

PERBANDINGAN KONSEP PENGEKSTRAKAN BERDASARKAN GLOSARI
ISTILAH ISLAMIK

NURFATIN ATHIRAH BINTI RAHMAT

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT MEMPEROLEH IJAZAH SARJANA TEKNOLOGI
MAKLUMAT

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

PENGAKUAN

Saya mengakui bahawa karya penulisan ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya.

26 Februari 2018

NURFATIN ATHIRAH BINTI RAHMAT
GP04132

PENGHARGAAN

Syukur Alhamdulillah ke hadrat Illahi kerana limpah kurnia dan ehsan darinya telah memberikan saya masa, kesihatan dan keterangan hati untuk menyiapkan projek ini.

Jutaan terima kasih juga kepada penyelia projek saya iaitu yang berbahagia Dr. Saidah Saad yang telah memberikan banyak buah fikiran, bimbingan serta teguran yang membina sepanjang tempoh projek ini dijalankan. Tidak lupa juga jasa baik penyelaras program Prof. Madya Dr Kamsuriah Ahmad yang memberikan peluang untuk tambah-baik projek ini.

Ucapan terima kasih yang tidak terhingga kepada ahli keluarga yang tercinta khususnya ibu saya Hajjah Norzaili binti Ahmad dan ayah saya Haji Rahmat bin Rosdi kerana memberikan sokongan yang tidak berbelah bahagi dari segi sokongan emosi dan kewangan. Terima kasih kerana tidak pernah berputus asa dengan diri ini.

Ucapan penghargaan juga ditujukan kepada bekas majikan terdahulu dan sekarang kerana memahami peranan saya sebagai pelajar separuh masa dan juga memberikan ruang dan masa bilamana saya memerlukannya.

Buat rakan-rakan seperjuangan, terima kasih kerana selalu ada di saat diri ini memerlukan bantuan dari segi akademik dan saling sama menyokong memberi semangat di sepanjang pengajian ini. Semoga Allah mengurniakan kebahagiaan buat kalian. Amin.

ABSTRAK

Pengekstrakan maklumat merupakan satu proses bagi mengekstrak maklumat dari dokumen yang tidak berstruktur. Dokumen tidak berstruktur yang digunakan dalam kajian ini sebagai dataset ialah Glosari Istilah Islam yang telah dibangunkan oleh kumpulan DEED (*Dependable Entrepreneurial Engineering Division*) dari Universiti Antarabangsa Islam Malaysia (UIAM). Glosari merupakan senarai istilah berserta takrifnya yang disusun mengikut abjad dalam sesuatu bidang pengkhususan. Kajian ini menjalankan perbandingan antara dua sistem pengekstrakan maklumat yang menggunakan teknik pengecaman entiti nama untuk mengekstrak data. Berdasarkan keputusan keseluruhan pengujian bagi kedua-dua sistem, didapati kedua-dua sistem nilai dapatan keseluruhan bagi sistem pengekstrakan *Stanford CoreNLP* ialah 65% bagi dapatan, 61% bagi kejituan dan 62% bagi *F-Measure* manakala untuk keputusan pengujian sistem pengektrakan GATE ialah 63% bagi dapatan, 59% bagi kejituan dan 70% bagi *F-Measure*. Secara keseluruhannya, keputusan untuk sistem pengekstrakan berdasarkan *Stanford CoreNLP* telah menghasilkan keputusan yang baik. Hasil dari kajian ini diharapkan dapat memberi maklumbalas tentang pendekatan terbaik dalam menghasilkan senarai istilah yang lebih tepat dan betul.

COMPARISON OF CONCEPT EXTRACTION USING ISLAMIC TERM GLOSSARY

ABSTRACT

Information extraction (IE) is a process of extracting information or data from unstructured documents. Unstructured document is Islamic Term Glossary which is the dataset for the project and it has been developed by DEED (Dependable Entrepreneurial Engineering Division) group from International Islamic University Malaysia (IIUM). Glossary can be defined as a list of terms which contains meanings and arranged alphabetically order and it is based on certain concept or area. This project is particularly comparing two extraction systems technique which is named-entity recognition. Based on the result obtained from both systems, it can be concluded that recall, precision and F-Measure value for extraction system of Stanford CoreNLP is 65, 61% and 62% simultaneously, whereas recall, precision and F-Measure value for extraction system of GATE is 63%, 59% and 70% simultaneously. Overall, extraction system of Stanford CoreNLP has achieved good result. Hopefully, the result obtained from this project would give feedback on which approach is better to produce accurate and precise term list.

