



**PENCANTAS PERKATAAN MELAYU UNTUK AKSARA JAWI
BERASASKAN PETUA**

SULIANA BINTI SULAIMAN



**TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH IJAZAH
DOKTOR FALSAFAH**

**FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
BANGI**

2013





PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang setiap satunya telah saya jelaskan sumbernya.



1 Ogos 2013

SULIANA SULAIMAN
P 47840





PENGHARGAAN

Dengan nama Allah yang Maha Pengasih lagi Maha Penyayang.

Segala puji bagi Allah S.W.T tuhan sekalian alam. Selawat dan salam ke atas junjungan besar Nabi Muhammad S.A.W. Alhamdulillah, setinggi kesyukuran ke hadrat Allah S.W.T kerana limpah kurnia dan rahmatNya, dapat saya meyempurnakan tesis ini.

Ucapan penghargaan dan jutaan terima kasih ditujukan kepada barisan peyelia saya iaitu, Prof. Khairuddin Omar, Dr Nazlia Omar dan Encik Zamri Murah yang banyak memberi tunjuk ajar, saranan dan bantuan dari awal sehingga ke peringkat akhir tesis ini disiapkan. Penghargaan ini juga ditujukan buat sarjana tamu Tuan Haji Hamdan Abdul Rahman yang banyak membantu dan mencurahkan ilmu dalam menjayakan tesis ini. Semoga segala bentuk bantuan yang dicurahkan akan diberi ganjaran yang berlipat kali ganda oleh Allah S.W.T.

Penghargaan yang tidak terhingga juga diberikan kepada suami tercinta Mohd Nazir yang telah banyak mendoakan kejayaan dan memberikan dorongan serta galakan buat diri ini. Juga buat anakanda Harith Hakimi yang selalu memahami keadaan mama sebagai seorang pelajar. Tidak lupa juga buat emak dan ayah yang sentiasa mendoakan kejayaan saya. Semoga Allah sentiasa merahmati kalian di dunia dan di akhirat.

Selain itu sekalung penghargaan ditujukan buat Kementerian Pengajian Tinggi dan juga Universiti Pendidikan Sultan Idris kerana telah membiayai segala kos pengajian sepanjang menyempurnakan tesis ini melalui Skim Latihan Bumiputera (SLAB).

Akhir sekali, ucapan terima kasih buat rakan-rakan dan pensyarah Kumpulan Penyelidikan Pengecaman Pola atas segala bantuan dan kerjasama yang diberikan.





ABSTRAK

Pencantas perkataan merupakan proses membuang imbuhan pada perkataan dan menghasilkan perkataan tercantas ataupun kata dasar. Pencantas perkataan boleh digunakan dalam capaian dokumen, transliterasi, pengkelasan teks dan penterjemahan mesin. Pencantas perkataan yang dihasilkan dalam kajian yang lepas bagi Bahasa Melayu lebih tertumpu kepada tulisan Rumi. Set petua yang dihasilkan untuk mencantas imbuhan tidak sesuai untuk kata terbitan Jawi. Perbezaan ketara boleh dilihat pada petua pembuangan akhiran '-an'. Contohnya, untuk tulisan Rumi, akhiran '-an' dieja dengan menggunakan satu cara manakala untuk tulisan Jawi ianya boleh dieja sebagai 'اَن', 'ان', 'ءن' dan 'ن'. Oleh yang demikian, set petua Jawi diperlukan untuk membuang imbuhan pada kata terbitan Jawi. Selain itu pencantas perkataan Bahasa Melayu yang menggunakan kamus kata dasar perlu sentiasa dikemas kini untuk memastikan setiap perkataan yang dicantas sama dengan perkataan di dalam kamus untuk mengurangkan ralat. Objektif bagi tesis ini adalah untuk menghasilkan petua cantasan serta membangun dan menilai pencantas perkataan Jawi yang digunakan untuk mencantas kata terbitan dan menghasilkan kata dasar yang merangkumi kata jati melibatkan satu, dua dan tiga suku kata. Set data yang digunakan dalam kajian ini telah ditransliterasi ke dalam Jawi dan dibahagi kepada dua set, iaitu artikel-artikel daripada Utusan Melayu dan Berita Harian yang dipilih secara rawak di antara September 2009 - November 2010. Pangkalan data yang digunakan juga termasuklah Al-Quran terjemahan Sheikh Abdullah Basmeih yang telah digunakan dalam kajian yang lepas. Dalam penghasilan algoritma pencantas perkataan ini, terdapat dua komponen penting telah dihasilkan iaitu petua *nyah-imbuhan* untuk mencantas imbuhan dan petua *pengesanan kesalahan ejaan Jawi (SEDR)* yang digunakan untuk menyemak perkataan yang dicantas. Petua *nyah-imbuhan* melibatkan beberapa proses yang memerlukan pembuangan, penggantian dan penambahan aksara dalam setiap kata terbitan Jawi. Petua *SEDR* pula melibatkan susunan corak ejaan untuk membentuk suku kata Jawi. Sebanyak enam eksperimen telah dilakukan bermula dengan pengiraan ketepatan petua *SEDR*, pengiraan ketepatan petua *Rule Application Order* dan *Rule Frequency Order* menggunakan data Jawi, turutan pembuangan imbuhan, pengiraan ketepatan pencantas berasaskan penilaian Frakes dan Paice serta penilaian algoritma signifikasi berasaskan statistik. Hasil keseluruhan daripada kajian mendapati bagi nilai *min purata ketepatan (MPK)* dokumen Jawi yang dicantas adalah 8.43% manakala nilai *MPK* dokumen Jawi yang tidak dicantas adalah 5.14%. Pencantas perkataan Melayu untuk aksara Jawi ini dapat membantu meningkatkan ketepatan dalam capaian dokumen Jawi.





