 56.98%, 46.23% dan 51.04%. Ini adalah kerana terdapat sebilangan nama organisasi yang turut mempunyai rujukan silang dengan entiti nama lain seperti individu dan lokasi. Sebagai contoh, perkataan "Hospital Tun Aminah" yang memisahkan "Tun Aminah" dari entiti organisasi dan "Balai Polis Negeri Sabah" yang memisahkan "Negeri Sabah" dari entiti organisasi. Selain itu, kesilapan pengecaman terhadap perkataan akronim seperti "USJ" yang sebenarnya merupakan entiti bagi lokasi turut menyumbang kepada penurunan kadar kejituan ini.

Bagi entiti lokasi, nilai dapatan yang diperolehi adalah tinggi iaitu 89.3% bagi recall, 70.85% bagi precision dan 79.01% bagi F-measure. Ini adalah kerana kata kunci bagi negara, negeri dan daerah telah disenaraikan dan memudahkan proses pengecaman dilakukan. Namun begitu, terdapat sebilangan nama lokasi yang turut mempunyai entiti nama lain seperti individu iaitu "Jalan Sultan Ismail" yang memisahkan "Sultan Ismail" dari entiti lokasi. Begitu juga dengan kata kunci yang mempunyai dua perkataan seperti "Pulau Pinang", pengecaman kata kunci hanya dilakukan pada perkataan pertama sahaja iaitu "Pulau". Oleh itu, pengecaman hanya dapat dilakukan separa sahaja bagi nama lokasi yang mempunyai lebih dari satu perkataan.

Bagi entiti tarikh, nilai dapatan adalah 74.56% bagi recall, 73.28% bagi precision dan 73.91% bagi F-measure. Penurunan kadar kejituan ini terjadi kerana terdapat kesilapan pengecaman bagi kata kunci untuk hari, sebagai contoh "pada Khamis" yang telah diklasifikasikan kepada kategori entiti individu. Selain itu, terdapat kesilapan pengecaman bagi tarikh yang membawa kepada keputusan separa tepat. Sebagai contoh, sistem prototaip hanya dapat mengecam "20 April" sahaja bagi tarikh "20 April 2014".

Seterusnya bagi entiti masa, nilai dapatan yang diperolehi adalah tinggi iaitu 94.5% bagi recall, 90.6% bagi precision dan 92.51% bagi F-measure. Namun begitu, terdapat beberapa kesilapan dalam pengecaman entiti masa seperti dalam ayat "seberat 3.16 kg" yang turut dikenalpasti sebagai entiti masa.

Bagi entiti kewangan, nilai dapatan bagi recall, precision dan F-measure juga tinggi iaitu 91.2%, 87.69% dan 89.41%. Kebanyakan entiti kewangan dapat dikenalpasti dengan baik namun masih terdapat kesilapan dalam pengecaman bagi matawang asing seperti "S$170,000", "US$2,100" dan "Rial Iran 1.26 juta".

Seterusnya bagi entiti peratus, nilai dapatan yang diperolehi agak sederhana iaitu 66.67% bagi ketiga-tiga recall, precision dan F-measure. Dalam pengujian ini, hanya terdapat tiga entiti peratus yang dikenalpasti secara manual tetapi hanya dua entiti yang dikenalpasti oleh sistem prototaip. Perkataan yang gagal dikenalpasti ialah "lima peratus" yang mana nombor ditulis dalam format perkataan.

Entiti jenayah turut menunjukkan nilai dapatan yang tinggi iaitu 93.54% bagi recall, 91.26% bagi precision dan 92.39% bagi F-measure. Ini kerana pengecaman entiti nama bagi jenayah dikenalpasti melalui senarai kata kunci. Namun, senarai ini perlu ditambah dari semasa ke semasa bagi memastikan pengecaman yang lebih baik. Sebagai contoh, bagi kata dasar "curi" perlu ditambah perkataan imbuhan sama ada untuk ayat aktif ataupun ayat pasif seperti "mencuri" dan "dicuri" bagi mendapatkan pengecaman yang lebih baik. Selain itu, permasalahan yang sama seperti entiti lokasi turut terjadi bagi kata kunci yang mempunyai lebih dari satu perkataan. Sebagai contoh, perkataan "samun berkumpulan" hanya dapat dikenalpasti sebagai "samun".