KANDUNGAN

		Halaman
PENGAKUAN		ii
PENGHARGAAN		iii
ABSTRAK		iv
ABSTRACT		v
KANDUNGAN		vi
SENARAI JADUAL		ix
SENARAI ILUSTRASI		x
SENARAI SINGKATAN		xi
BAB I	PENDAHULUAN	2
1.1	Pengenalan	2
1.2	Latar Belakang Kajian	3
1.3	Penyataan Masalah	5
1.4	Objektif Kajian	6
1.5	Skop Kajian	6
1.6	Metodologi Kajian	8
1.7	Organisasi Kajian	9
BAB II	KAJIAN LITERASI	11
2.1	Pengenalan	11
2.2	Pemprosesan Bahasa Tabii (NLP)	11
2.3	Pengeskrakan Maklumat (IE)	13
2.4	Pengecaman Entiti Nama	15
2.5	Pendekatan Pengecaman Entiti Nama	15
	2.5.1 Pendekatan Berdasarkan Peraturan	18
	2.5.2 Pendekatan Berdasarkan Statistik	28
	2.5.3 Perbandingan Pendekatan Berdasarkan Peraturan dan Statistik	19
2.6	<i>Stanford CoreNLP</i>	21
2.7	<i>GATE (General Architecture For Text Engineering)</i>	2

2.8	Kajian Lepas	24
BAB III	METODOLOGI KAJIAN	29
3.1	Pengenalan	29
3.2	Fasa Pengumpulan Maklumat	29
3.3	Fasa Pembangunan Teknik Pregekstrakan	31
	3.3.1 Teknik Pengekstrakan Bagi <i>Stanford CoreNLP</i>	38
	3.3.2 Teknik Pengekstrakan Bagi GATE	41
3.4	Fasa Definasi Fungsi	45
3.5	Fasa Dokumentasi	45
3.6	Kesimpulan	46
BAB IV	IMPLEMENTASI DAN PENGUJIAN	47
4.1	Pengenalan	47
4.2	Keperluan Pembangunan Sistem	47
4.3	Perbandingan Persediaan Data	48
4.4	Sistem Pengekstrakan GATE	49
4.5	Sistem Pengekstrakan Stanford CoreNLP	53
	4.5.1 Pembangunan Sistem Stanford CoreNLP	55
4.6	Pengujian	57
4.7	Analisis Pengujian	61
4.8	Kesimpulan	64
BAB V	PERBINCANGAN DAN KESIMPULAN	65
5.1	Pengenalan	65
5.2	Rumusan Dan Penemuan Kajian	65
5.3	Sumbangan Kajian	66
5.4	Kekangan Kajian	67
5.5	Cadangan Kajian Lanjutan	68
5.6	Penutup	69
RUJUKAN		70

SENARAI JADUAL

No.Jadual		Halaman
Jadual 2.1	Perbandingan antara pendekatan peraturan dan statistik	19
Jadual 2.2	<i>POS Tagging Treebank</i>	20
Jadual 2.3	Perbandingan kajian-kajian lepas	27
Jadual 3.1	Input kepada Sistem Pengekstrakan	34
Jadual 4.1	Senarai keperluan perisian dan perkakasan	36
Jadual 4.2	Sampel data yang telah diesktrak	48
Jadual 4.3	Data dari Glosari Istilah Islamik	51
Jadual 4.4	Fungsi bagi sistem pengecaman entiti nama	55
Jadual 4.5	Jadual pengujian <i>Stanford CoreNLP</i>	56
Jadual 4.6	Jadual pengujian GATE	58
Jadual 4.7	Sebahagian data untuk kejituan (<i>Stanford CoreNLP</i>)	58
Jadual 4.8	Sebahagian data untuk kejituan (<i>GATE</i>)	59
Jadual 4.9	Keputusan pengujian sistem <i>Stanford CoreNLP</i>	60
Jadual 4.10	Keputusan pengujian sistem GATE	61
		62

SENARAI ILUSTRASI

No.Rajah		Halaman
Rajah 1.1	Ringkasan Metodologi Kajian	8
Rajah 2.1	<i>Stanford CoreNLP</i>	22
Rajah 2.2	GATE	23
Rajah 3.1	Rangka kajian	35
Rajah 3.2	Perbandingan persediaan data	36
Rajah 3.3	Proses penyalinan glosari	37
Rajah 3.4	Persediaan menginput data <i>Stanford CoreNLP</i>	42
Rajah 3.5	Persediaan menginput data GATE	43
Rajah 4.1	<i>Interface GATE</i>	50
Rajah 4.2	Senarai anotasi GATE	50
Rajah 4.3	Contoh output GATE	52
Rajah 4.4	Output sistem pengekstrakan <i>Stanford CoreNLP</i>	54

SENARAI SINGKATAN

CO	<i>Co-Reference Identifcation or Resolution</i>
GB	<i>Giga Byte</i>
IDE	<i>Integrated Development Environment</i>
IE	<i>Information Extraction</i> (Pengekstrakan Maklumat)
IR	<i>Information Retrieval</i> (Capaian Maklumat)
MUC	<i>Message Understanding Conference</i> (Persidangan Pemahaman Mesej)
NE	<i>Named Entity</i> (Entiti Nama)
NER	<i>Named Entity Recognition</i> (Pengecaman Entiti Nama)
NLP	<i>Natural Language Processing</i> (Pemprosesan Bahasa Tabii)
POS	<i>Part Of Speech</i> (Penandaan Golongan Kata)
HTML	<i>Hypertext Markup Language</i>
RAD	<i>Rapid Application Development</i> (Pembangunan Aplikasi Pantas)
RAM	<i>Random Access Memory</i>
SDK	<i>Software Development Kit</i>

BAB I

PENDAHULUAN

1.1 PENGENALAN

Menurut Dewan Pustaka dan Bahasa, maklumat ditakrifkan sebagai keterangan, perincian, berita, khabar, mesej, taklimat, laporan, fakta, data, dokumen dan pengetahuan. Dewasa ini, begitu banyak maklumat yang wujud di persekitaran kita melibatkan maklumat yang sah dan sebaliknya. Dengan kepesatan dan keghairahan teknologi dan juga manusia dalam mencipta kebenaran maklumat dan relevan sesuatu maklumat haruslah dititikberatkan. Oleh itu, pengekstrakan maklumat yang tepat dan jitu merupakan sesuatu yang sukar dilakukan oleh manusia mahupun sistem komputer.

