PENGECAMAN ENTITI NAMA (NER) MENGGUNAKAN WIKIPEDIA BAHASA MELAYU DALAM DOMAIN SEJARAH KEMERDEKAAN TANAH MELAYU

MUHAMMAD ARIF SYAMIL BIN MOHD RAHIMI
SAIDAH SAAD

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Setiap perkataan mempunyai kebolehan untuk menyimpan pelbagai makna tersurat dan tersirat bagi menyampaikan mesej atau pengajaran. Dalam konteks sastera, setiap komponen tatabahasa akan digabungkan menjadi sebuah ayat atau frasa dan boleh membawa makna yang baharu atau berlainan daripada kata asal. Dalam membentuk ayat atau kata yang sesuai, setiap perkataan perlu melalui proses pengekstrakan makna di mana terpisah subjek kata dengan maksud kata tersebut. Proses pengekstrakan makna perkataan yang efektif terutamannya dalam dokumen-dokumen teks Bahasa Melayu akan menjadi fokus kajian ini. Hal ini demikian kerana kajian berkaitan pengekstrakan maklumat Bahasa Melayu masih lagi dalam peringkat awal dan pendekatan ini hanya terhad kepada kajian para sasterawan dan pakar linguistik sahaja. Kajian ini akan mengambil teknik pendekatan yang sering digunakan secara berleluasa iaitu pengecaman entiti nama terhadap artikel yang ditulis dalam laman web Wikipedia. Segala teks yang dipilih berdasarkan subjek sejarah Tanah Melayu akan diperoleh serta ditafsir mengikut jenis-jenis entiti yang wujud dalam teks. Penemuan jenis entiti seperti manusia, lokasi dan organisasi akan digunakan bagi melatih satu model pembelajaran mesin untuk memahami tren serta struktur ayat yang bersesuaian. Model yang terlatih ini kemudian akan digunakan semula bagi membuat pengecaman entiti nama bagi dokumen teks Bahasa Melayu yang baru dan luar daripada data teks yang diperolehi daripada laman web Wikipedia. Hasil keputusan yang dapat dikumpulkan daripada kajian ini dapat menentukan pendekatan model yang paling baik dalam menjalankan pengekstrakan entiti nama terhadap teks Bahasa Melayu. Dapatan ini juga dapat membantu sasterawan untuk memahami bagaimana pemahaman struktur sesebuah ayat berbanding pemahaman manusia dan seterusnya menambah baik proses pengekstrakan maklumat dalam teks.

1 PENGENALAN

Bahasa Melayu merupakan bahasa lingua franca bagi negara-negara di Asia Tenggara seperti Malaysia, Brunei, Singapura. Kewujudan Bahasa Melayu yang lahir di bawah keluarga bahasa Austronesia telah dituturkan sejak 1000 tahun lalu, dan telah berkembang dari segi dialek mengikut negara masing-masing. Negara Indonesia, yang juga menuturkan bahasa Melayu tetapi telah mentransformasikan bahasa tersebut sehingga menjadi bahasa Indonesia di bawah rumpun yang sama. Bahasa Melayu kini mempunyai jumlah penutur sebanyak 30 juta dan lebih daripada 300 juta termasuk penutur bahasa Indonesia(Anon, 2021). Jelas di sini bahawa bahasa ini telah meliputi ramai penutur yang fasih daripada segenap budaya dari pelbagai negara, apatah lagi bahasa Melayu pernah dijadikan sebagai lingua franca atau bahasa perantaraan utama di seluruh dunia pada zaman kegemilangan Kesultanan Melayu Melaka. Bahasa Melayu masih berkembang sehingga kini dan mendapat perhatian segenap

