PENCANTAS PERKATAAN BAHASA MELAYU BAGI MENGHASILKAN KATA DASAR DENGAN MENGGUNAKAN KAEDAH PERATURAN

Nur Hazrina Asyikin binti Hamdan, Nazlia Omar

[1,2]*Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 UKM Bangi,, Selangor Darul Ehsan, Malaysia*

Abstrak

Pencantas perkataan merupakan salah satu teknik yang membuang mahupun memisahkan kata imbuhan pada sesuatu perkataan bagi menghasilkan perkataan dasar, atau dikenali sebagai perkataan asalnya. Pencantas perkataan Bahasa Melayu akan memfokuskan kepada perkataan yang mempunyai kata imbuhan pada perkataan dasarnya bagi membentuk ayat yang lebih gramatis. Dengan adanya pencantas perkataan ini, perkataan yang mempunyai imbuhan akan menghasilkan semula kata dasar tersebut. Saban hari, teknologi semakin lama semakin berkembang. Namun begitu, masih lagi tiada pencantas bagi perkataan bahasa Melayu yang lengkap dengan baik yang boleh membantu di dalam pemprosesan teks bahasa Melayu. Projek ini bertujuan membangunkan pencantas bahasa Melayu berasaskan kaedah peraturan. Projek ini menfokuskan kepada tiga kategori sahaja iaitu imbuhan awalan, imbuhan akhiran dan imbuhan apitan dimana projek ini akan mencantas atau memisahkan perkataan imbuhan bagi menghasilkan kata dasar tersebut. Bahasa Pengaturcaraan Python akan digunakan bagi melaksanakan projek ini. Metode projek ini terbahagi kepada beberapa Peringkat. Peringkat yang pertama adalah pengumpulan data. Peringkat seterusnya adalah pra-pemprosesan, diikuti dengan penyusunan peraturan dan pembangunan algoritma dan akhir sekali adalah Peringkat pengujian dan analisa. Dengan adanya pencantas perkataan Bahasa Melayu ini, ianya diharapkan boleh membantu pemprosesan teks dalam pelbagai aplikasi seperti dalam pencarian maklumat dan terjemahan mesin.

Kata kunci: [Bahasa Melayu, pencantas, peraturan]

Pengenalan

Pencantas ini merupakan proses bagi menghasilkan pelbagai morfologi bagi kata dasar sesuatu perkataan. Kebiasaannya, pencantas ini dirujuk sebagai satu pencantas algoritma. Pencantas atau namanya di dalam bahasa inggeris iaitu “Stemming” juga merupakan salah satu teknik pemprosesan bahasa untuk mengurangkan perkataan tertentu iaitu imbuhan kepada akhiran dan awalan bagi menghasilkan atau membentuk kata dasar bagi perkataan tersebut yang dikenali sebagai lemma. Pencantas ini juga turut membantu dalam prapremprosesan teks, perkataan dan juga dokumen untuk penormalan teks. Selain itu, peranan pencantas ini amat penting dalam pemahaman bahasa semula jadi (NLU) dan juga pemprosesan bahasa semula jadi (NLP).

Selain itu, pencantas adalah sebahagian daripada kajian linguistik yang berperanan dalam pencarian dan pengekstrakan maklumat morfologi dan juga kecerdasan buatan (AI). Pencentas dan kecerdasan buatan (AI) mengekstrak maklumat yang penting dan bermakna daripada sumber yang luas seperti big data atau internet. Hal ini disebabkan corak bentuk perkataan tambahan yang berkaitan dengan subjek mungkin perlu dicari untuk mendapatkan hasil yang terbaik.

Pada tahun 1968, seorang cendekiawan bernama Julie Beth Lovins merupakan orang pertama yang menerbitkan idea pencantas ini, khususnya untuk Bahasa Inggeris. Pencantas bagi Bahasa Inggeris telah banyak dibangunkan sejak awal pencantas ini diterbitkan. Pada masa yang sama, pencantas bagi Bahasa Melayu turut dihasilkan oleh salah seorang cendekiawan berbangsa Melayu iaitu Asim Othman pada tahun 1993. Walaupun begitu, pencantas perkataan terutamanya Bahasa Melayu yang telah dihasilkan oleh para penyelidik masih lagi tidak mencapai keputusan atau hasil yang tepat dan sahih. Oleh hal yang demikian, peluang bagi penambahbaikan pencentas perkataan Bahasa Melayu diberikan kepada para penyelidik yang baharu. Penyelidik baharu boleh membuat penambahbaikan kepada pencantas perkataan yang digunakan pada zaman kini, bagi menghasilkan lagi keputusan yang lebih baik dan sahih serta memberi impak yang lebih baik untuk kegunaan pengguna.