Pengekstrakan maklumat ataupun *information extraction* (IE) merupakan salah satu bidang penting dalam kumpulan Sains Maklumat selain daripada Capaian Maklumat (IR), Ontologi, Pemprosesan Bahasa Tabii (NLP) dan lain-lain lagi. Mereka saling berkaitan untuk menghasilkan trend carian maklumat yang optimum. Menurut Alfred, terdapat tiga jenis dokumen yang digunakan sebagai data *input* iaitu dokumen berstruktur seperti pengkalan data, dokumen separa struktur seperti fail HTML dan dokumen yang tidak berstruktur seperti dokumen teks berita. Kewujudan kemudahan seperti inilah yang membenarkan komunikasi secara digital

antara manusia dan dunia digital. Manusia boleh menerima dan mengubah serta menyimpan data atau informasi secara senang dan berhemah.

Pengekstrakan maklumat (IE) merupakan satu proses bagi mengekstrak maklumat dari dokumen yang tidak berstruktur. Salah satu contoh dokumen yang tidak berstruktur yang bakal digunakan sebagai dataset mahupun domain ialah glosari istilah Islamik yang telah dibangunkan oleh kumpulan DEED (*Dependable Entrepreneurial Engineering Division*) dari Universiti Antarabangsa Islam Malaysia (UIAM). Pembangunan projek glosari istilah islamik ini merupakan projek yang telah diuruskan oleh Bahagian Teknologi Maklumat, Fakulti Kejuruteraan UIAM. Pada dasarnya glosari merupakan senarai istilah berserta takrifnya yang disusun mengikut abjad dan sesuatu bidang pengkhususan.

Namun ianya masih menimbulkan kesukaran ketika pengguna mahu mencari atau mengetahui sesuatu terma tersebut. Pengguna sedikit sebanyak haruslah mempunyai latar belakang pengetahuan tentang sesuatu topik sebelum memulakan pencarian maklumat. Oleh itu, pelbagai konsep pengesktrakan maklumat telah diperkenalkan dan dibangunkan untuk memudahkan pencarian maklumat dilakukan secara optimum dan maklumat dapat dipaparkan seterusnya dapat diguna oleh pengguna. Tetapi masih begitu banyak maklumat serta pengetahuan manusia yang tersirat dan tersurat dan masih lagi tidak berkemampuan untuk sistem mentafsirkannya.

Kebiasaannya, glosari ataupun kamus mempunyai persamaan di mana maklumat disusun mengikut abjad dan disertakan dengan terma serta maksudnya. Ini mempunyai perkaitan dari segi konsep dan hubungan antara mereka. Menurut Asuncion, prestasi penggunaan kamus berdasarkan kamus boleh-baca mesin bagi mengekstrak konsep dan hubungan yang relevan adalah berdasarkan pembelajaran ontologi dimana ontologi merupakan salah satu cara juga untuk menghasilkan pengeluaran dari sistem pengekstrakan.

Sehubungan dengan itu, wujudnya konsep pengekstrakan menggunakan perisian sumber terbuka *Stanford Parser* dan GATE yang secara amnya dibangunkan khas untuk mengesktrak pengetahuan yang terkandung dalam korpus maklumat yang semestinya besar ini. Dengan lebih tepat lagi, projek ini akan menerangkan lebih lanjut tentang perbandingan konsep pengekstrakan menggunakan GATE (*General Architecture for Text Engineering*) dan *Stanford CoreNLP*.

1.1 LATAR BELAKANG KAJIAN

Letupan maklumat di Internet dalam kuantiti yang banyak menjadikan pencarian dan pemilihan terhadap maklumat yang berkaitan menjadi kritikal dan amat penting. Kaedah untuk mencari dokumen atau maklumat yang berkaitan menjadi isu yang sentiasa dititikberatkan. Oleh itu, penggunaan konsep pengekstrakan maklumat bagi menstrukturkan kembali maklumat perlu dilaksanakan supaya maklumat ini dapat dilihat sebagai informasi yang berguna.

Terdapat banyak cara untuk mengenalpasti pengekstrakan maklumat antaranya ialah pengecaman entiti nama (NER), menerusi ontologi, pakar domain dan lain-lain. Tetapi untuk projek ini, konsep pengekstrakan maklumat akan diberikan penekanan. Konsep pengekstrakan maklumat itu sendiri merangkumi aspek-aspek seperti penandaan golongan kata (*part-of-speech tagging*), penghurai (*parser*), pengecaman entiti nama (*named entity recognition*), pengtokenan, *coreference* turut digunakan sepanjang kajian ini dijalankan. Selain itu, terma-terma yang diguna pakai dalam pengecaman entiti nama juga akan diketengahkan.