A MALAY STEMMEER FOR JAWI CHARACTERS BASED ON RULE

ABSTRACT

Stemming is the process used to remove affixes from words to produce stemmed words; or root words. Stemmers can be used in document retrieval, transliteration, text classification and machine translation. Malay stemmers produced in previous studies focus more on Roman (Rumi) script. The rule set to produce stem affixes is unsuitable for derived Jawi words. Significant differences can be seen in the deaffixation of suffix rule '-an'. For example, for Roman (Rumi) script, the suffix '-an' is spelt using one technique; whilst for Jawi script, it can be spelled as 'اَن', 'ان', 'ءن' and 'ن'. Therefore, a new Jawi rule set is needed to remove affixes from Jawi derived words. Additionally, Malay stemmers that use the root word dictionary need to be updated constantly, to ensure every stem word matches with the corresponding dictionary word to reduce the errors. The objective of this thesis is to produce stemmer rule and also develop and evaluate of a new Jawi stemmer used to stem derived words and to produce root words that encompass pure words involving one, two, and three syllables. The data set used in this work, has been transliterated into Jawi and has been divided into two sets. This study use articles from Utusan Melayu and Berita Harian which had been randomly picked between September 2009 – November 2010. The database also includes the translation of Al-Quran by Sheikh Abdullah Basmeih which has been used in one of the previous studies. In developing the proposed stemmer algorithm, two important components had been presented, namely a deaffixation rule to stem affixes and a Spelling Error Detector Rule (SEDR), to validate the stemmed word. The deaffixation rule involves several processes that require the removal, replacement, and character addition, within the Jawi derived word. SEDR involves the arrangement of a spelling pattern to form Jawi syllables. Several experiments were performed to prove that the Jawi stemmer could assist in Jawi document retrieval. A total of six experiments have been carried out starting with accuracy calculation of the SEDR rule, accuracy calculation of *Rule Application Order* and *Rule Frequency Order* using Jawi word, sequencing of the deaffixation, calculation of the stemmer accuracy based on Frakes and Paice assessments, and significant assessment of algorithms based on statistics. The overall result of this study found that the Mean Average Precision (MAP) of the stemmed Jawi documents is 8.43%, while the MAP value of the non-stemmed Jawi documents is 5.14%. Therefore, the Malay stemmer for Jawi characters could help to improve the accuracy of Jawi document retrieval.





KANDUNGAN

	Halaman
PENGAKUAN	ii
PENGHARGAAN	iii
ABSTRAK	iv
ABSTRACT	v
KANDUNGAN	vi
SENARAI JADUAL	xi
SENARAI ILUSTRASI	xiv

BAB 1 PENGENALAN

1.1	Pendahuluan	1
1.2	Pernyataan Masalah	1
1.3	Objektif Kajian	4
1.4	Kepentingan Kajian	4
1.5	Skop Kajian	5
1.6	Struktur Organisasi Tesis	6
1.7	Kesimpulan	7

BAB 2 KAJIAN LITERATUR

2.1	Pengenalan	9
2.2	Pencantas Perkataan	10
	2.2.1 Pencantas perkataan pembuangan imbuhan	13
	2.2.2 Pencantas perkataan varieti pengganti	14
	2.2.3 Pencantas perkataan carian jadual	14
	2.2.4 Pencantas perkataan n-gram	15
2.3	Pencantas Perkataan Bahasa Inggeris	16
	2.3.1 Algoritma Lovins	16
	2.3.2 Algoritma Dawson	17
	2.3.3 Algoritma Porter	18
	2.3.4 Algoritma Paice/Husk	19