Pernyataan Masalah

Di antara bahasa yang digunakan di pentas global, Bahasa Melayu merupakan salah satu bahasa yang agak sulit dari sudut perkataan di mana ia mempunyai banyak imbuhan yang ditambah pada sesuatu perkataan bagi membentuk satu perkataan dan maksud yang baharu serta menghasilkan penghasilan ayat yang lebih kompleks dan teratur. Imbuhan yang digunakan dalam perkataan Bahasa Melayu mempunyai perbezaan sama ada dari segi maksud atau penggunaan berbanding imbuhan yang digunakan dalam Bahasa Inggeris. Imbuhan dalam Bahasa Melayu terbahagi kepada tiga jenis iaitu imbuhan awalan, akhiran dan juga apitan.

Membina sebuah sistem atau aplikasi untuk sesebuah pencantas perkataan khususnya Bahasa Melayu memerlukan strategi yang penting serta perancangan yang teliti. Hal ini disebabkan, perkara ini bukanlah sesuatu yang boleh dianggap mudah. Setakat ini, tiada lagi peraturan yang dispesifikkan untuk mencantas imbuhan yang ada pada perkataan tersebut untuk menghasilkan kata dasar itu. Selain itu, pencantas yang sedia ada merupakan pencantas yang tidak lengkap dan memerlukan penambahbaikan serta menambah lagi beberapa maklumat penting bagi memberi hasil yang baik kepada para pengguna. Kebanyakan pencantas Bahasa Melayu yang digunapakai pada masa kini tidak memberi hasil yang baik. Bukan itu sahaja, malah pencantas yang sekarang menggunakan peraturan yang berkonsepkan kamus sebagai rujukan.

Akhir sekali, bagi mendapatkan hasil yang baik, peraturan yang digunapakai dalam pada zaman kini perlu ditambahbaik. Hal ini bagi memastikan perkataan yang mempunyai imbuhan sama ada awalan, akhiran atau apitan dapat dicantas dengan lebih tepat dan baik. Perkara ini juga mampu mengelak daripada berlakunya ralat sama ada terlebih cantas atau terkurang cantas yang boleh memberi

kesilapan pada akhir hasil sistem dimana kata dasar tidak dapat dikeluarkan oleh sistem dengan lebih tepat.

Cadangan Penyelesaian

Cadangan penyelesaian yang akan dilaksanakan melalui projek ini adalah dengan mengurangkan penggunaan kamus. Selain itu, kajian perlu dibuat dan penting dalam mendapatkan maklumat berkenaan dengan pencantas perkataan Bahasa Melayu yang pernah dilakukan oleh penyelidik-penyelidik yang lepas. Hal ini untuk mengelakkan sebarang kesilapan yang boleh berlaku. Peraturan yang sedia ada akan ditambah baik dan peraturan-peraturan yang baharu dan penting juga akan dimasukkan dalam pencantas Bahasa Melayu jika perlu, bagi menghasilkan *output* yang lebih tepat dan lengkap.

Objektif Kajian

Objektif utama projek ini adalah:

1. Menambah peraturan yang baru dan sesuai bagi mencantas perkataan Bahasa Melayu dengan lebih sahih dan tepat.
2. Mengurangkan penggunaan kamus melalui peraturan yang dibina dalam sistem pencantas dan ditambah baik.
3. Menguji dan menilai keberkesanan dari aspek ketepatan bagi hasil akhir selepas pelaksanaan mencantas perkataan Bahasa Melayu yang telah ditambah baik.

Pengarang, Tahun	Objektif	Kaedah	Kajian	Ketepatan
Asim Othman (1993)	Membina algoritma pencantas bagi Bahasa Melayu	Kaedah Berasaskan Peraturan *(Rule Based Approach)*	Mencantas kesemua imbuhan Bahasa Melayu, namun mempunyai had panjang minimum bagi sesuatu perkataan dasar	Berjaya mencatat peratus ketepatan sebanyak 97%
Fatimah (1995)	Mencantas pelbagai perkataan berimbuhan pada kata dasar yang bertujuan untuk pengindeksan dan capaian dokumen Melayu.	Peraturan Susunan Aplikasi *(Rules Application Order)*	Menguji algoritma dengan terjemahan Al-Quran.	Hasil keputusannya adalah 61.35%. Menghasilkan beberapa ralat seperti terlebih cantas sebanyak 18.2%, ralat terkurang cantas sebanyak 4.6%, ralat tiada perubahan sebanyak 22.7%, ralat pengecualian ejaan sebanyak 4.6% dan ralat lain sebanyak 50%
Norisma Idris (2001)	Menghasilkan algoritma pencantas yang bersesuaian dengan Bahasa Melayu.	Menggunakan asas daripada algoritma Fatimah (1995)	Mencantas imbuhan Bahasa Melayu yang ada pada sesuatu perkataan	Mencapai ketepatan sebanyak 90%