Tujuan utama pengekstrakan maklumat ini dijalankan adalah kerana untuk mendapatkan konsep yang relevan bagi sesuatu terma. Dengan menggunakan NER atau aspek-aspek lain, pengguna akan mengenal pasti kata nama ataupun terma mengikut klasifikasi dan konsep yang ditetapkan.

Hal ini kerana, menurut glosari yang sedia ada, penerangan yang diberikan untuk setiap terma terlalu panjang dan ada yang terlalu pendek. Oleh itu, konsep pengekstrakan maklumat membantu untuk mengekstrak sesuatu yang relevan untuk ketepatan maklumat dan pengetahuan seseorang individu itu.

Penggunaan terma-terma dalam glosari mahupun makna-makna yang telah tersedia kebanyakannya menggunakan klasifikasi mengikut kategori individu, organisasi, lokasi, peristiwa, tarikh atau masa. Sebagai contoh, nama peristiwa Lailatul Qadar, manakala lokasi Al-Qadr iaitu nama surah untuk mendapatkan lebih banyak informasi dan maklumat yang berkaitan dengan malam Lailatul Qadar.

Proses pengekstrakan maklumat berdasarkan Glosari Istilah Islamik turut bergantung kepada pengkhususan domain atau subjek untuk bahasa tertentu. Sebagai contoh menurut Glosari Istilah Islamik bahasa yang diguna pakai ialah Bahasa Inggeris dan mudah untuk dikenalpasti dengan penggunaan kata nama atau kata kerja untuk mendefinisikan sesuatu terma terbabit.

Terdapat dua badan penyelidikan yang aktif dalam bidang ini iaitu *Message Understanding Conference (MUC)* dan program *Automatic Content Extraction (ACE)* (Cunningham 2006; Darwich 2014). Kedua-dua badan ini telah menganjurkan persidangan dan menggariskan panduan bagi proses pengekstrakan maklumat. Pengecaman entiti nama boleh dibangunkan melalui tiga teknik yang diperkenalkan iaitu pendekatan berasaskan peraturan, pendekatan berasaskan statistic dan juga gabungan kedua-dua pendekatan ini (Alfred et al.2014).

Pengecaman dan pengklasifikasian pengecaman entiti nama turut bergantung kepada pengkhususan domain dan juga nahu bahasa tertentu. Sebagai contoh, domain islamik memerlukan kata kunci khusus bagi mengenalpasti perseorangan yang berbeza sama ada Muhammad, Daud dan lain-lain. Begitu juga dengan nahu bahasa, setiap bahasa mempunyai sifat dan tatabahasa yang berbeza untuk melakukan pengecaman. Sebagai contoh, kata nama bahasa universal iaitu Bahasa Inggeris adalah mudah dikenalpasti dengan pengenalan huruf besar di awalan perkataan.

Penyelidikan pengecaman entiti nama turut dilakukan berdasarkan domain yang berbeza-beza. Sebagai contoh, antara domain yang dilakukan ialah domain perubatan dan juga domain jenayah. Ini kerana pegawai perubatan mahupun pegawai penguatkuasa serta pegawai penyelidik memerlukan maklumat penting berdasarkan sesuatu maklumat yang tersimpan dalam mana-mana dokumen. Capaian yang dilakukan secara manual terhadap dokumen-dokumen ini mengambil masa yang lama dan tidak praktikal untuk memproses satu kumpulan yang banyak. Para

penyelidik seperti yang disebutkan diatas memerlukan suatu sistem untuk memproses semua maklumat secara pantas dan tuntas. .

1.2 PENYATAAN MASALAH

Sejak kebelakangan ini, ledakan maklumat telah berlaku secara berleluasa dan ini menyebabkan berlaku lambakan maklumat. Maklumat ini perlu diurus dan ditakbir bagi memudah capaian maklumat termasuk maklumat berkaitan pengetahuan Islam. Oleh itu, dengan bantuan Glosari Istilah Islamik, pengguna akan lebih peka dan akan menggunakan penerangan untuk memperbetulkan atau menyebarkan sebarang maklumat yang bersifat islamik. Oleh itu, pengurusan data bagi membolehkan pemprosesan dan pengekstrakan maklumat berlaku dengan cepat dan berkesan.

Selain itu, terma-terma yang mempunyai maksud lain juga menimbulkan kekeliruan sewaktu pencarian dilakukan. Oleh itu, penggunaan konsep pengekstrakan maklumat sedikit sebanyak membantu dalam melakukan pencarian yang tepat kerana penggunaan konsep serta hubungan yang jelas maksudnya dan penerangan yang mudah difahami. Ini kerana terma-terma dalam glosari ini merupakan sesuatu yang menyakinkan kerana ianya adalah isi maklumat dari dalam Al-Quran itu sendiri. Pengekstrak ini adalah bagi memudahkan mesin memahami akan maklumat yang akan diproses.