Akhir sekali, entiti senjata juga menunjukkan nilai dapatan yang tinggi iaitu 92.55% bagi recall, 87.88% bagi precision dan 90.16% bagi F-measure. Entiti senjata juga dikenalpasti melalui senarai kata kunci, oleh itu penambahan kata kunci perlu dilakukan dari semasa ke semasa bagi meningkatkan lagi kejituan pengecaman oleh sistem prototaip. Sebagai contoh, perkataan "pistol Bul M5" hanya dapat dikenalpasti sebagai "pistol" sahaja.

KESIMPULAN

Terdapat beberapa kekangan yang telah dikenalpasti sewaktu melakukan kajian ini. Antara kekangan yang dihadapi semasa menjalankan kajian ini ialah sistem prototaip ini dibangunkan ke atas korpus berita jenayah bahasa Melayu sahaja. Oleh itu, teks berita selain dari bahasa Melayu tidak dapat dijalankan oleh sistem prototaip ini.

Kekangan seterusnya berkaitan dengan saringan berdasarkan kata kunci yang mempunyai lebih dari satu perkataan. Sebagai contoh, kata kunci "pecah rumah" dan "Pulau Pinang" hanya dapat disaring pada perkataan pertama sahaja iaitu "pecah" dan "Pulau". Oleh itu, keputusan yang diperolehi adalah separa tepat jika dibandingkan dengan pengecaman keseluruhan perkataan yang menghasilkan keputusan tepat.

Kekangan yang terakhir adalah melibatkan pengecaman terhadap satu entiti yang mempunyai perkataan entiti yang lain. Sebagai contoh, entiti tersebut merupakan entiti individu atau entiti organisasi yang mempunyai nama lokasi seperti "Ketua Polis Daerah Kota Bharu" ataupun "Hospital Kuala Lumpur". Begitu juga dengan entiti lokasi yang mempunyai nama individu seperti "Jalan Sultan Ismail". Selain itu, pengujian ini turut mengklasifikasikan negara sebagai lokasi walaupun terdapat di dalam ayat tersebut menerangkan negara sebagai warga kepada individu seperti "warga Iran".

Kajian yang dijalankan ini merupakan kajian peringkat awal yang dilakukan dalam domain jenayah bahasa Melayu. Dari kajian ini, terdapat beberapa kekangan dan permasalahan yang telah dikenalpasti sebagaimana yang dinyatakan pada bahagian sebelum ini. Kekangan dan permasalahan ini boleh diatasi dan diperbaiki menerusi cadangan bagi kajian-kajian yang akan datang.

Cadangan pertama ialah penambahan ciri saringan perkataan seperti tanda pengklasifikasian perkataan (POS) bagi domain jenayah bahasa Melayu sebagaimana yang telah dilakukan oleh sistem prototaip InNER oleh Indra Budi dan RPOS oleh Rayner Alfred (Alfred, Mujat & Obit, 2013; Budi et al., 2005). Begitu juga dengan penambahbaikan teknik pengekestrakan maklumat dengan menggunakan teknik co-reference sebagaimana yang dilakukan oleh Mohammad Darwich (Darwich, 2014).

Cadangan seterusnya ialah pemurnian proses pengecaman entiti nama agar tidak berlaku rujukan silang di antara entiti seperti "Ketua Polis Negeri Selangor" yang memisahkan "Ketua Polis" sebagai entiti individu dan "Negeri Selangor" sebagai entiti lokasi. Begitu juga dengan keupayaan untuk mengenalpasti warganegara di dalam teks berita (Darwich 2014).

Namun tujuan utama kajian ini dijalankan adalah untuk membantu para penyelidik serta para penyiasat menyelesaikan kes-kes jenayah yang kian bertambah dengan melakukan pemprosesan maklumat ke atas dokumen-dokumen Bahasa Melayu serta laporan polis yang diterima dengan lebih cepat dan berkesan.

Melalui kajian ini, entiti-entiti yang telah diekstrak mengikut kategori klasifikasi dapat dimasukkan ke dalam pangkalan data untuk dijadikan sebagai kata kunci bagi carian lanjut. Ini akan membolehkan proses capaian maklumat dilakukan dengan lebih pantas berbanding dengan carian yang dilakukan secara menyeluruh di dalam teks.