Taufik (2009)	Menambahbaik petua '*Rule Application Order'* bagi mengurangkan kesilapan semasa proses cantasan berlaku	Peraturan Susunan Aplikasi *(Rules Application Order)*	Menambah baik pencantas Fatimah dan memperkenalkan pencantas baru iaitu *Rule Frequency Order* (RFO).	Berjaya menghasilkan satu peraturan ketepatan yang tinggi dengan menggunakan data yang sama dalam algoritma pencantas Fatimah (1995).

Jadual 1 Sorotan Susastera

Jadual 1 memaparkan sorotan susastera berkenaan dengan kajian-kajian lepas yang telah dijalankan oleh para penyelidik berkenaan dengan tajuk kajian.

Metodologi Kajian

Penggunaan model pembangunan yang relevan adalah signifikan bagi memastikan pelancaran sebuah kajian itu berjalan dengan lancar dan menghasilkan sebuah kerja yang berpotensi dan berprestasi tinggi. Kajian ini terdiri daripada empat peringkat iaitu Peringkat Pernyataan Masalah, Peringkat Pengumpulan Data, Peringkat Pembangunan dan Peringkat Pengujian.

1. Peringkat Pernyataan Masalah

Antara item-item yang dilaksanakan dalam Peringkat ini adalah:

i) Sorotan susastera
ii) Mengenalpasti masalah
iii) Mengenalpasti objektif kajian
iv) Perancangan kajian

Item-item yang disenaraikan di atas merupakan pelan perancangan projek pada awal kajian. Peringkat ini dimulakan dengan sorotan susastera bagi memahami berkenaan dengan tajuk kajian yang dilakukan sebelum ini dengan lebih terperinci. Sorotan susastera ini juga memberi maklumat tambahan dalam mengenalpasti metodologi kajian dengan lebih jelas. Berdasarkan sorotan susastera, isu-isu yang dikenalpasti akan diambil bagi mengenalpasti masalah yang dihadapi semasa pembangunan kajian ini dijalankan. Melalui masalah kajian tersebut, objektif kajian akan dikesan. Dalam perancangan kajian pula, proses dan metod untuk membina kajian ini akan dirancang dengan baik.

2. Peringkat Pengumpulan Data

Projek ini mempunyai dua jenis data set yang berbeza dimana ia akan digunakan sepanjang proses pembangunan kajian ini, iaitu data set latihan dan juga data set ujian. Data yang dilabel sebagai data latihan diambil secara rawak daripada artikel-artikel yang dipaparkan secara atas talian di

laman sesawang Awani. Data ujian pula diambil secara rawak daripada artikel-artikel di laman sesawang Kosmo. Keseimbangan kedua-dua data ini disusun kepada 80% data latihan dan selebihnya iaitu 20% kepada data ujian. Data set latihan ini berperanan untuk digunakan dalam pembangunan prototaip. Data set latihan ini merupakan data set yang melatih prototapi berdasarkan peraturan-peraturan yang dibina. Selepas prototaip berjaya dihasilkan, data set ujian akan digunakan bagi meguji potensi dan keberkesanan prototaip yang telah dilatih menggunakan data set latihan. Data set ujian yang digunakan mempunyai data set yang berbeza dengan data set latihan.

Sebelum proses pencantasan perkataan, perkataan itu perlu diproses menggunakan teknik pemprosesan bahasa tabii. Perkataan akan melalui proses tokenisasi (*tokenization)* agar perkataan tersebut disusun secara berasingan dan menjadi kata tunggal. Selepas proses tokenisasi, penormalan teks (*normalization)* akan dilaksanakan terhadap perkataan-perkataan tersebut untuk menjadi huruf kecil. Kemudian, penyingkiran tanda baca akan dibuat. Tanda baca merupakan simbol yang digunakan dalam suatu ayat seperti “-,!.:;”

3. Peringkat Pembangunan

Peringkat ini merupakan Peringkat dimana prototaip dan antara muka dibina. Data set latihan akan digunakan bagi melaksanakan pembangunan terhadap prototaip tersebut. Terdapat dua bahagian utama iaitu pembangunan peraturan dan pembangunan algoritma pencantas perkataan Bahasa Melayu.

4. Peringkat Pengujian

Peringkat pengujian merupakan Peringkat yang terakhir bertujuan untuk menjalankan pengujian dan penilaian keberkesanan ke atas algoritma yang telah dibina. Prestasi prototaip akan dicatat

meggunakan kejituan (*precision).* Kejituan adalah ukuran peratus ketepatan terhadap maklumat yang diperolehi adalah tepat.