masyarakat Barat dan Timur untuk mempelajari bahasa ini. Oleh sebab itu, perkembangan bahasa Melayu telah digiatkan oleh penutur-penutur fasih bukan sahaja melalui pengajaran secara lisan dan bercetak, malah secara digital seperti laman web dan aplikasi selari dengan mengharungi era globalisasi. Laman web yang terkenal seperti Wikipedia juga menyambut baik perkembangan ini dengan mencipta satu ensiklopedia yang diterbitkan pada 26 Oktober 2002 dan merupakan ensiklopedia dalam talian daripada versi asal iaitu Wikipedia. Ensiklopedia percuma ini juga adalah salah satu inisiatif bagi pihak Wikipedia untuk memberikan kemudahan akses kepada maklumat bagi pengguna yang tidak fasih dalam bahasa Inggeris dimana kebanyakkan data disimpan dan disunting oleh komuniti berdasarkan bahasa utama masing-masing dan berasaskan teknologi web 3.0.

Pencarian maklumat dalam bahasa Melayu kini lebih mudah dengan laman web ini. Namun begitu, proses pencarian maklumat dilihat tidak efisien dan kadangkala tidak memberikan makna signifikan kepada pencari. Oleh itu, proses ini boleh ditambahbaik dengan menggunakan kaedah pengkelasan bagi setiap perkataan. Salah satu teknik yang dipraktikkan pada masa kini adalah teknik pengecaman entiti nama. Teknik pengecaman entiti nama merupakan kaedah dalam mengenalpasti konsep utama yang dibincangkan di dalam sesuatu dokumen dan ini dapat memberikan kesimpulan atau menjelaskan intipati yang terkandung dalam dokumen tersebut. Teknik ini juga digunakan untuk mengekstrak serta melombong maklumat semantik yang telah ditetapkan daripada data teks dan bertanggungjawab untuk mengenalpasti konsep yang berkaitan dengan mentafsir makna teks dan mengklasifikasikan makna teks mengikut sesuatu set kategori.

Pelbagai pendekatan yang boleh diaplikasikan dalam proses pengecaman entiti nama iaitu dengan model-model terkini yang berjaya dalam mengekstrak entiti daripada setiap perkataan dalam dokumen teks. Antara model tersebut ialah model Bidirectional Encoder Representative for Transformer (BERT), FastFormer, ALBERT dan XLNET. Fokus projek ini adalah untuk menggunakan model-model ini dan membina satu model baharu berdasarkan pembelajaran mesin K Nearest Neighbors (KNN) bagi mencari pendekatan terbaik dalam mengekstrak entiti nama berdasarkan teks sejarah Tanah Melayu di dalam laman web Wikipedia.

2 PERNYATAAN MASALAH

Dalam projek ini, terdapat pernyataan masalah yang dapat dikenal pasti iaitu:

1. Bagaimana untuk membangunkan satu pendekatan yang dapat mengenalpasti entiti utama dalam bahasa Melayu berdasarkan dokumen-dokumen teks sejarah Tanah Melayu.
2. Apakah model terbaik yang boleh diimplementasikan dalam proses pengecaman entiti nama bagi dokumen teks sejarah Tanah Melayu

3 OBJEKTIF KAJIAN

Melalui pembacaan kajian yang dilakukan, terdapat beberapa solusi yang boleh digunakan bagi menyelesaikan masalah ini, iaitu:

1. Membangunan sistem pengekstrakan entiti nama bahasa Melayu secara automatik dan mudah.
2. Mengenalpasti model terbaik dalam mengklasifikasi dokumen teks sejarah Tanah Melayu mengikut entiti nama dan menghasilkan data hierarki yang merujuk kepada entiti nama tersebut.

4 METOD KAJIAN

Kajian ini dibangunkan menggunakan Model Tokokan (Incremental Model) yang mudah untuk difahami dan digunakan. Dengan menggunakan kaedah ini, setiap fasa akan dibahagikan kepada 6 fasa dan perlu dilengkapkan serta melakukan pengujian sebelum fasa seterusnya dimulakan. Setiap fasa akan dijadikan sebagai satu blok fasa dan dijalankan secara berperingkat bagi memastikan projek berjalan seperti yang dirancang.