Selain itu, hasil dari kajian ini turut dapat menyumbang pengetahuan mengenai keberkesanan teknik pengecaman entiti nama bagi berita jenayah bahasa Melayu. Dengan ini, dapat membantu para penyelidik dalam bidang Pengekstrakan Maklumat (IE) untuk memperkembangkan lagi penyelidikan dalam domain jenayah bahasa Melayu pada masa hadapan. Seterusnya dapat membantu para penyelidik, polis, peguam serta pihak berkuasa yang terlibat dalam bidang jenayah menyelesaikan jenayah dengan lebih berkesan.

RUJUKAN

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

Alfred, R., Leong, L., On, C. & Anthony, P. (2013). A Rule-Based Named-Entity Recognition for Malay Articles. *Advanced Data Mining.* 288-299.

Alfred, R., Mujat, A. & Obit, J. (2013). A ruled-based part of speech (RPOS) tagger for malay text articles. Asian Conference on Intelligent Information and Database Systems, LNCS. 7803, 50-59.

Alruily, M., Ayesh, A. & Zedan, H. (2009). Crime Type Document Classification from Arabic Corpus. 2009 Second International Conference on Developments in eSystems Engineering, 153–159. doi:10.1109/DeSE.2009.50

Amin M.B., Rahim M.K. & Ayu G.M.S. (2014). A Trend Analysis of Violent Crimes in Malaysia. *Health and the Environment Journal. Vol. 5*(2), 41-56.

Asharef, M. M. A. A. (2012). Rule Base Arabic Named Entity Recognition For Crime Documents. Unpublished Master Thesis, Universiti Kebangsaan Malaysia, Bangi, Malaysia

Bikel, D. M., Miller, S., Schwartz, R. & Weischedel, R. (1997). Nymble: A high-performance learning name-finder. Proceedings of the Fifth Conference on Applied Natural Language Processing, 194–201.

Budi, I.,Bressan, S.,Wahyudi, G.,Hasibuan, Z.A. & Nazief, B.A.A. (2005). Named Entity Recognition for the Indonesian language: Combining contextual, morphological and part-of-speech features into a knowledge engineering approach. Proceedings of the 8th international conference on Discovery Science. LNCS 3735 LNAI : 57-69.

Chau, M., Xu, J. & Chen, H. (2002). Extracting Meaningful Entities from Police Narrative Reports. Proceedings of the 2002 annual national conference on Digital government research, 1–5.

Croft, W. B., Metzler, D. & Strohman, T. (2010). *Search Engines Information Retrieval in Practice*. Pearson.

Cunningham, H. (2005). Information extraction, automatic. *Encyclopedia of Language and Linguistics*.

Darwich, M. (2014). Probabilistic Reference To Suspect Or Victim In Nationality Extraction From Unstructured Crime News Documents. Unpublished Master Thesis, Universiti Kebangsaan Malaysia, Bangi, Malaysia.

Douthat, A. (1998). The Message Understanding Conference Scoring Software User's Manual. 7th Message Understanding Conference (MUC-7).

Eikvil, L. (1999). Information Extraction from the World Wide Web: A Survey. Norwegian Computer Center, Report no. 945, July 1999.

Esmaail, N. F. M. (2012). Arabic Named Entity Recognition Using Neural Network. Unpublished Master Thesis, Universiti Kebangsaan Malaysia, Bangi, Malaysia.

Faizah Md Latif. (2015). Ke Arah Pengurangan Indeks Jenayah Jalanan di Pusat Bandar Kuala Lumpur. *GEOGRAFIA Malaysian Journal of Society and Space. Vol. 11*(4), 97-107.

Fong, Y., Ranaivo-Malançon, B. & Wee, A. (2011). NERSIL-the Named-Entity Recognition System for Iban Language. 25th Pacific Asia Conference on Language, Information and Computation, pages 549–558.

Fukuda, K., Tsunoda, T., Tamura, A. & Takagi, T. (1998). Toward Information Extraction: Identifying protein names from biological papers. Pacific Symposium on Biocomputing PSB'98, 707–718.

Grishman, R. & Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. Proceedings of the 16th International Conference on Computational Linguistics, 466–471.