Keputusan dan Perbincangan

Dalam kajian ini, algoritma ini akan dibentuk menjadi salah satu program yang mempunyai antara muka bagi memudahkan para pengguna mengakses algoritma ini tanpa melihat kod di belakang sistem ini. Antara muka atau dipanggil sebagai ‘dashboard’ akan digunakan menggunakan flask. Visual Studio Code telah dipilih untuk menjadi platform bagi membina algoritma ini.

Rajah 1 Antara muka pencantas Bahasa Melayu

Bagi mencapai keputusan, perancangan pengujian dibuat untuk menguji algoritma yang dibangunkan menggunakan data-data yang dicapai di akhbar Kosmo secara atas talian di laman web sesawang https://www.kosmo.com.my/. Data yang diambil di laman web sesawang Kosmo ini adalah secara rawak dan diambil daripada artikel-artikel yang dipaparkan di laman web sesawang tersebut. Tiga jenis ujian yang akan diteliti melalui perancangan pengujian ini, iaitu imbuhan awalan, imbuhan akhiran dan imbuhan apitan.

1. Eksperimen I (Mencantas Imbuhan Awalan)

Ujian pertama iaitu eksperimen 1 bagi pencantas Bahasa Melayu hanya memfokuskan kepada imbuhan awalan sahaja. Terdapat beberapa proses yang perlu diambil kira dalam membangunkan

peraturan imbuhan awalan iaitu proses penambahan, penggantian dan pembuangan huruf jika perlu pada suatu kata dasar tersebut.

Terdapat dua jenis imbuhan awalan yang berkongsi peraturan yang sama iaitu meN- dan peN-. Hal ini kerana berlakunya penambahan huruf T kepada perkataan yang telah dicantas sekiranya huruf awal kata dasar tersebut dimulai dengan huruf T.

Contohnya: menarik --> tarik

Seterusnya ialah peraturan imbuhan awalan bagi meM- dan peM-. Kedua-dua imbuhan awalan ini mempunyai peraturan yang sama dimana berlakunya penambahan huruf P dan T selepas awalan meM- dan peM- dicantas. Walaubagaimanapun, peraturan ini didapati agak sukar untuk menetapkan sama ada huruf P atau T yang perlu ditambah.

Contohnya: memukul --> pukul atau tukul.

Imbuhan awalan menG- dan penG- adalah peraturan yang ketiga. Jika kata dasar yang bermula dengan huruf A dan dieja bersama dengan dua imbuhan tersebut, maka imbuhan awalan itu akan dibuang. Dalam proses pencantasan itu, huruf k perlu ditambah bagi perkataan yang bersesuaian dengan peraturan yang dibina.

Contohnya: mengamuk --> amuk

mengecut --> kecut

Peraturan imbuhan awalan yang terakhir ialah menY- dan penY-. Sekiranya perkataan yang dimasukkan oleh pengguna mempunyai imbuhan awalan meny- dan peny-, maka imbuhan awalan tersebut akan dicantas berdasarkan peraturan di atas dan berlakunya penambahan huruf iaitu huruf S di awal perkataan yang dicantas. Jadual 6.1 menunjukkan beberapa hasil akhir selepas berlakunya proses pencantasan perkataan imbuhan awalan serta kata dasar yang betul.

Perkataan	**Kata Dasar**	**Hasil Proses Pencantasan**
berlari	lari	lari

mengetuk	ketuk	ketuk
melakar	lakar	lakar
mengecil	kecil	kecil
sepanjang	panjang	panjang

Jadual 2 Hasil proses pencantasan perkataan imbuhan awalan

2. Eksperimen II (Mencantas Imbuhan Akhiran)

Bagi mencantas imbuhan akhiran yang ada pada hujung atau akhir perkataan, maka peraturan imbuhan akhiran untuk melaksanakan pencantasan tersebut. Sewaktu proses cantasan akhiran berlaku, kata dasar perlu dititikberatkan sama ada ianya perlu dicantas ataupun dikekalkan seperti yang sedia ada. Oleh hal yang demikian, peraturan imbuhan akhiran yang dibina boleh digunakan ke atas beberapa imbuhan akhiran seperti -lah, -i, -an, -kan dan -nya. Hasil proses pencantasan perkataan terhadap imbuhan akhiran serta kata dasar yang betul ditunjukkan dalam Jadual 3.