Pengecaman entiti nama merupakan salah satu teknik pengekstarakan 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 pengecaman entiti nama adalah untuk mengklasifikasikan entiti nama seperti individu, organisasi, lokasi dan lain-lain lagi.

Pada kebiasaannya, teknik pengecaman entiti nama dipengaruhi oleh domain yang dikaji. Setiap domain memerlukan pengecaman yang tersendiri bagi memperoleh keputusan yang baik. Sama juga dengan penggunaan bahasa bagi setiap domain, bahasa mempunyai sifat dan tatabahasa yang unik dan tersendiri (Alfred et al.2014). Ini kerana pengecaman eniti nama melibatkan proses pengecaman

perkataan seperti morfologi, penandaan golongan kata, dan klasifikasi tesaurus serta penggunaan senarai kata kunci atau kamus.

Dalam pembangunan projek ini, perbandingan antara konsep pengekstrakan maklumat berdasarkan glosari akan dilakukan menggunakan perisian Stanford Parser dan juga GATE. Atas dasar itu juga, pernyataan masalah yang boleh dimajukan ialah:

- a) Adakah konsep pengekstrakan menggunakan pendekatan *Stanford CoreNLP* dan GATE yang sedia ada dalam mengesktrak maklumat daripada Glosari Istilah Islamik sebagai input memberi hasil yang sepatutnya?

1.3 OBJEKTIF KAJIAN

Objektif kajian yang dibangunkan ialah:

- i. Pengestrakan entiti nama berdasarkan *Stanford CoreNLP* dan GATE untuk mengekstrak maklumat berdasarkan Glosari Istilah Islamik
- ii. Menilai keberkesanan teknik pengekstrakan yang dijalankan ke atas Glosari Istilah Islam

1.4 SKOP KAJIAN

Konsep pengekstrakan berdasarkan GATE merupakan perisian sumber terbuka yang berupaya untuk menyelesaikan hampir semua masalah berkenaan pemprosesan teks. Perisian ini diuruskan oleh komuniti dari *University of Shefflied* yang terdiri daripada pemaju, pengguna, pendidik, pelajar dan para saintis.

Kajian penyelidikan tentang pembaharuan dalam perisian ini giat dijalankan (GATE, 2017). Kajian terbaru yang menarik minat ialah mengenai rangka kerja untuk pengumpulan dan menganalisis kandungan media dalam skala besar (Maynard et al. 2017). Badan ini telah melakukan analisis tentang penyalahgunaan media sosial (*Twitter*) terhadap ahli politik ketika kempen pemilihan dijalankan. Proses pengekstrakan maklumat ini bermula daripada analisis teks masa nyata (*real-time*

text analysis) dan melalui proses-proses seperti tokenisasi, *normalizer*, penandaan golongan kata dan pengecaman entiti nama (Bontcheva, 2017).

Seterusnya mengenai konsep pengekstrakan menggunakan *Stanford CoreNLP*. Perisian ini menyediakan alatan teknologi untuk bahasa manusia. Perisian ini berfungsi menyediakan asas untuk kepelbagaian perkataan dan penandaan golongan kata sama ada ayat yang diutarakan oleh manusia itu mempunyai kata nama seperti kata nama am, kata nama khas, kata nama tempat, tarikh, masa, kuantiti angka dan juga lain-lain (Manning,2014).

Perisian ini merupakan senjata utama bagi para penyelidik yang sememangnya arif dalam penggunaan bahasa tabii dan ingin menggunakan sistem analisis bahasa kearah yang lebih mendalam lagi. Perisian ini sememangnya memberikan kemudahan untuk aplikasi penggunaan bahasa yang sekarang ini menyebabkan terjadinya letupan maklumat yang menyebabkan kecelaruan maklumat.

Oleh itu, setidaknya dengan adanya perisian-perisian pengekstrakan maklumat seperti ini dapat membantu dalam mengurangkan kecelaruan maklumat. Secara umumnya, kajian ini tertumpu kepada beberapa skop sebagaimana yang dinyatakan dibawah:

- i. Glosari Istilah Islamik dalam Bahasa Inggeris diperolehi dari sumber Kumpulan DEED dari Universiti Antarabangsa Islam Malaysia (UIAM).
- ii. Dokumen dalam format teks.
- iii. Penggunaan konsep pengekstrakan berdasarkan *Stanford CoreNLP*.
- iv. Penggunaan konsep pengekstrakan berdasarkan GATE.

1.5 METODOLOGI KAJIAN

Metodologi kajian yang akan dilaksanakan dalam kajian ini terbahagi kepada 4 tahap utama seperti ditunjukkan pada rajah 1.1 dibawah iaitu, tahap pengumpulan maklumat, tahap pembangunan teknik konsep pengekstrakan, fasa pengujian, dan tahap dokumentasi.



