# PENCANTAS PERKATAAN DALAM BAHASA MELAYU BERASASKAN PERATURAN

Nur Anis Putri Zulkiflee
Nazlia Omar

*Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia*

**ABSTRAK**

Pencantas perkataan merupakan satu teknik untuk membuang kata imbuhan pada perkataan dan menghasilkan kata dasar. Pencantas perkataan digunakan secara meluas dalam bidang capaian maklumat di mana terdapat kepelbagaian morfologi atau pembentukan kata pada sesuatu perkataan yang ingin dicapai. Bahasa Melayu juga mempunyai pelbagai kata imbuhan yang kompleks dan selari dengan kata dasarnya. Perkara ini menyebabkan banyak kesilapan dalam pencantas Bahasa Melayu yang terdahulu. Hal ini memerlukan penambahbaikan untuk kegunaan pada masa sekarang. Pembangunan pencantas perkataan bagi Bahasa Melayu dalam kajian ini dibangunkan menggunakan Bahasa Pengaturcaraan Python. Pencantas perkataan ini menggunakan lima fasa metodologi kajian iaitu fasa pengumpulan data, fasa pra-pemprosesan, fasa penyusunan peraturan dan pembangunan algoritma dan fasa pengujian dan analisa. Peraturan bagi pencantas perkataan ini akan disusun semula daripada hasil kajian yang lepas.



## 1 PENGENALAN

Pencantas merupakan salah satu teknik yang digunakan unutk mencantaskan kata imbuhan kepada kata dasar. Pencantas adalah sebahagian daripada kajian linguistik dalam morfologi dan kepintaran buatan (AI) dalam pengambilan dan pengekstrakan maklumat. Kegunaan pencantas perkataan tidak terhad kepada bidang capaian dokumen sahaja tetapi ia juga amat penting dalam bidang transliterasi, serta penyemakan ejaan. Pencantas dipandang sesuatu yang mudah namun hakikatnya ianya bukanlah suatu yang mudah dan proses yang bebas masalah.

Konsep pencantas perkataan telah diperkenalkan oleh Julie Beth Lovins pada tahun 1968(Lovins, 1968). Terdapat banyak pencantas perkataan yang telah dibina bagi Bahasa Inggeris sejak awal tahun penghasilannya begitu juga pencantas perkataan bagi Bahasa Melayu. Pencantas perkataan pertama bagi Bahasa Melayu telah dihasilkan oleh Asim Othman pada tahun 1993 (Asim, 1993). Namun, pencantas perkataan yang dihasilkan oleh pengkaji-pengkaji terdahulu masih lagi kurang tepat dan jitu. Ini membuka ruang kepada pengkaji-pengkaji baharu untuk melakukan menambah baik terhadap pencantas perkataan yang sedia ada agar proses pencantasan menjadi lagi tepat dan jitu sesuai untuk kegunaan semasa.

## 2 PENYATAAN MASALAH

Bahasa Melayu merupakan satu bahasa yang kompleks dalam aspek morfologi dan mempunyai banyak imbuhan. Imbuhan yang digunakan berbeza berbanding penggunaan imbuhan dalam Bahasa Inggeris. Bahasa Melayu mempunyai empat kumpulan imbuhan iaitu awalan, apitan dan akhiran.

Bagi membangunkan sesebuah pencantas perkataan bagi Bahasa Melayu bukanlah sesuatu yang mudah berbanding pencantas Bahasa Inggeris. Pelbagai fasa diperlukan untuk memastikan pencantas Bahasa Melayu dapat menghasilkan dengan tepat. Bagi Bahasa Melayu tiada peraturan yang khusus yang ditetapkan harus mencantas imbuhan yang mana dahulu walaupun perkataan terbitan yang sama. Selain itu, kebanyakan peraturan terdahulu menggunakan kamus sebagai rujukan.

Akhir sekali, peraturan perlu ditambahbaik untuk memastikan perkataan yang baru dicantas dapat ditemui untuk mengelak daripada berlakunya ralat terlebih cantas dan ralat terkurang cantas.

## 3 OBJEKTIF KAJIAN

Objektif utama kajian ini adalah untuk menambah peraturan bagi mencantas perkataan Bahasa Melayu, menguji keberkesanan mencantas dari segi kejituan setelah mencantas perkataan Bahasa Melayu yang ditambah baik dan mengurangkan kegunaan kamus melalui peraturan yang dibangunkan dan ditambahbaik.

## 4 METOD KAJIAN

Penggunaan model pembangunan yang sesuai penting untuk memastikan perjalanan kajian berjalan dengan lancar dan menjamin hasil kerja yang berkualiti. Kajian ini terdiri daripada empat fasa iaitu Fasa Pernyataan Masalah, Fasa Pengumpulan Data, Fasa Pembangunan dan Fasa Pengujian.