Perkataan	**Kata Dasar**	**Hasil Proses Pencantasan**
hidangan	hidang	hidang
bukunya	buku	buku
buktikan	bukti	bukti

larilah	lari	lari

Jadual 3 Hasil proses pencantasan perkataan imbuhan akhiran

3. Eksperimen III (Mencantas Imbuhan Apitan)

Imbuhan apitan ini merupakan gabungan diantara imbuhan awalan dan imbuhan akhiran yang ada pada sesuatu perkataan tersebut dan membentuk satu perkataan terbitan. Oleh itu, peraturan imbuhan apitan ini melibatkan pencantasan imbuhan daripada bahagian hadapan perkataan dan belakang perkataan untuk melaksanakan pencantasan. Jadual 4 menjelaskan beberapa hasil proses pencantasan perkataan imbuhan apitan serta kata dasar yang betul.

Contohnya:menghiburkan --> hibur

Perkataan	**Kata Dasar**	**Hasil Proses Pencantasan**
pertandingan	tanding	tanding
menyekolahkan	sekolah	sekolah
penghargaan	harga	harga

Jadual 4 Hasil proses pencantasan perkataan imbuhan apitan

Kesimpulan

Pencantas perkataan Bahasa Melayu yang dihasilkan ini adalah penting untuk digunakan dalam semua capaian dokumen dan maklumat. Bukan itu sahaja, malah penghasilan pencantas perkataan Bahasa Melayu yang dibangunkan ini adalah satu satu penambahbaikan dan salah satu projek baharu yang dapat membantu pengguna untuk membuat pencantasan perkataan Bahasa Melayu mengikut kamus Dewan Bahasa dan Pustaka yang bersesuaian mengikut teknologi dan peredaran zaman berbanding dengan pencantasan perkataan Bahasa Melayu yang sedia ada.

Objektif utama kertas projek dan projek ini dibangunkan adalah bagi menghasilkan satu pencantas perkataan yang memfokuskan hanya kepada Bahasa Melayu agar dapat mengeluarkan kata terbitan ke dalam bentuk kata dasar mengikut kamus Dewan Bahasa dan Pustaka. Projek ini memberi fokus sepenuhnya untuk membangunkan tiga peraturan sahaja iaitu peraturan imbuhan awalan, peraturan imbuhan akhiran dan peraturan imbuhan apitan. Harapan yang tinggi terhadap penghasilan pencantas ini agar ianya dapat membantu dalam melaksanakan proses capaian dokumen dan maklumat pada era kini.

Pencantas perkataan bagi Bahasa Melayu yang dibangunkan ini mempunyai beberapa kelebihan dan keutamaan yang tersendiri berbanding dengan pencantas yang telah dibina sebelumnya. Antara karakteristik yang ada pada pencantas perkataan bagi Bahasa Melayu adalah seperti yang tertera:

1. Peraturan dan algoritma yang disusun dan diimplemen dalam pencantas ini lebih ringkas dan mudah untuk difahami.

2. Tidak hanya memfokuskan kepada penggunaan kamus dimana ia juga turut menghasilkan hasil proses akhir yang memuaskan.

3. Mempunyai antaramuka pengguna yang mudah menarik perhatian pengguna dan cara penggunaannya yang mudah difahami.

4. Selain mengadakan pencantasan daripada pendekatan peraturan, turut disediakan API AI (mesin kecerdasan bagi membandingkan keputusan dengan pencantas yang berasaskan peraturan

Setiap projek yang dibangunkan oleh manusia pasti ada kelemahannya. Begitu juga dengan pencantas perkataan Bahasa Melayu yang dibangunkan dalam projek ini dimana ia turut mempunyai beberapa kelemahan yang tidak dapat dielakkan. Antaranya ialah:

1. Terdapat beberapa peraturan dan algoritma yang digunakan masih tidak dapat menghasilkan hasil cantasan yang betul bagi beberapa perkataan.

2. Berlakunya isu terlebih cantas dan terkurang cantas terhadap hasil akhir projek.

3. Hasil akhir berkemungkinan boleh menghasilkan kata dasar yang membawa maksud lain. Contohnya adalah "seksa" akan menjadi "seks", "sekolah" akan menjadi "seko" dan "lelaki" akan menjadi "lelak".

Dalam membangunkan membangun pencantas perkataan Bahasa Melayu ini, terdapat beberapa kekangan yang tidak dapat dielakkan. Kekangan yang berlaku menyukarkan masa dan proses dalam membangunkan pencantas perkataan ini. Antara kekangan yang dihadapi adalah kekangan data, kekangan masa dan juga kekangan dalam mencari sumber rujukan.

1. Kekangan data

Maksud kekangan data disini ialah kekangan data dimana data yang diperolehi dari laman web sesawang Kosmo dan Astro Awani perlu disemak terlebih dahulu. Hal ini kerana mungkin akan mengganggu hasil kajian dan ianya perlu dilakukan secara manual. Sebagai contoh, terdapat beberapa perkataan yang digunakan dalam bahasa pasar seperti "weh", "tauke" dan lain-lain.