2.4	Pencantas Perkataan Bahasa Perancis	20
2.5	Penemberengan Perkataan Bahasa Cina	22
2.6	Pencantas Perkataan Parsi	23
	2.6.1 Algoritma Kazem Tagva et. al	23
	2.6.2 Algoritma Somayye Estahbani et. al	24
2.7	Pencantas Perkataan Bahasa Arab	24
	2.7.1 Algoritma Belal	24
2.8	Pencantas Perkataan Bahasa Indonesia	26
2.9	Pencantas Perkataan Bahasa Melayu	28
	2.9.1 Algoritma pencantas perkataan Asim	28
	2.9.2 Algoritma pencantas Fatimah	29
	2.9.3 Penganalisis morfologi Melayu	31
	2.9.4 Algoritma pencantas perkataan Idris	33
	2.9.5 Algoritma pencantas perkataan Taufik	35
2.10	Masalah Adaptasi Pencantas Rumi Ke Atas Jawi	37
2.11	Ralat Pencantas Perkataan	37
2.12	Penilaian Pencantas Perkataan	38
	2.12.1 Kaedah pengiraan ralat Paice	38
	2.12.2 Kaedah Frakes	41
	2.12.3 Penilaian pencantas perkataan menggunakan capaian dokumen	42
2.13	Kesimpulan	44

BAB 3 METODOLOGI

3.1	Pengenalan	50
3.2	Metodologi Kajian	50
	3.2.1 Reka bentuk kajian	50
	3.2.2 Kerangka kerja kajian	53
	3.2.3 Reka bentuk eksperimen	55
3.3	Sukatan Prestasi	57
3.4	Ujian Signifikasi Berasaskan Statistik	57
3.5	Alatan Kajian	58
3.6	Kesimpulan	59



BAB 4 BAHASA MELAYU DAN TULISAN JAWI

4.1	Pengenalan	60
4.2	Pembentukan Kata Tunggal	62
4.3	Perbezaan Ejaan Rumi dan Jawi	64
4.3.1	Suka kata yang menggunakan [e] Pepet	64
4.3.2	Suka kata yang menggunakan vokal [a]	65
4.3.3	Suku kata yang menggunakan vokal [i] atau [e]	66
4.3.4	Suku kata yang menggunakan vokal [u] atau [o]	66
4.4	Pembentukan Kata Terbitan	66
4.4.1	Awalan menN- dan peN-	67
4.4.2	Awalan se-, ke- dan di-	69
4.4.3	Awalan beR-, teR- dan peR-	69
4.4.4	Akhiran -an	69
4.4.5	Akhiran -i	70
4.4.6	Akhiran-akhiran lain	70
4.4.7	Apitan	70
4.5	Pembentukan Kata Ganda	71
4.6	Partikel	71
4.7	Kesimpulan	72

BAB 5 PEMBANGUNAN PETUA PENCANTAS PERKATAAN BAHASA MELAYU UNTUK AKSARA JAWI

5.1	Pengenalan	75
5.2	Petua Pengesanan Kesalahan Ejaan Kata Dasar Jawi	76
5.2.1	Petua pengesanan kesalahan ejaan ekasuku	76
5.2.2	Petua pengesanan kesalahan ejaan dwisuku	77
5.2.3	Petua pengesanan kesalahan ejaan trisuku	81
5.3	Proses Penyemakan Ejaan Jawi	81
5.4	Eksperimen I: Menentukan Nilai Ketepatan Petua Pengesanan Kesalahan Ejaan	83
5.5	Eksperimen II: Mengira Nilai Ketepatan Petua Nyah-Imbuhan Fatimah dan Taufik ke Atas Data Jawi	86
5.6	Petua Nyah-Imbuhan	88
5.6.1	Petua nyah-imbuhan awalan	90
5.6.2	Petua nyah-imbuhan akhiran	94

5.6.3	Petua nyah-imbuhan apitan	96
5.6.4	Petua nyah-imbuhan sisipan	97
5.7	Eksperimen III: Menentukan Turutan Keutamaan Dalam Proses Pembuangan Imbuhan	97
5.8	Kesimpulan	99

BAB 6 PEMBANGUNAN ALGORITMA PENCANTAS PERKATAAN BAHASA MELAYU UNTUK AKSARA JAWI

6.1	Pengenalan	101
6.2	Kata Henti (Stop Word)	102
6.3	Algoritma Pencantas Perkataan Jawi	103
6.4	Ekperimen IV: Menentukan Ketepatan Pencantas Perkataan Jawi	107
6.4.1	Peratus nilai ketepatan bagi pencantas perkataan Jawi	108
6.4.2	Penilaian Berdasarkan Kaedah Paice	110
6.5	Ekperimen V: Penilaian Berdasarkan Kaedah Frakes	112
6.6	Kesimpulan	116

BAB 7 KEBERKESANAN ALGORITMA PENCANTAS DALAM CAPAIAN DOKUMEN JAWI

7.1	Pengenalan	117
7.2	Set Data	118
7.3	Set Pertanyaan (<i>Query</i>)	119
7.4	Penilaian Releven (<i>Relevance Judgement</i>)	120
7.5	Kaedah Analisis Data	120
7.6	Indri 1.0	121
7.7	Eksperimen VI: Penilaian Algoritma Pencantas Terhadap Capaian Dokumen Dan Penilaian Ujian Signifikasi Berasaskan Statistik	125
7.7.1	Ketepatan dan perolehan kembali dan min purata ketepatan (MPK)	125
7.7.2	Ujian Signifikasi	128
7.8	Kesimpulan	134