Rajah 1.1: Ringkasan metodologi kajian

Fasa pertama iaitu fasa pengumpulan maklumat dimana berlakunya proses mendapatkan maklumat yang akan dilaksanakan sebelum melakukan kajian lanjut. Pada tahap ini minat terhadap topik perlu utuh untuk memastikan kajian dijalankan dengan sukses. Penyelidik perlu melihat situasi secara menyeluruh agar pengenalpastian masalah dan pernyataan dapat diselesaikan. Maklumat ini dapat diperoleh dari pelbagai sumber yang diyakini seperti buku teks, jurnal atas talian, laman sesawang, dan lain-lain. Penggunaan laman jurnal elektronik memainkan peranan penting untuk memastikan penyelidik mempunyai maklumat dan informasi yang mencukupi untuk membangunkan projek ini. Jurnal yang relevan seperti Jurnal

Teknologi Maklumat & Multimedia memudahkan pencarian dilakukan untuk topik ini.

Seterusnya fasa kedua ialah tahap pembangunan teknik konsep pengestrakan. Fasa ini membabitkan penggunaan aplikasi Java untuk menilai teknik konsep pengestrakan menggunakan perisian yang disebutkan diawal permulaan penulisan. Fasa ini merupakan fasa yang memperlihatkan segala jenis peraturan atau pengkodan digunakan ke atas projek.

Selain itu, fasa ketiga merupakan fasa pengujian di mana pengumpulan teks islamik glosari akan dilaksanakan menjadi korpus dalam bentuk teks. Pemilihan teks akan berlaku secara berperingkat dimana tidak semua maklumat didalam islamik glosari akan dikumpulkan.

Terakhir sekali ialah fasa dokumentasi. Fasa ini membabitkan perincian proses penglibatan pengumpulan maklumat dari mula hingga fasa implementasi dan pengujian dilaksanakan. Seterusnya, konklusi dan sumbangan kepada bidang penyelidikan akan dinyatakan.

1.6 ORGANISASI KAJIAN

Organisasi kajian bermaksud bagaimana cara susunan penulisan dalam projek ini akan dijalankan iaitu pengenalan, kajian literasi, metodologi kajian, implementasi kajian dan pengujian serta perbincangan dan akhir sekali ialah kesimpulan. Penerangan lanjut tentang setiap bab akan dibincangkan seperti dibawah.

Pertamanya ialah bab 1 iaitu bab pengenalan yang menggariskan tentang latar belakang kajian, pernyataan masalah yang diutarakan serta objektif kajian yang mahu dicapai setelah berakhirnya projek ini. Selain itu, dalam bab 1 ini, penerangan tentang skop kajian dan metodologi kajian akan dilakukan.

Bab 2 pula membincangkan secara terperinci tentang kajian-kajian atau projek-projek terdahulu yang dijalankan oleh para penyelidik dalam membimbing arah tuju penulisan projek ini. Kajian literasi merupakan simpulan atau hasil kajian yang dapat dibuktikan seterusnya digunakan sebagai rujukan untuk penyelidik masa kini dan juga di masa hadapan. Pendekatan atau teknik pengestrakan yang

digunakan juga akan dikupas lanjut untuk menunjukkan bahawa begitu banyak cara untuk mengekstrak maklumat daripada sumber yang besar.

Seterusnya, bab 3 merupakan bab yang menerangkan secara eksklusif bagaimana terjadinya proses evaluasi sistem pengekstrakan maklumat itu sendiri. Proses terbabit meliputi proses pengumpulan maklumat, penilaian konsep pengekstrakan, pengujian dan perbandingan serta dokumentasi.

Bab 4 merupakan fasa pengujian dan penilaian kedua-dua perisian pengekstrakan maklumat menggunakan Glosari Istilah Islamik.

Terakhirnya, bab 5 merupakan fasa dimana perbincangan dan rumusan akan dilakukan setelah berakhirnya fasa pengujian dijalankan. Disamping itu juga, sumbangan kajian, masalah serta cadangan kajian untuk masa hadapan juga akan disertakan sebagai panduan untuk penyelidikan seterusnya.

BAB II

KAJIAN LITERASI

2.1 PENGENALAN

Bab dua penulisan projek ini ialah kajian literasi yang memperkatakan tentang pandangan-pandangan para penyelidik berkaitan dengan konsep pengekstrakan maklumat, penggunaan pendekatan-pendekatan yang dibahaskan seperti pengecaman entiti nama, dan juga domain yang digunakan serta menganalisa dapatan dari kajian yang dijalankan.

Secara keseluruhannya, bab 2 ini akan diterangkan secara berperingkat mengikut bahagian-bahagiannya seperti pemprosesan bahasa tabii (NLP), pengekstrakan maklumat (IE), pengecaman entiti nama dan lain-lain. Seterusnya, berkaitan dengan jenis set data dan tumpuan khusus yang akan diberikan kepada analisis kandungan dokumen iaitu Glosari Istilah Islamik dan akhir sekali kesimpulan pada bab ini.

2.2 PEMROSESAN BAHASA TABII (NLP)

Pemprosesan Bahasa tabii (NLP) merupakan satu kaedah untuk membangunkan model komputer bagi menganalisis dan memahami bahasa pertuturan dan penulisan manusia (Hadi 2011). Pemprosesan Bahasa tabii (NLP) juga boleh dikatakan sebagai pembelajaran tentang kebolehan komputer dalam melakukan penterjemahan yang melibatkan bahasa manusia dari segi percakapan mahupun bahasa-bahasa tertentu. Antara teknik-teknik pemprosesan bahasa tabii yang penting dan selalu digunakan dalam dunia pembelajaran mesin ialah pengkesktrakan maklumat, capaian maklumat, penterjemahan mesin, sistem soal-jawap dan lain lain lagi.