### 4.1 Fasa Pernyataan Masalah

Perkara-perkara yang dijalankan dalam fasa ini adalah:

i) Kajian literatur
ii) Mengenalpasti masalah
iii) Mengenalpasti objektif kajian
iv) Perancangan kajian

Langkah-langkah yang telah dinyatakan di atas merupakan perancangan projek pada awal kajian. Fasa ini bermula dengan kajian literatur untuk memahami berkenaan tajuk kajian yang dilakukan sebelum ini. Selain itu, kajian literatur dapat membantu untuk mengenalpasti metodologi kajian yang lebih jelas. Berdasarkan kaijan literatur, isu-isu yang dikenalpasti akan digunakan untuk mengenalpasti masalah yang dialami semasa membangunkan kajian ini. Seterusnya, berdasarkan masalah kajian tersebut, objektif-objektif kajian ini akan dikenalpasti. Dalam perancangan projek, proses dan metod untuk kajian akan dirancang.

## 4.2 Fasa Pengumpulan Data

Terdapat dua jenis set data yang akan digunakan dalam kajian ini, data set latihan dan juga data set ujian. Data yang digunakan diambil daripada akhbar Berita Harian secara atas talian di halaman www.beritaharian.com.my. Data latihan untuk kajian ini akan menggunakan 80% daripada set data manakala baki 20% akan digunakan sebagai data ujian. Data set latihan merupakan data yang digunakan dalam pembangunan prototaip tersebut. Ia adalah data yang melatih prototaip berdasarkan peraturan-peraturan yang dibangunkan. Setelah prototaip dibangunkan, data set ujian pula akan digunakan untuk menguji prestasi dan keberkesanan prototaip yang telah dilatih menggunakan data set latihan. Data set ujian yang digunakan adalah berlainan dengan data set latihan.

Sebelum mencantas perkataan, perkataan itu perlu diproses menggunakan teknik pemprosesan bahasa tabii. Pertama sekali, perkataan akan melalui proses tokenisasi (tokenization) agar menjadi kata tunggal. Selepas proses tokenisasi, penormalan teks (normalization) akan dilakukan terhadap perkataan-perkataan tersebut untuk menjadi huruf kecil. Selain itu, menyikirkan tanda baca. Tanda baca merupakan simbol yang digunakan dalam suatu ayat seperti “-, !, :”.

## 4.3 Fasa Pembangunan

Fasa ini adalah untuk membangunkan prototaip yang akan dilatih menggunakan data set latihan. Ia terdiri daripada 2 bahagian utama iaitu pembangunan peraturan dan pembangunan algoritma pencantas perkataan Bahasa Melayu.

### 4.3.1 Pembangunan Peraturan Imbuhan

Proses pembangunan peraturan imbuhan melibatkan proses pembentukan perkataan berdasarkan imbuhan. Imbuhan dibahagikan kepada tiga kategori utama iaitu peraturan imbuhan awalan, akhiran dan apitan. Imbuhan terdiri daripada morfem terikat manakala kata dasar merupakan morfem bebas yang boleh menerima imbuhan. Peraturan bagi pencantas perkataan Bahasa Melayu dihasilkan adalah untak memastikan setiap proses yang ada dalam sesebuah pencantas itu dapat dijalankan. Peraturan ini disusunkan di dalam sesebuah algoritma pencantas dan ianya mempengaruhi hasil cantasan yang dilakukan. Antara peraturan yang dicadangkan adalah peraturan imbuhan awalan, imbuhan akhiran dan imbuhan apitan.

### 4.3.2 Pembangunan Algoritma pencantas perkataan Bahasa Melayu

Pembangunan algoritma bagi pencantas Bahasa Melayu ini merupakan prosedur berkomputer yang akan mengurangkan kata terbitan ke dalam bentuk kata cantas atau kata dasar (Lovins, 1968). Algoritma pencantas ini juga boleh dikatakan sebagai proses untuk mendapatkan kata dasar daripada imbuhan.

Sebelum proses cantasan dilakukan, setiap perkataan akan melalui fasa pra-pemprosesan. Selepas itu, barulah proses cantasan bermula. Perkataan tersebut akan semak kata imbuhan, jika tiada perkataan tersebut terus keluarkan kata dasar. Jika ada semak peraturan akhiran dan semak peraturan awalan. Setiap kali semak peraturan akhiran dan awalan mestilah menyemak kamus dan imbuhan itu akan dicantas mengikut peraturan. Selepas itu, semak penggubahan ejaan sekiranya ada penambahan huruf atau pengurangan huruf, barulah keluarkan kata dasar dan proses tamat. Peraturan yang dibentuk akan diuji untuk memastikan ia berkualiti apabila diuji dengan set data latihan. Peraturan yang tidak berfungsi dengan baik akan dinilai semula dan dibaiki.