4.1 Fasa Perancangan

Fasa ini merupakan fasa yang terpenting dalam pembangunan sistem. Fasa ini selari dengan pernyataan masalah di mana setiap komponen dalam sistem ini haruslah membawa kepada penyelesaian terhadap masalah tersebut. Fasa ini memberikan gambaran menyeluruh bagi sistem. Objektif dan kekangan bagi membangunkan sistem pengecaman entiti nama telah dikenalpasti dalam fasa ini. Cadangan penyelesaian bagi pernyataan masalah juga akan dikenalpasti untuk membantu proses analisis.

4.2 Fasa Analisis

Fasa ini menfokuskan kepada analisa keperluan sistem. Keperluan fungsian dan bukan fungsian sistem akan dikenalpasti untuk memudahkan proses reka bentuk sistem. Selain itu, analisis terhadap model yang dibangunkan sendiri iaitu model KNN dan MultiOutputClassifier akan dijalankan untuk meningkatkan pemahaman tentang cara pengekstrakan entiti nama itu dilaksanakan. Selain itu, pemerhatian terhadap model-model yang terdapat dalam modul Malaya dan diimplementasikan dalam sistem ini akan dijalankan untuk membuat perbandingan antara model bagi menentukan model terbaik dalam pengekstrakan entiti nama.

4.3 Fasa Reka Bentuk

Fasa ini menentukan senibina sistem.yang akan digunakan. Aliran fungsi sistem pengecaman entiti nama akan dibincangkan dalam fasa ini. Antara muka sistem akan dibangunkan dan dipastikan mengikut kesesuaian permasalahan kajian ini supaya objekif projek dapat dicapai.

4.4 Fasa Implementasi

Fasa ini membincangkan tentang aspek pembangunan dan implementasi sistem yang dibangunkan. Segala fungsi kod yang kecil akan dikumpulkan dan digabungkan untuk menjadi sebuah sistem besar yang menyeluruh dan dapat mencapai objektif kajian. Fasa ini penting dalam menentukan kelemahan sistem selepas proses reka bentuk.

4.5 Fasa Pengujian

Sistem ini akan diuji sama ada dapat mencapai objektif atau sebaliknya. Sistem ini akan diuji melalui nilai-nilai statistik yang dibangunkan dalam kod program bagi melihat keberkesanan model KNN dan MultiOutputClassifier berbanding dengan model-model lain dalam modul Malaya. Maklum balas bagi kemudahan sistem dalam mengekstrak entiti nama akan diperolehi bagi melihat bagaimana sistem ini mesra pengguna dan dapat difahami ketika pengguna berinteraksi dengan sistem ini.

5 HASIL KAJIAN

Hasil pengujian dalam projek ini akan melibatkan perbandingan nilai-nilai statistik antara setiap model bagi menentukan model terbaik yang boleh digunakan untuk proses pengecaman entiti nama dalam dokumen teks sejarah Tanah Melayu. Nilai-nilai statistik bagi