Oleh kerana ledakan pelbagai jenis teknologi dan maklumat masa kini, pengeskrakan maklumat bertambah sukar kerana dek penggunaan pelbagai bahasa yang meluas. Oleh kerana itu, pemprosesan bahasa tabii merupakan cabang yang penting untuk merealisasikan sesuatu arahan pengkomputeran.

Salah satu cabang pemprosesan bahasa tabii ialah *regular expressions*. Penggunaan kalimah “Allah” dan “ALLAH” banyak digunakan dalam Glosari Istilah Islamik dan itu menyebabkan percanggahan maklumat sewaktu pencarian dilakukan sedangkan kedua-dua terma itu sama sahaja maksudnya. Di sinilah, kelebihan penggunaan *regular expressions* dimana proses yang perlu dilakukan adalah memadankan *strings* yang tidak padan dan tidak memadankan *strings* yang telah padan antara satu sama lain.

Selain itu, morfologi merupakan analisis yang dijalankan untuk menghuraikan perkataan kepada bahagian-bahagian yang membawa makna tertentu (Kamil 2015). Teknik ini penting untuk pengekstrakan maklumat dari set data yang melibatkan dwi-bahasa atau terma-terma yang membawa maksud yang pelbagai. Mengikut kajian terdahulu, sistem pengekstrakan terminologi telah banyak dibangunkan (Kit and Liu, 2008) oleh kerana sumber bahasa yang telah diperkaya dan perkembangan pesat bidang pemprosesan bahasa tabii. Pendekatan pengekstrakan terminologi yang selalu digunakan menurut Zhang ialah linguistik, statistik dan pendekatan campuran.

Dalam kajian ini, pendekatan pengekstrakan terminologi mungkin tidak sesuai dijalankan kerana glosari itu sendiri terdiri daripada gabungan-gabungan terma-terma dari dalam Al-Quran dan telah diterjemahkan kepada bahasa Inggeris. Ini akan menimbulkan kesamaran dalam pencarian maklumat. Keputusan yang baik boleh diperolehi jika set data yang kecil digunakan dan kebolehsesuaian sesuatu bahasa diambil kira.

Pemprosesan bahasa tabii mula digunakan pada era 1940-an di mana mesin penterjemah mula-mula diperkenalkan bagi menyahkod maklumat musuh dalam perang dunia kedua. Namun, kajian dalam bidang ini tidak banyak dilakukan sehinggalah pada tahun 1980-an (Alfred et al.2014). Berbeza sekali keadaannya dengan zaman sekarang dimana, segalanya-segalanya menggunakan teknologi berasaskan bidang sains komputer dan memerlukan komputer mempunyai kebolehan

untuk memahami bahasa manusia. Antara bidang-bidang yang memerlukan aplikasi NLP ialah pengekstrakan maklumat (IE), capaian maklumat (IR) dan lain-lain.

2.3 PENGEKSTRAKAN MAKLUMAT (IE)

Untuk membincangkan hal –hal seterusnya yang berkait rapat dengan bidang pemprosesan bahasa tabii, pengekstrakan maklumat (IE) merupakan salah satu cabang terpenting dalam bidang pemprosesan bahasa tabii. Pengekstrakan maklumat menggunakan pemprosesan bahasa tabii untuk mencapai maklumat dan mengecam entiti-entiti tertentu dari teks dokumen yang tidak berstruktur tanpa sebarang kefahaman mendalam tentang makna teks tersebut (Kamil 2015).

Pengekstrakan maklumat menggunakan peraturan-peraturan tertentu ataupun pendekatan yang berbeza untuk melaksanakan proses mengeskrak maklumat dari sebarang dokumen. Maklumat yang diekstrak kemudiannya boleh dipaparkan, disimpan ke dalam pangkalan data ataupun digunakan untuk pengindeksan oleh enjin carian seperti Google (Cunningham 2006). Fungsi untuk menyimpan dan boleh digunakan semula untuk tujuan masa depan merupakan satu kelebihan yang membuatkan pengekstrakan maklumat menjadi terkini dan sentiasa berubah-ubah.

Menurut Kovacevic dalam penulisannya, sistem pengekstrakan maklumat boleh dilakukan berdasarkan pembelajaran mesin dengan melaksanakan pengekstrakan automatik dan proses pengklasifikasi menggunakan metadata berpaksikan lapan kategori pratakrif (predefined). Lapan kategori yang telah diklasifikasikan dalam *metadata* ialah tajuk (*title*), penulis (*authors*), penggabungan (*affiliation*), alamat (*address*), email, abstrak, kata kunci, dan penerbitan dan nota. *Metadata* boleh difahami sebagai sesuatu data yang menghubungkan kepada satu data yang lain. Seperti yang boleh dilihat, kelapan-lapan kategori yang disebutkan boleh melampirkan maklumat deskriptif mengenai sesuatu konteks. Ia juga boleh dilihat sebagai elemen yang memudahkan pemahaman tentang sesuatu dokumen.