Selain itu, perlu juga membuat penyemakan bahasa yang ada pada setiap data. Maksudnya disini ialah membuat semakan data sekiranya ayat yang diambil dari laman web sesawang itu mempunyai dua bahasa yang berbeza. Contohnya, "*Manchester City*", "*Mobile Legend"* dan banyak lagi.

2. Kekangan masa

Kekangan masa ini berlaku apabila projek akhir ini dilaksanakan dalam semester 2 pada tahun 3, dimana banyak tugasan dan projek lain yang perlu disiapkan dalam tempoh yang sama dalam menyiapkan projek akhir ini.

3. Kekangan dalam mencari sumber rujukan

Kekangan yang terakhir adalah kekangan dalam mencari sumber rujukan. Sumber rujukan yang sedia ada di dalam internet dan spesifik untuk pencantas Bahasa Melayu merupakan kajian yang sudah lama dibangunkan, contohnya kajian pada tahun 1990, tahun 2000 dan sebagainya. Kurang sumber rujukan yang dipaparkan oleh internet yang bermula pada 2018 hingga 2023. Hal ini menyebabkan projek ini terpaksa mencari sumber rujukan lain iaitu sumber rujukan dari bahasa lain. Sebagai contoh, pencantas bagi bahasa Arab, pencantas bahasa Perancis dan lain-lain bagi menjayakan projek akhir ini.

Terdapat beberapa cadangan untuk menjayakan kajian masa hadapan bagi pencantas perkataan Bahasa Melayu bagi menghasilkan kata dasar dengan menggunakan kaedah peraturan ini. Antara cadangan yang boleh dilakukan bagi melaksanakan pencantas perkataan yang lebih tepat dan jitu adalah seperti berikut:

i. Membuat penambahan perkataan di dalam kamus yang sedia ada.

ii. Membuat penambahan peraturan dari kata pinjaman daripada bahasa asing seperti Bahasa Inggeris, bahasa Arab dan bahasa Indonesia.

Pembangunan projek ini iaitu Pencantas Perkataan Bahasa Melayu bagi menghasilkan kata dasar dengan menggunakan kaedah peraturan ini dibina untuk menghasil, membangun dan menilai penggunaan pencantas Bahasa Melayu untuk mencantas imbuhan-imbuhan yang ada pada sesuatu perkataan bagi menerbitkan kata dasar. Pencantas ini bukanlah salah satu komponen yang baru diwujudkan terutamanya dalam bidang penterjemahan, bahasa, capaian maklumat dan dokumen. Oleh itu, kajian ini mempunyai potensi dalam memajukan dan mempertingkatkan lagi penggunaannya pada masa akan datang mengikut kesesuaian teknologi. Kelebihan dan juga kekurangan yang ada dalam pencantas ini sewaktu proses pembangunan pencantas telah dijelaskan dengan lebih terperinci dalam bab ini. Kelebihan yang ada diharapkan dapat memberi manfaat yang baik kepada mereka yang menggunakan pencantas ini dan kepada mereka yang memerlukan kajian ini. Diharapkan juga kekurangan yang ada pada pencantas ini dapat diperbaiki, diolah dan ditambahbaik untuk masa hadapan bagi meningkatkan lagi kualiti pencantas ini.

Penghargaan

Alhamdulillah, bersyukur ke hadrat Illahi kerana dengan izin dan berkat-Nya dapatlah saya menyiapkan laporan penyelidikan bagi memenuhi syarat Ijazah Sarjana Muda Sains Komputer dengan Kepujian dalam tempoh masa yang ditetapkan. Selain itu, saya juga bersyukur kerana segala masalah dan pelbagai kekangan yang timbul sepanjang kajian ini dijalankan berjaya diatasi tanpa sebarang sekatan. Semua permasalahn ini menjadikan saya sebagai seorang insan yang mengenal erti kesabaran dan manisnya sebuah kejayaan.

Setinggi-tinggi ucapan terima kasih dan sejuta penghargaan ditujukan kepada Prof. Madya. Dr Nazlia Omar selaku penyelia projek tahun akhir yang telah banyak memberi tunjuk ajar dan juga bimbingan kepada saya sepanjang kajian ini dilaksanakan. Beliau juga memberi banyak dorongan dan kata-kata semangat kepada saya bagi menjayakan penulisan kajian ini. Penghargaan dan ucapan ini juga turut ditujukan kepada Dr. Saidah Saad dan Dr. Sabrina Tiun yang membantu dalam memberi idea dan tunjuk ajar dalam sepanjang kajian ini. Tidak lupa juga setinggi-tinggi penghargaan dan ucapan terima kasih kepada keluarga yang amat disayangi, Hamdan bin Mohd Ali (Ayahanda), Zarinah binti Ismail (Bonda) dan kakak tersayang iaitu Nur Sabrina binti Hamdan yang tidak putus-putus mendoakan dan memberi dorongan serta semangat untuk meneruskan penyelidikan ini.