Hadi, S. binti A. (2011). Pendekatan Pengecaman Nama Entiti Bagi Capaian Berita Berbahasa Inggeris di Malaysia. Unpublished Master Thesis, Universiti Kebangsaan Malaysia, Bangi, Malaysia

Humphreys, K., Demetriou, G. & Gaizauskas, R. (2000). Two Applications of Information Extraction to Biological Science Journal Articles. Enzyme Interactions and Protein Structures. Pacific Symposium on Biocomputing, 5, 502–513.

Ishak, S. & Bani, Y. (2017). Determinants of Crime in Malaysia: Evidence from Developed States. *Int. Journal of Economics and Management. Vol. 11,* 607-622.

Kamus Dewan Edisi Ketiga. (2002). Dewan Bahasa dan Pustaka.

Ku, C. H., Iriberri, A. & Leroy, G. (2008). Crime Information Extraction from Police and Witness Narrative Reports. 2008 IEEE Conference on Technologies for Homeland Security, 193–198. doi:10.1109/THS.2008.4534448

Liddy, E. D. (2001). Natural Language Processing. *Encyclopedia of Library and Information Science*, hlm.2nd Ed. New York, New York, USA: Marcel Decker, Inc.

Masnizah Mohd, Wan Fariza Paizi@Fauzi, Amri Jasin. (2018). Teknik Pengukuhan Perangkak Tumpuan melalui Modul Pengesan Bahasa bagi Capaian Web Bahasa Melayu. *GEMA Online® Journal of Language Studies. Vol. 18*(3).

Mohammed, N. F. & Omar, N. (2012). Arabic Named Entity Recognition Using Artificial Neural Network. *Journal of Computer Science. Vol. 8*(8), 1285-1293.

Nik Safiah Karim, Onn, F. M., Musa, H. H. & Mahmood, A. H. (2013). *Tatabahasa Dewan Edisi Ketiga*. Dewan Bahasa dan Pustaka.

Proux, D., Rechenmann, F., Julliard, L., Pillet, V. & Jacq, B. (1998). Detecting Gene Symbols and Names in Biological Texts : A First Step toward Pertinent Information Extraction. *Genome Informatics. Vol. 9,* 72-80.

Rindflesch, T. C., Rajan, J. V. & Hunter, L. (2000). Extracting Molecular Binding Relationships from Biomedical Text. ANLC '00 Proceedings of the sixth conference on Applied natural language processing, 188–195

Sari, Y., Hassan, M. F. & Zamin, N. (2010). Rule-based pattern extractor and named entity recognition: A hybrid approach. 2010 International Symposium on Information Technology, 563–568. doi:10.1109/ITSIM.2010.5561392

Sekine, S., Grishman, R. & Shinnou, H. (1998). A Decision Tree Method for Finding and Classifying Names in Japanese Texts. Proceedings of the Sixth Workshop on Very Large Corpora, Montreal, Canada.

Shahrul Azman Mohd Noah, Nazlena Mohamad Ali, Mohd Sabri Hasan. (2018). Penentuan Fitur bagi Pengekstrakan Tajuk Berita Akhbar Bahasa Melayu. *GEMA Online® Journal of Language Studies. Vol. 18*(2).

Sommerville, I. (2011). *Software Engineering*. 9th Edition. Pearson.

Tang, C. (2009). The linkages among inflation, unemployment and crime rates in Malaysia. *International Journal of Economics and Management. Vol. 3*(1), 50-61.

PENULIS

Saidah Saad mendapat PhD daripada Universiti Teknologi Malaysia. Merupakan Pensyarah Kanan di Fakulti Teknologi dan Sains Maklumat UKM. Fokus bidang penyelidikan beliau ialah teknologi web semantik dan ontologi merangkumi perwakilan maklumat, pemprosesan bahasa tabii dan capaian maklumat.

Mohamed Kamil Mansor memperoleh Sarjana Teknologi Maklumat (Sains Maklumat) daripada Fakulti Teknologi dan Sains Maklumat, UKM pada tahun 2014. Bidang penyelidikan beliau adalah capaian maklumat bagi Bahasa Melayu yang memfokus kepada berita-berita jenayah dan sekarang beliau bertugas sebagai Lead Engineer di System Consultancy Services Sdn Bhd.

eISSN: 2550-2131
ISSN: 1675-8021