## 4.4 Fasa Pengujian

Fasa Pengujian adalah fasa terakhir yang bertujuan untuk menguji dan menilai keberkesanan pendekatan yang telah dicadangkan. Fasa ini dijalankan untuk menguji dan menilai keberkesanan pendekatan yang telah dicadangkan. Prestasi prototaip akan diukur menggunakan kejituan (*precision*). Kejituan ialah ukuran peratus ketepatan maklumat yang diperolehi adalah tepat.
Formula kejituan ialah:

Kejituan (*Precision*) = $\frac{h}{}$

## 5 HASIL KAJIAN

Sebanyak tiga eksperimen telah dijalankan untuk kajian ini. Kesemua eksperimen dijalankan dengan menggunakan set data latihan dan ujian.

Pencantas perkataan untuk eksperimen pertama hanya melibatkan peraturan imbuhan awalan sahaja. Peraturan ini mengambil kira proses penambahan, penggantian dan pembuangan huruf dalam kata dasar. Eksperimen kedua hanya melibatkan peraturan imbuhan akhiran sahaja dan eksperimen ketiga adalah imbuhan apitan iaitu mewakili imbuhan awalan dan akhiran.

Kejituan dokumen yang dicantas dengan betul mampu mencapai tahap 0.97. Pengujian dilakukan bagi mengenalpasti jika wujud masalah peraturan dan algoritma yang mungkin timbul dalam setiap dibangunkan dalam capaian maklumat. Setiap ralat yang ditemui harus dibetulkan bagi memastikan keberkesanan capaian maklumat tidak terjejas. Setiap pengujian dilakukan bagi memastikan capaian maklumat yang dibangun memenuhi keperluan pengguna, menepati objektif kajian dan setiap spesifikasi seperti yang dibincangkan di bab-bab sebelumnya.

## 6 KESIMPULAN

Kesimpulannya, pembangunan Pencantas Bahasa Melayu Berasaskan Peraturan ini dihasilkan adalah untuk menghasil, membangun dan menilai pencantas Bahasa Melayu yang diguna untuk mencantas kata terbitan kepada kata dasar. Pencantas ini bukanlah perkara baru terutamanya dalam bidang penterjemahan dan bahasa. Kelebihan yang ada diharapkan dapat membantu mereka yang memerlukan kajian ini dan juga kekurangan yang dimiliki pada pencantas ini dapat diperbaiki untuk masa hadapan.

## 7 RUJUKAN

Asim Othman. 1993. Pengakar perkataan melayu untuk sistem capaian dokumen. Tesis Sarjana, Jabatan Sains Komputer, Universiti Kebangsaan Malaysia.

Chiranjibi Sitaula .2013. A Hybrid Algorithm for Stemming of Nepali Text. *Intelligent Information Management*, 1(5): 136-139.

Fatimah Ahmad. 1995. Sistem capaian dokumen Bahasa Melayu: satu pendekatan eksperimen & analisis. Tesis Dr. Falsafah, Jabatan Sains Komputer, Universiti Kebangsaan Malaysia.

Jasmeet Singh, Vishal Gupta. 2016. A systematic review of text stemming techniques. *Artificial Intelligence Review*, 48(2): 157-217.

Lovins, J.B. 1968. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics* 11(1):22-31.

Mangalam Sankupellay, Subbu Valliappan. 2006. Malay-language stemmer. *Sunway Academic Journal* 3(1): 147 -153.

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

N Idris, "Automated Essay Grading System using Nearest Neighbour Technique in Information Retrieval," Universiti of Malaya, Kuala Lumpur,2001.

Smirnov I. 2008. Overview of stemming algorithms. *In: Mechanical Translation*, 52.

Sock Yin Tai, Cheng Soon Ong & Noor Aida Abdullah. 2000. *On designing an* automated Malaysian stemmer for the Malay language. *Proceedings of the fifth international workshop on Information retrieval with Asian languages*, hlm:207 –208.

Sulaina Sulaiman. 2013. Pencantas Perkataan Melayu Untuk Aksara Jawi Berasaskan Petua. Tesis Dr. Falsafah, Jabatan Sains Komputer, Universiti Kebangsaan Malaysia.

Syed Abdullah Fadzli, A Khairani Norsalehan, I. Ahmad Syarilla, Hassan Hasni, M Satar Siti Dahlia, Simple Rules Malay Stemmer. Faculty of Informatics Universiti Sultan ZainalAbidin (UniSZA). Hlm:1-8.