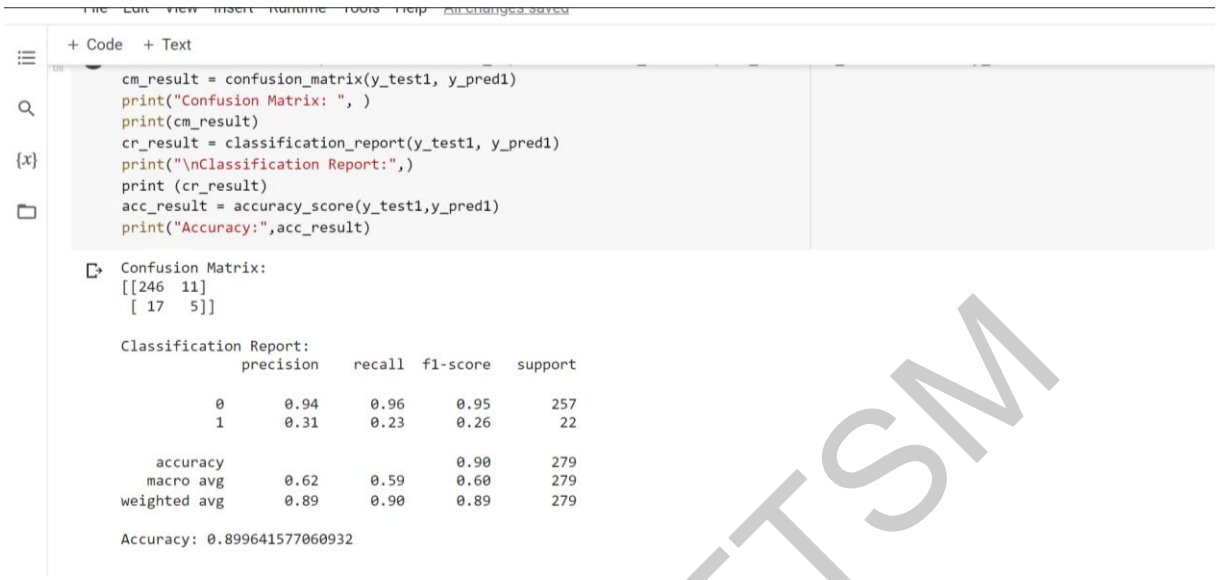
setiap model yang terdapat dalam modul MALAYA boleh diperolehi daripada dokumentasi modul MALAYA (Husein, Zolkepli, 2022). Bagi model KNN dan MultiOutputClassifier, nilai statistik akan diperolehi melalui barisan kod-kod yang telah dibina dalam projek seperti fungsi confusion_matrix() dan classification_report().

[5]:

	Size (MB)	Quantized Size (MB)	macro precision	macro recall	macro f1-score
bert	425.4	111.00	0.99291	0.97864	0.98537
tiny-bert	57.7	15.40	0.98151	0.94754	0.96134
albert	48.6	12.80	0.98026	0.95332	0.96492
tiny-albert	22.4	5.98	0.96100	0.90363	0.92374
xlnet	446.6	118.00	0.99344	0.98154	0.98725
alxlnet	46.8	13.30	0.99215	0.97575	0.98337
fastformer	446.6	113.00	0.95031	0.94018	0.94498
tiny-fastformer	77.3	19.70	0.93574	0.89979	0.91640

Rajah 1: Nilai kejituan, dapatan semula dan skor-F1 bagi setiap model dalam modul MALAYA

Berdasarkan Rajah 5.8, nilai ketepatan yang diperolehi ialah 0.899641577060932 melalui perbandingan antara dataset sebenar (y_test) dengan dataset ramalan (y_pred). Ini menunjukkan bahawa model ini dapat mengecam entiti nama dengan baik dan mempunyai persamaan dengan dataset sebenar sebanyak 89.9%. Bagi fungsi kod confusion matrix, model ini telah berjaya mendapatkan ramalan yang betul melalui perbandingan dengan dataset sebenar iaitu nilai *true positive* sebanyak 246 dan *true negative* bernilai 5 menjadikan jumlah ramalan betul adalah 251 daripada 279. Bagi ramalan yang salah, iaitu terdiri daripada *false positive* dan *false negative*, model ini meramalkan sebanyak 11 dan 17 menjadikan jumlah tersebut adalah 28 daripada 279 data. Dalam bahagian classification report pula, label 0 iaitu nilai yang menunjukkan bahawa model ini dapat mengekstrak jenis entity LAIN-LAIN memberikan nilai kejituan 0.94, dapatan semula 0.96 dan skor-F 0.95. Manakala bagi label 1 iaitu nilai yang menunjukkan bahawa model ini dapat mengekstrak jenis entity LOKASI, MANUSIA dan ORGANISASI memberikan nilai kejituan 0.31, dapatan semula 0.23, dan skor-F 0.26. Nilai-nilai statistik yang rendah ini adalah disebabkan oleh format output yang telah dipilih. Fungsi kod confusion_matrix() serta classification_report() tidak menyokong ramalan data MultiOutputClassifier() atau data yang memiliki lebih daripada satu lajur bagi kelas y. Maka, untuk menggunakan fungsi kod ini, dataset y_test dan y_pred perlu diubah bentuk daripada tiga dimensi (279 ,3) kepada satu dimensi (279, 1).