Kovacevic juga menyebut dalam penulisannya bahawa sistem pengekstrakan maklumat yang dibina melalui proses klasifikasi yang menggunakan model klasifikasi seperti *decision tree (DT)*, *Naïve Bayes (NB)*, dan mesin vektor sokongan (*support vector machine*).

Antara sistem pengekstrakan maklumat yang dikenal pasti sepanjang penulisan kajian ialah, pengekstrakan terminologi dwi-bahasa menggunakan kepelbagaian-level *termhood*. Terminologi ditakrifkan sebagai kumpulan perkataan teknikal atau penggunaan konteks yang spesifik dalam sesuatu bidang atau teras yang sama. Selalunya digunakan dalam pembelajaran mesin, capaian maklumat dan juga pengekstrakan maklumat. Persamaan kajian ini dengan kajian yang dilakukan oleh saya ialah penggunaan dua bahasa iaitu Bahasa Inggeris dan Bahasa Arab.

Menurut kajian terdahulu, banyak sistem pengekstrakan terminology telah dibangunkan disebabkan kekayaan sumber bahasa-bahasa (Kit and Liu 2008). Kenyataan ini disokong oleh Bourigalt (1992) yang mengatakan bahawa pendekatan linguistik merupakan salah satu jenis pengekstrakan terminologi yang mengeksploitasi metod penandaan golongan kata dan juga penghurai untuk menapis terminologi. Selain itu, Ido dan Ward (1994) juga menyakini bahawa terminologi boleh mewakili corak perkataan kata nama dalam sebuah terminologi. Semua kajian di atas menunjukkan bahawa kadar penyesuaian sesuatu bahasa itu boleh diboleh-ubah mengikut cara linguistik yang ditetapkan.

Selain itu, konsep pengekstrakan maklumat menggunakan pengecaman entiti antara yang popular masa kini kerana ianya merangkumi semua aspek seperti entiti nama iaitu individu, organisasi, lokasi, tarikh atau masa dan maklumat kewangan serta peratusan. Berpatutan kepada senarai entiti nama tersebut, dapat dikonklusikan bahawa pengecaman entiti nama tidak bergantung pada domain tertentu untuk melaksanakannya (Chau et al. 2002). Kenyataan diatas disokong sepenuhnya oleh Darwich (2014) yang mengatakan konsep pengekstrakan melalui pengecaman entiti nama mampu mengesktrak maklumat tersebut hampir menyamai keupayaan manusia sebanyak 95% peratusan ketepatan.

2.4 PENGECAMAN ENTITI NAMA

Pengecaman entiti nama telah digunakan secara meluas dan memainkan tugas penting khususnya dalam bidang berkaitan pemprosesan bahasa tabii. Pengecaman entiti nama ditakrifkan sebagai satu tugas atau fungsi bagi mengecam, mengekstrak serta

mengklasifikasikan entiti yang terdapat di dalam teks yang tidak berstruktur. Entiti ini boleh dikenalpasti melalui kata nama khas, kata hubung serta kata sendi. Sebagai contoh, bagai 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 yang pertama *ENAMEX* yang mengandungi kelas *PERSON* bagi nama individu, kelas *LOCATION* bagi kelas lokasi dan kelas *ORGANIZATION* bagi organisasi. Kategori yang kedua ialah kelas *DATE* bagi tarikh atau hari dan kelas *TIME* untuk masa dan seterusnya kategori ketiga ialah *NUMEX* iaitu kelas *PERCENTAGE* bagi nilai peratusan dan kelas *MONETARY* bagi nilai kewangan.

Pada peringkat awalnya, teknik pengecaman entiti nama banyak diaplikasikan dalam bahasa Inggeris. Kemudian pada tahun 2002, satu persidangan *Conference on Computational Natural Language Learning* telah dianjurkan untuk memberi fokus kepada pengecaman entiti nama dalam bahasa selain daripada bahasa Inggeris. Dalam persidangan yang terbabit, bahasa Belanda dan Sepanyol dijadikan korpus kajian (Sang 2002) manakala pada tahun 2003 bahasa German dan Inggeris dijadikan korpus sebagai bahan kajian. Hasil dari kedua-dua persidangan ini telah menggariskan panduan bagi pengkategorian pengecaman entiti nama seperti berikut:

- 1) *PER*: kategori nama individu, sama seperti kelas *PERSON* dalam MUC-6.
- 2) *LOC*: kategori nama tempat atau lokasi, sama seperti kelas *LOCATION* dalam MUC-6.
- 3) *ORG*: kategori nama organisasi, sama seperti kelas *ORGANIZATION* dalam MUC-6.
- 4) *MISC*: kategori bagi entiti nama lain yang tidak termasuk dalam kategori di atas.

Selain itu, terdapat beberapa lagi kumpulan penyelidik yang turut melakukan kajian berkaitan dengan pengestrakan maklumat iaitu *Automated Content Extraction* (ACE). ACE telah ditubuhkan pada tahun 2000 bagi memberi fokus kepada penyelidikan dalam bidang pengestrakan maklumat. Pada mulanya, tumpuan penyelidikan tertumpu kepada teknik pengecaman entiti nama dan