BAB 8 KESIMPULAN DAN SUMBANGAN

8.1	Pengenalan	135
8.2	Kesimpulan Kajian	135
8.3	Dapatan kajian	136
8.3.1	Objektif 1: Menghasilkan petua penyemakan kesalahan ejaan (SEDR) yang dapat menyemak ejaan kata dasar bagi aksara Jawi	136
8.3.2	Objektif 2: Menghasilkan petua nyah-imbuhan yang digunakan untuk mencantas kata terbitan Jawi	137
8.3.3	Objektif 3: Membangunkan algoritma pencantas perkataan Bahasa Melayu bagi aksara Jawi yang dapat mencantas kata terbitan kepada kata dasar.	138
8.4	Sumbangan Kajian	139
8.5	Cadangan Perluasan Kerja	139
8.6	Kesimpulan	140
	RUJUKAN	141



LAMPIRAN

A	Senarai Penerbitan	149
B	Senarai Petua Nyah-Imbuhan	150
C	Senarai Petua SEDR Dwisuku	152
D	Senarai Petua SEDR Trisuku	157
E	Set Pertanyaan	163
F	Set Releven	165
G	Ujian Sampel Pasangan t	190
H	104 Kata Terbitan Dalam Jawi	197



**SENARAI JADUAL**

No. Jadual		Halaman
2.1	Kesilapan yang dilakukan oleh algoritma Fatimah dengan menggunakan petua daripada set A.	30
2.2	Hasil kesilapan yang dilakukan oleh algoritma Fatimah dengan menggunakan petua daripada set B.	31
2.3	Petua ejaan yang digunakan untuk menghasilkan imbuhan awalan dan imbuhan apitan dalam Bahasa Melayu.	32
2.4	Perbandingan hasil eksperimen di antara algoritma N.Idris dan Fatimah.	34
2.5	Hasil kesilapan yang dilakukan oleh RFO stemmer	36
2.6	Perbandingan di antara setiap pengakar perkataan Bahasa Melayu	47
3.1	Alatan dan Bahasa Pengaturcaraan yang Digunakan.	58
4.1	Bentuk aksara dan cara penulisan aksara Jawi (Rusli, 2008)	61
4.2	Senarai aksara perangkai, pemutus dan perumah	62
4.3	Pelambangan vokal dalam ejaan Jawi	63
4.4	Penggunaan suku kata vokal [a].	63
4.5	Jenis-jenis suku kata bagi SKB dan SKT	64
4.6	Pola persekutuan kata.	65
4.7	Syarat dan gaya ejaan kata dasar yang menggunakan awalan meN- dan peN.	68
4.8	Gaya ejaan bagi akhiran -an	69
4.9	Penggunaan vokal [e] pepet, [a], [e], [i], [o] dan [u] dalam ejaan Jawi berdasarkan suku kata.	72
4.10	Petua-petua asas bagi imbuhan dalam Jawi.	74
5.1	Petua ejaan ekasuku	77
5.2	Petua ejaan Dwisuku yang bermula dengan vokal [e] pepet.	78





5.3	Petua ejaan Dwisuku yang bermula dengan vokal [a].	79
5.4	Petua ejaan Dwisuku yang bermula dengan vokal [i], [e], [o] dan [u]	80
5.5	Hasil semakan ejaan menggunakan petua SEDR	83
5.6	Peratus Ketepatan Petua SEDR	84
5.7	Contoh ralat yang dihasilkan oleh petua SEDR	85
5.8	Hasil ketepatan bagi petua nyah-imbuhan RAO dan RFO ke atas skrip Jawi	86
5.9	Jenis ralat yang dihasilkan oleh petua nyah-imbuhan RAO dan RFO	86
5.10	Ralat yang dihasilkan oleh RAO	87
5.11	Ralat yang dihasilkan oleh RFO	88
5.12	Bilangan imbuhan berdasarkan kumpulan bagi 1200 perkataan berimbuhan secara unik	98
5.13	Peratus ketepatan untuk D1-D6	99
6.1	Bilangan ralat yang dihasilkan dalam Ujian A dan Ujian B.	109
6.2	Prestasi Algoritma Pencantas Menggunakan Kaedah Paice	111
6.3	Prestasi Algoritma Pencantas Menggunakan Kaedah Frakes	113
7.1	Surah dan Ayat Al-Quran Yang Terlibat Dalam Penyediaan Set Dokumen.	118
7.2	Purata Ketepatan Dan Perolehan Kembali Bagi Dokumen Yang Dicantas	126
7.3	Purata Ketepatan Dan Perolehan Kembali Bagi Dokumen Yang Tidak Dicantas	126
7.4	MPK bagi dokumen Jawi Yang Dicantas Dan Dokumen Jawi Yang Tidak Dicantas.	128
7.5	Bilangan Relevan Dokumen Yang Dicapai Bagi 36 Pertanyaan Untuk Dua Jenis Dokumen Yang Berbeza (20 titik pemisah)	128
7.6	Statistik Sampel Pasangan	129