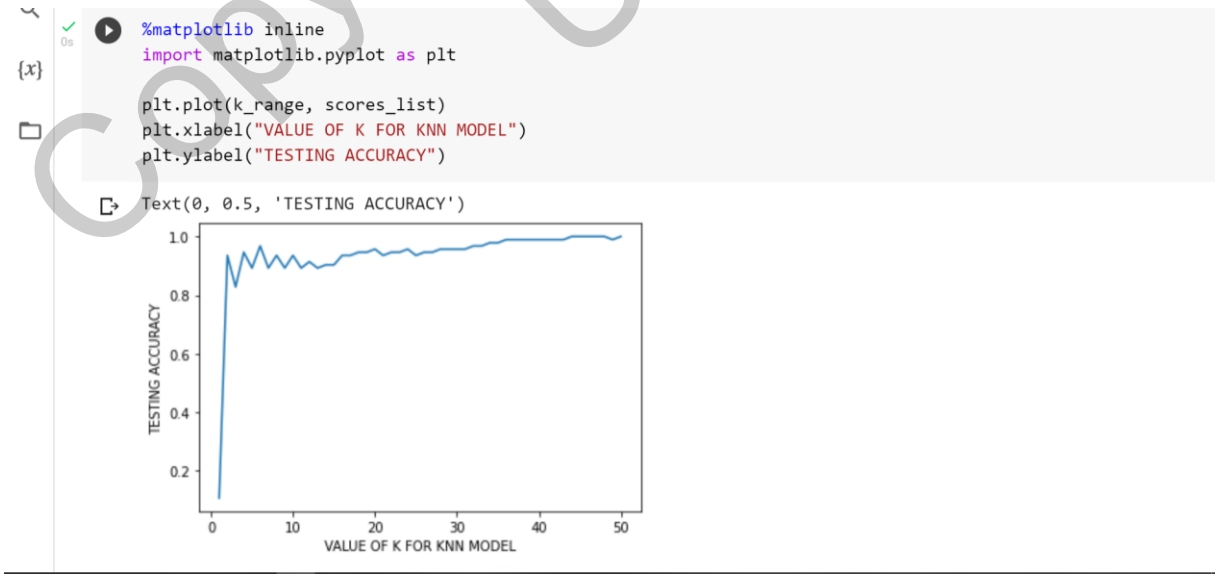


```
cm_result = confusion_matrix(y_test1, y_pred1)
print("Confusion Matrix: ", )
print(cm_result)
cr_result = classification_report(y_test1, y_pred1)
print("\nClassification Report:",)
print (cr_result)
acc_result = accuracy_score(y_test1,y_pred1)
print("Accuracy:",acc_result)
```

```
Confusion Matrix:
[[246  11]
 [ 17   5]]

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.96      0.95       257
           1       0.31      0.23      0.26        22

    accuracy                           0.90       279
   macro avg       0.62      0.59      0.60       279
weighted avg       0.89      0.90      0.89       279

Accuracy: 0.899641577060932
```