Di samping itu, saya juga turut ingin mengucapkan terima kasih kepada rakan-rakan seperjuangan saya iaitu Aisyah Hanis, Nurin Nabilah, Afiq Aiman, Nuruddin Naim, Izqalan, Muhammad Abu Bakar As Siddiq, Nurbalqis Qistina, Syafiah Iman dan Rabindranath atas segala pertolongan, dorongan dan sokongan moral kalian yang tidak berbelah bahagi. Tak lupa juga turut ingin mengucapkan kepada rakan-rakan fakulti Haziq Ruslan, Aina Athirah, Nur Samahati, Nur Syahirah Nabilah dan Muhammad Amiruddin atas segala sokongan yang diberikan dalam

membantu kajian ini. Ribuan terima kasih diucapkan kepada rakan jauh iaitu William Lim yang memberi semangat sepanjang kajian ini dijalankan.

Sekalung penghargaan dan sekalung budi buat semua yang terlibat di Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan malaysia sepanjang pengajian saya di sini.

Last but not least, I wanna thank me. I wanna thank me for believing in me. I wanna thank me for doing all this hard work. I wanna thank me for having no days off. I wanna thank me for never quitting.

RUJUKAN

Abdullah, M. T. [Muhammad T. A., Ahmad, F. [Fatimah A., Mahmod, R. [Ramlan M., & Tengku Sembok, T. M. [Tengku M. T. S. (2009). Rules Frequency Order Stemmer for Malay Language. *International Journal of Computer Science and Network Securirty*, *9*(2).

AlSerhan, H. M., Alqrainy, S., & Ayesh, A. (2008). Is paice method suitable for evaluating Arabic stemming algorithms? *2008 International Conference on Computer Engineering &Amp; Systems*. https://doi.org/10.1109/icces.2008.4772981

Anjali G.Jivani (2011). A Comparative Study of Stemming Algorithms. *International Journal of Computer Technology and Applications*, *02*(06), 1930–1938.

Asim Othman. 1993. *Pengakar perkataan melayu untuk sistem capaian dokumen*

automated Malaysian stemmer for the Malay language. Proceedings of the fifth

Bangi: Penerbit Universiti Kebangsaan Malaysia.

Chiranjibi Sitaula (2013). A Hybrid Algorithm for Stemming of Nepali Text. *Intelligent Information Management*, *05*(04), 136–139. https://doi.org/10.4236/iim.2013.54014

Chua, Y. (2011). Kaedah dan statistik penyelidikan: kaedah penyelidikan. *McGraw-Hill Education EBooks*. http://eprints.um.edu.my/10246/

Contributor, T. (2018, January 26). *stemming*. SearchEnterpriseAI. https://www.techtarget.com/searchenterpriseai/definition/stemming

Deepika Sharma (2012). Stemming Algorithms: A Comparative Study and their Analysis. *International Journal of Applied Information Systems*, *4*(3), 7–12. https://doi.org/10.5120/ijais12-450655

eksperimen & analisis. Bangi: Penerbit Universiti Kebangsaan Malaysia.

Fadzli, S. A., Norsalehen, A. K., Syarilla, I. A., Hasni, H., & Siti Dhalila, M. S. (2012). SIMPLE RULES MALAY STEMMER. *Faculty of Informatics, UnisZa*, 28–35.

Fatimah Ahmad (1995). *Sistem capaian dokumen Bahasa Melayu: satu pendekatan*

international workshop on Information retrieval with Asian languages, hlm:207 –

Ismail, N. K. Nurazzah A. R., Abu Bakar, Z. & Tengku Sembok, T. M. (2007). TERMS VISUALIZATION FOR MALAY TRANSLATED QURAN DOCUMENT. *Proceedings of the International Conference on Electrical Engineering and Informatics, 19.*

Ismailov, A., Jalil, M. A., Abdullah, Z., & Rahim, N. A. (2016). A comparative study of stemming algorithms for use with the Uzbek language. *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*. https://doi.org/10.1109/iccoins.2016.7783180

Jumadi, J., Maylawati, D. S., Pratiwi, L. D., & Ramdhani, M. A. (2021). Comparison of Nazief-Adriani and Paice-Husk algorithm for Indonesian text stemming process. *IOP Conference Series: Materials Science and Engineering*, *1098*(3), 032044. https://doi.org/10.1088/1757-899x/1098/3/032044

Karaa, W. B. A., & Gribâa, N. (2013). Information Retrieval with Porter Stemmer: A New Version for English. *Advances in Intelligent Systems and Computing*, 243–254. https://doi.org/10.1007/978-3-319-00951-3_24

Kean, H. A. (2016). Mengenai penyelidikan dan kajian kes: Satu tinjauan literatur. *Malaysian Journal of Society and Space*, *12*(10). https://doi.org/10.6084/m9.figshare.3803583.v1

Kraaij, W., & Pohlmann, R. (1994). Porter's stemming algorithm for Dutch. -.