7.7	Ujian Sampel Pasangan	130
7.8	Purata Ketepatan Pada Pelbagai Titik Pemisahan Untuk 36 Pertanyaan Bagi Dokumen Yang Dicantas.	131
7.9	Menunjukkan Purata Ketepatan Pada Pelbagai Titik Pemisahan Untuk 36 Pertanyaan Bagi Dokumen Yang Tidak Dicantas.	132
7.10	Keputusan ujian sampel pasangan -T bagi setiap titik pemisahan.	133





SENARAI ILUSTRASI

No. Rajah		Halaman
2.1	Algoritma pencantas perkataan mengikut bahasa	13
2.2	Contoh jadual yang mengandungi indeks kata dan perkataan yang dicantas bagi pencantas perkataan carian jadual	15
2.3	Perkataan <i>engineering</i> dan <i>engineers</i> yang dipecahkan dalam bentuk diagram	15
2.4	Carta Alir Bagi Algoritma Belal (Belal, 2001)	26
2.5	Kedudukan setiap kumpulan imbuhan apabila dicantumkan dengan kata dasar	28
3.1.	Reka bentuk kajian	51
3.2	Kerangka kerja kajian	54
5.1	Proses Penyemakan Ejaan Jawi	82
5.2	Perkataan yang berjaya disemak dan ralat yang dihasilkan	85
5.3	Bilangan petua Rumi dan petua Jawi	90
6.1.	Carta Alir Algoritma Pencantas Perkataan Untuk Aksara Jawi	105
6.2.	Graf Peratus Ketepatan Untuk Setiap Algoritma Cantasan	110
6.3.	Perbandingan MWC untuk Ujian A dan Ujian B	113
6.4.	Perbandingan Di Antara WCF Untuk Ujian A Dan Ujian B	114
6.5	ICF untuk setiap pencantas bagi Ujian A dan Ujian B	115
6.6.	Min CR Bagi Setiap Pencantas	115
7.1.	Penggunaan Indri bagi penghasilan indeks	123
7.2	Indeks yang telah dihasilkan oleh IndexerUI	123
7.3	Hasil capaian ke atas dokumen yang tidak dicantas Menggunakan Indri	124
7.4	Graf Ketepatan Dan Dapatan Bagi Dokumen Jawi Yang Dicantas Dan Dokumen Jawi Yang Belum Dicantas	127





BAB I

PENGENALAN

1.1 PENDAHULUAN

Pencantas perkataan digunakan untuk mencantas kata terbitan dan menghasilkan kata dasar. Pencantas perkataan yang terawal dihasilkan oleh Julie Beth Lovins pada tahun 1968 (Lovins, 1968). Kegunaan pencantas perkataan tidak terhad kepada bidang capaian dokumen sahaja tetapi ia juga amat penting dalam bidang transliterasi, serta penyemakan ejaan. Pencantas perkataan yang pertama bagi Bahasa Melayu telah dihasilkan oleh Asim Othman pada tahun 1993 (Asim, 1993). Seterusnya pencantas perkataan yang ada telah dilakukan penambahbaikan dan dapat mencantas perkataan Bahasa Melayu dengan baik. Kajian yang dilakukan telah membuktikan bahasa pencantas perkataan Bahasa Melayu ini dapat membantu dalam capaian dokumen. Walau bagaimanapun tumpuan utama penghasilan pencantas perkataan Bahasa Melayu tertumpu kepada aksara Rumi dan tidak merangkumi ejaan bagi aksara Jawi.

1.2 PERNYATAAN MASALAH

Bahasa Melayu boleh ditulis dengan menggunakan dua jenis aksara yang berbeza iaitu Rumi dan Jawi. Tulisan Jawi pernah digunakan sebagai tulisan utama untuk berkomunikasi dan kemudiannya tulisan ini telah digantikan dengan tulisan Rumi.

Morfologi bagi Bahasa Melayu lebih rumit jika dibandingkan dengan morfologi Bahasa Inggeris (Fatimah, 1995). Kajian mengenai pencantas perkataan dalam Bahasa Melayu telah dilakukan oleh beberapa orang penyelidik seperti Asim (1993), Fatimah (1995), Sock (2000), Idris (2001) dan Taufik (Muhammad Taufik,





2009). Namun hanya Sock (2000) sahaja yang menggunakan kaedah N-gram dalam penghasilan pencantas perkataannya manakala penyelidik yang lain menggunakan kaedah petua morfologi.

Dalam kajian Asim (1993), penyelidik telah menggunakan kamus untuk menyemak imbuhan yang telah dicantas. Walau bagaimanapun dalam kajian Fatimah (1995), telah menggantikan penggunaan kamus yang digunakan oleh Asim dengan 'kamus kata dasar'. Penggunaan kamus kata dasar dapat meningkatkan lagi ketepatan cantasan (Fatimah, 1995). Hasil kajian Idris (2001), penyelidik telah menambah satu lagi kamus khas iaitu 'kamus tempatan' yang dapat mengurangkan ralat dalam cantasan. Fungsi kamus khas adalah untuk mengelakkan ralat terlebih cantas dalam sesuatu subjek tertentu. Dalam kajian Idris (2001), kamus tempatan yang digunakan adalah kamus bagi mata pelajaran Sejarah.