Maheswari S,K.Arthi (2019). Rule Based Morphological Variation Removable Stemming Algorithm. *International Journal of Recent Technology and Engineering (IJRTTE)*, *8*(4). https://doi.org/10.35940/ijrte.c6200.118419

McNamee, P., & Mayfield, J. (2004). Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval*, *7*(1/2), 73–97. https://doi.org/10.1023/b:inrt.0000009441.78971.be

Melucci, M., & Orio, N. (2003). A novel method for stemmer generation based on hidden markov models. *Proceedings of the Twelfth International Conference on Information and Knowledge Management - CIKM '03*. https://doi.org/10.1145/956863.956889

Memon, S., Ali, G., -, K., Shaikh, A., K.Aasoori, S., & Ul, F. (2020). Comparative Study of Truncating and Statistical Stemming Algorithms. *International Journal of Advanced Computer Science and Applications*, *11*(2). https://doi.org/10.14569/ijacsa.2020.0110272

Muchtar, M. A., Nababan, E. B., Nababan, M., Andayani, U., Simanjuntak, T., & Sitompul, O. S. (2019a). Implementation of Porter Stemmer Algorithm to Obtain Basic Words in Toba Batak Language Documents with the Two-Level Morphological Method. *IOP Conference Series: Materials Science and Engineering*, *648*(1), 012025. https://doi.org/10.1088/1757-899x/648/1/012025

Muchtar, M. A., Nababan, E. B., Nababan, M., Andayani, U., Simanjuntak, T., & Sitompul, O. S. (2019b). Implementation of Porter Stemmer Algorithm to Obtain Basic Words in Toba Batak Language Documents with the Two-Level Morphological Method. *IOP Conference Series: Materials Science and Engineering*, *648*(1), 012025. https://doi.org/10.1088/1757-899x/648/1/012025

Muhammad Taufik Abdullah (2009). *Rules frequency order stemmer for malay language.* International Journal of Computer Science and Network Security 1(9):433-438

Norisma Idris, & Syed Mustapha. (2001). Stemming for Term Conflation in Malay Texts. *International Conference of Artificial Intelligence (ICAI2001)*.
Nur Anis Putri binti Zulkiflee (2020). *Pencantas Perkataan Dalam Bahasa Melayu Berasaskan Peraturan.* Bangi: Penerbit Universiti Kebangsaan Malaysia

Nur Anis Putri binti Zulkiflee (2020). PENCANTAS PERKATAAN DALAM BAHASA MELAYU BERASASKAN PERATURAN. *Faculty of Information Science and Technology, UKM*.

Prashant Sharma. (2022, September 1). *An Introduction to Stemming in Natural Language Processing*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/11/an-introduction-to-stemming-in-natural-language-processing/

Pravesh Koirala, & Aman Shakya. (2020). A Nepali Rule Based Stemmer and its performance on different NLP applications. *ArXiv: Computation and Language*.

Rohit Kansal, Vishal Goyal, & Gurpreet Singh Lehal. (2012). Rule Based Urdu Stemmer. *International Conference on Computational Linguistics*, 267–276.

Sock Yin Tai, Cheng Soon Ong & Noor Aida Abdullah (2000). *On designing an*

Stemming in Data Mining - Javatpoint. (n.d.). www.javatpoint.com. https://www.javatpoint.com/stemming-in-data-mining

Team, T. A. I. (2022, September 2). *Stemming: Porter Vs. Snowball Vs. Lancaster*. Towards AI. https://towardsai.net/p/l/stemming-porter-vs-snowball-vs-lancaster

Zakeri Rad, H., Tiun, S., & Saad, S. (2018). VBS Stemmer: A vocabulary-based stemmer. *International Journal of Engineering &Amp; Technology*, *7*(2.14), 551. https://doi.org/10.14419/ijet.v7i2.9192

Nur Hazrina Asyikin binti Hamdan (A190375)
Prof. Madya. Dr Nazlia Omar
Fakulti Teknologi & Sains Maklumat,
Universiti Kebangsaan Malaysia