Kamus perlu dikemaskini untuk memastikan perkataan yang baru dicantas dapat ditemui sekaligus mengelak daripada berlakunya ralat terlebih cantas dan ralat terkurang cantas (Kazem, 2005; Somayye Estahbani & Reza Javidan, 2011). Selain itu kajian yang dilakukan terhadap pencantas perkataan Bahasa Melayu lebih tertumpu kepada aksara Rumi (Asim, 1993; Fatimah, 1995; Sock, 2000; Idris, 2001, Muhammad Taufik, 2009).

Walaupun tulisan Rumi dan Jawi digunakan untuk mewakili satu bahasa yang sama tetapi terdapat perbezaan dalam aksara dan gaya ejaan. Aksara Rumi sama seperti aksara yang digunakan dalam Bahasa Inggeris manakala aksara Jawi pula sebahagiannya sama seperti aksara Arab. Terdapat enam penambahan aksara lain dalam Jawi menyebabkan hanya sebahagian aksara Jawi menyerupai aksara Arab.

Bentuk ejaan bagi Rumi dan Jawi juga berbeza. Dalam Rumi, vokal diwakili oleh lima aksara yang berbeza iaitu a, e, i, o dan u manakala bagi Jawi pula lima bunyi vokal hanya diwakili oleh tiga aksara yang berbeza iaitu ا, و dan ي. Setiap suku kata dalam ejaan Rumi mempunyai aksara vokalnya yang tersendiri. Tetapi bagi ejaan Jawi ada suku kata yang dieja tanpa menggunakan sebarang aksara vokal.





Selain itu terdapat perbezaan di antara petua yang digunakan untuk mencantas imbuhan. Perbezaan yang ketara boleh dilihat pada imbuhan awalan, akhiran dan apitan. Contohnya untuk mencantas perkataan *binaan* dan *bukaan* bagi aksara Rumi, akhiran *-an* perlu dipanggil untuk menghasilkan kata dasar *bina* dan *buka*. Tetapi untuk mencantas kata terbitan بينان dalam aksara Jawi, akhiran -ءن perlu dipanggil untuk menghasilkan kata dasar بينا. Dalam sesetengah situasi, untuk mencantas kata terbitan بوکان akhiran -ان perlu dipanggil untuk menghasilkan kata dasar بوك. Oleh yang demikian untuk mencantas akhiran *-an* dalam Jawi memerlukan perhatian sama ada untuk memanggil petua -ان atau -ءن.

Perbezaan juga wujud dalam kata serapan Bahasa Inggeris. Berdasarkan buku Daftar Kata Rumi-Sebutan-Jawi terbitan Dewan Bahasa dan Pustaka edisi ke-2 dinyatakan, ejaan kata serapan Bahasa Inggeris yang mempunyai dua aksara Jawi di akhir perkataan yang membentuk kelompok konsonan tidak perlu ditulis secara bersambung. Contohnya perkataan "golf" perlu ditulis sebagai كولف dan bukannya كولف. Aksara ل dan ف dianggap sebagai dua aksara konsonan dan perlu ditulis secara berasingan tetapi sekiranya aksara konsonan tadi ditambah dengan aksara vokal di akhir perkataan seperti perkataan 'fakta' maka suku kata akhir perlu dieja sebagai suku kata terbuka seperti فكتا. Kesemua perkataan Bahasa Inggeris dan Eropah yang memerlukan penggunaan aksara [g] dieja dengan menggunakan aksara 'ga' dalam Jawi. Contohnya perkataan seperti agenda dan gimnasium. Untuk aksara [k] pula dieja dengan menggunakan aksara ك dan bukannya ق.

Bagi kata serapan Bahasa Arab pula, aksara vokalnya merujuk kepada lambang kepanjangan atau dikenali juga sebagai mad. Aksara vokal yang sama boleh dijumpai dalam dua perkataan yang berbeza tetapi mempunyai jenis kepanjangan yang berbeza. Perkataan yang mempunyai vokal yang panjang akan dieja dengan menggunakan salah satu aksara vokal ي, و, ا manakala perkataan yang mempunyai aksara vokal pendek akan dieja dengan tanpa menggunakan aksara vokal contohnya 'dam' (vokal pendek) dieja sebagai دم dan 'am' (vokal panjang) dieja sebagai عام (Hamdan Abdul Rahman, 1999).





Untuk memastikan kata terbitan dapat dicantas dengan tepat, pencantas perkataan Melayu untuk aksara Jawi telah dibangunkan. Objektif bagi kajian ini akan dibincangkan seperti berikut.

1.3 OBJEKTIF KAJIAN

Objektif bagi kajian ini adalah seperti berikut:

1. Menghasilkan petua penyemakan kesalahan ejaan (SEDR) yang dapat menyemak ejaan kata dasar bagi aksara Jawi.
2. Menghasilkan petua nyah-imbuhan yang digunakan untuk mencantas kata terbitan Jawi.
3. Membangunkan algoritma pencantas perkataan Bahasa Melayu bagi Aksara Jawi yang dapat mencantas kata terbitan Jawi kepada kata dasar.

1.4 KEPENTINGAN KAJIAN



Pencantas perkataan diperlukan sebagai asas untuk menyokong capaian maklumat sesuatu bahasa dan digunakan dalam aplikasi terjemahan dokumen dan carian sesawang (Jelita Asian, 2005). Berdasarkan Van Rijsbergen (1979), teknik pencantas perkataan dapat mengurangkan saiz indeks dan membantu capaian maklumat yang lebih relevan. Menurut beliau lagi proses pencantas perkataan dapat mengurangkan saiz perwakilan dokumen kepada 20%-50% jika dibandingkan dengan perwakilan penuh perkataan. Dalam tesis Belal (2001), penyelidik menyatakan kepentingan pengakar perkataan sebagai fungsi utama dalam sistem capaian dokumen kerana ianya berupaya mengurangkan bilangan perkataan yang berbeza dan sekaligus mengurangkan saiz kamus.

Kepentingan kajian yang dilakukan boleh dilihat dalam beberapa faktor iaitu pencantas perkataan yang dihasilkan merupakan pencantas perkataan untuk mencantas kata terbitan Melayu bagi aksara Jawi. Petua nyah-imbuhan yang dihasilkan bersesuaian untuk digunakan dalam mencantas imbuhan Jawi dan kajian ini juga memperkenalkan penggunaan petua SEDR iaitu petua yang digunakan untuk menyemak kembali ejaan kata dasar dalam Jawi setelah proses pembuangan imbuhan





dilakukan. Dalam kajian pencantas perkataan yang lepas, kamus digunakan sebagai proses penyemakan setelah cantasan dilakukan. Perkataan yang dijumpai di dalam kamus akan dijadikan sebagai kata dasar. Terdapat perbagai jenis kamus yang digunakan seperti kamus kata dasar yang mengandungi senarai lengkap semua kata dasar dan juga kamus tempatan yang mengandungi perkataan khusus bagi sesuatu subjek contohnya kamus sejarah. Peratus ketepatan pencantas bergantung kepada masukan kamus. Lebih tepat kamus membuat padanan ke atas kata cantasan bermakna lebih tinggi ketepatan yang akan dihasilkan oleh pencantas. Walau bagaimanapun untuk memastikan cantasan adalah tepat, kamus perlulah sentiasa dikemaskini untuk memastikan perkataan yang dicantas dapat ditemui dalam kamus (Kazem Taghva, Russek Beckley dan Mohammad Sadeh, 2005; Somayye Estahbani dan Reza Javi, 2011).

Selain itu, kajian ini juga telah membuktikan bahawa penggunaan pencantas perkataan Jawi mampu untuk membantu capaian dokumen Jawi. Dengan adanya pencantas Jawi carian ke atas dokumen Jawi boleh dilakukan dengan mudah tanpa perlu ditukarkan ke dalam ejaan Rumi. Selain itu pencantas Jawi juga penting untuk digunakan dalam proses transliterasi disebabkan ejaan Jawi yang semakin kurang dikuasai oleh generasi muda. Proses penyemakan ejaan bagi skrip Jawi juga dapat dilakukan secara automatik.

1.5 SKOP KAJIAN

Skop kajian yang dijalankan merangkumi kesemua imbuhan-imbuhan Bahasa Melayu dalam tulisan Jawi termasuk imbuhan awalan, apitan, akhiran dan sisipan yang digabungkan dengan kata dasar satu suku kata, dua suku kata dan tiga suku kata untuk membentuk kata terbitan yang ditransliterasi ke dalam skrip Jawi. Kesemua set data yang akan diuji diambil daripada artikel-artikel yang terkandung dalam Utusan Melayu dan ditukar dalam ejaan Jawi melalui proses transliterasi Rumi-Jawi menggunakan sistem TERUJA (2011) dan e-Jawi Converter (2011) untuk menghasilkan 1200 kata terbitan dalam Jawi. Berdasarkan kajian Yon Hendri (2009) TERUJA menghasilkan ketepatan sebanyak 70.7%. Seterusnya untuk melihat sejauh mana pencantas perkataan Jawi ini dapat membantu capaian dokumen, tafsir Al-Quran





seperti yang digunakan dalam kajian Fatimah (Fatimah Ahmad, 1995) dan Taufik (Muhammad Taufik, 2006) telah digunakan.

1.6 STRUKTUR ORGANISASI TESIS

Tesis ini telah ditulis dan disusun dalam lapan bab yang utama. BAB I dimulai dengan perbincangan mengenai latar belakang kajian yang dijalankan. Seterusnya bab ini juga membincangkan mengenai pernyataan masalah, objektif kajian, kepentingan kajian dan juga skop. Struktur tesis yang berikutnya dibincangkan dalam bab yang berikut.

Bab II menerangkan kepentingan pencantas perkataan serta jenis pencantas perkataan yang biasa digunakan. Pencantas-pencantas perkataan untuk bahasa lain seperti Bahasa Inggeris, Bahasa Perancis, Bahasa Arab, Bahasa Indonesia dan Bahasa Melayu turut dibincangkan dengan teliti. Selain itu jenis ralat yang ditemui dalam pencantas perkataan serta pengujian pencantas perkataan juga dimuatkan dalam bahagian ini.

Bab III membincangkan perbezaan yang wujud antara ejaan Rumi dan Jawi. Pembentukan perkataan melalui kata tunggal, kata terbitan, kata ganda dan juga partikel diterangkan dengan lengkap dalam bab ini.

Bab IV menerangkan metodologi tesis yang dijalankan. Perkara yang dibincangkan adalah berkaitan dengan reka bentuk kajian, kerangka kerja kajian dan reka bentuk eksperimen. Bab ini juga turut memuatkan sukatan prestasi, jenis ujian signifikansi dan alatan kajian yang digunakan sepanjang membangunkan tesis ini.

Bab V membincangkan proses yang terlibat dalam pembangunan petua pencantas perkataan melayu bagi aksara Jawi. Penghasilan petua pengesanan kesalahan ejaan (SEDR) dan petua nyah-imbuhan Jawi diterangkan dengan lebih lanjut. dalam bab ini. Eksperimen I dijalankan untuk menentukan turutan cantasan imbuhan yang bersesuaian dengan ejaan Jawi. Selain itu juga petua mengenai SEDR yang dihasilkan turut dinilai untuk menentukan peratus ketepatan petua SEDR. Hasil



dapatan daripada eksperimen dibincangkan dan digunakan untuk membangunkan algoritma cantasan seperti yang diterangkan dalam bab VI.

Bab VI menerangkan mengenai pembangunan algoritma pencantas perkataan Melayu bagi aksara Jawi. Algoritma pencantas yang telah dibangunkan kemudiannya dinilai dengan menentukan ketepatan cantasan yang dilakukan. Selain itu, penilaian menggunakan kaedah Paice (1990) dan Frakes & Fox (2003) turut dilakukan untuk melihat keberkesanan algoritma cantasan yang dihasilkan. Hasil eksperimen yang dijalankan turut dibincangkan.

Bab VII menerangkan eksperimen sistem capaian maklumat yang dijalankan untuk menguji sama ada algoritma cantasan yang dihasilkan dapat membantu capaian dokumen Jawi atau sebaliknya. Dalam ujian ini, set dokumen daripada tafsir Al-Quran yang sama seperti kajian Fatimah (Fatimah Ahmad, 1995) dan Taufik (Muhammad Taufik, 2006) telah digunakan. Ujian pasangan sampel-t juga turut dilakukan untuk menentukan sama ada terdapat perbezaan yang signifikan antara min purata ketepatan (MPK) bagi dokumen yang dicantas dengan dokumen yang tidak dicantas dan hasil eksperimen turut dibincangkan dalam bab ini.

Bab VIII menerangkan mengenai kesimpulan bagi kajian yang telah dijalankan. Bab ini juga turut membincangkan dapatan kajian, sumbangan kajian serta cadangan perluasan kerja kajian.

1.7 KESIMPULAN

Bab ini menyentuh secara ringkas dan padat mengenai objektif, pernyataan masalah, kepentingan kajian, skop dan struktur organisasi mengenai kajian. Setiap aspek penting dibincangkan dalam bab ini. Bab ini mengemukakan gambaran menyeluruh mengenai kajian yang dijalankan dan membantu para penyelidik yang ingin memahami kajian ini.

Objektif bagi kajian ini adalah untuk menghasilkan petua penyemakan kesalahan ejaan (SEDR) yang dapat menyemak ejaan kata dasar bagi aksara Jawi,



menghasilkan petua nyah-imbuhan yang digunakan untuk mencantas kata terbitan Jawi dan juga membangunkan algoritma pencantas perkataan Bahasa Melayu bagi Aksara Jawi yang dapat mencantas kata terbitan Jawi kepada kata dasar. Untuk memastikan kesemua objektif dapat dicapai, kajian untuk menghasilkan petua SEDR serta petua nyah imbuhan yang baru telah dilakukan dengan melihat penggunaan setiap petua menerusi buku-buku seperti “Panduan Membaca dan Menulis Jawi”, “Tatabahasa Dewan”, ditulis oleh Nik Safiah Karim, Farid M. Onn, Hashim Haji Musa dan Abdul Hamid Mahmood terbitan Dewan Bahasa dan Pustaka, “The Morphology of Malay”, “Daftar Ejaan Rumi Jawi” dan “Daftar Kata Bahasa Melayu Rumi-Sebutan-Jawi” terbitan Dewan Bahasa dan Pustaka. Seterusnya enam eksperimen telah dilakukan untuk menentukan objektif telah dicapai.