Rajah 2: Confusion matrix dan classification report

Pada peringkat seterusnya, pada Rajah 5.9 menunjukkan tren menaik sehingga statik secara melintang pada suatu nilai ketepatan yang sama. Tren ini menunjukkan bahawa semakin tinggi nilai kluster k dalam model KNN yang digunakan, nilai ketepatan juga akan bertambah, sehingga ia mencapai nilai ketepatan bersamaan dengan 1.0 . Graf ini menunjukkan bahawa model ini akan bertambah keberkesanannya apabila nilai kluster k semakin banyak.



```
%matplotlib inline
import matplotlib.pyplot as plt

plt.plot(k_range, scores_list)
plt.xlabel("VALUE OF K FOR KNN MODEL")
plt.ylabel("TESTING ACCURACY")
```

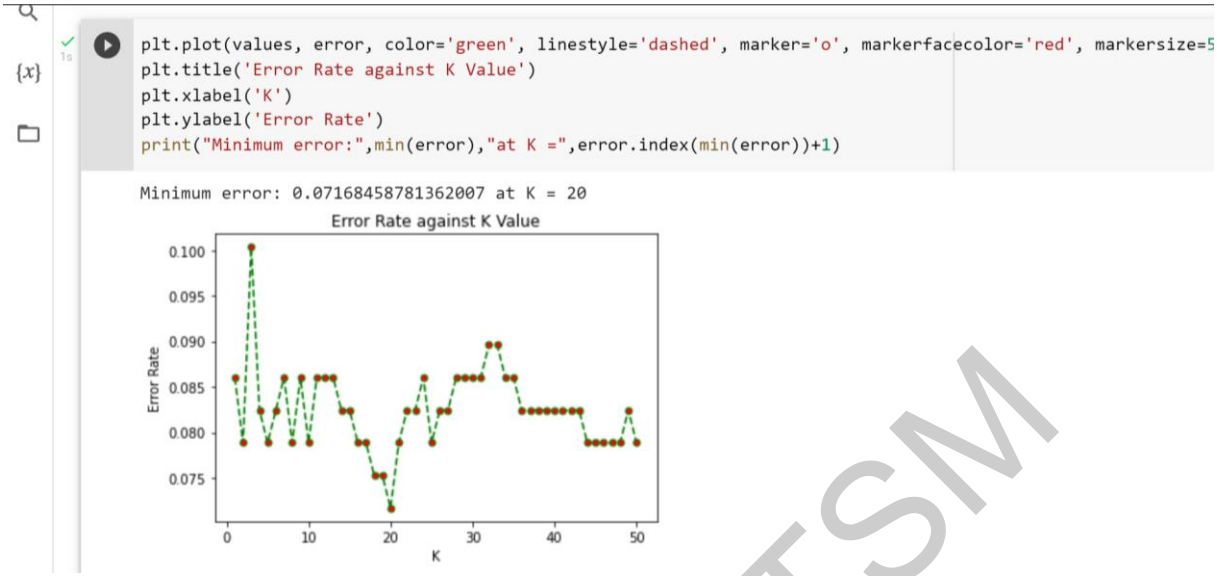
```
Text(0, 0.5, 'TESTING ACCURACY')
```


Rajah 3: Graf bagi skor ketepatan bagi setiap pertambahan nilai k KNN

Di samping itu, pada Rajah 5.10 di bawah adalah berkenaan nilai ketepatan antara training dataset dan testing dataset menunjukkan tren yang sama seperti graf pada Rajah 5.8. Graf testing dataset mempunyai lebih nilai ketepatan yang lebih tinggi pada nilai k kurang daripada 10 berbanding training dataset. Kemudian, kedua-dua graf ini akan bertembung dari segi nilai ketepatan sehingga kedua-dua graf mencapai nilai ketepatan bersamaan dengan 1.0. Tren ini menunjukkan bahawa nilai ketepatan bagi training dataset akan meningkat selari dengan testing dataset.

Rajah 4: Perbezaan graf bagi skor ketepatan training dataset dan testing dataset

Akhir sekali, merujuk kepada Rajah 5.11, graf yang dibina tidak menunjukkan tren yang stabil kerana berlaku peningkatan dan penurunan berkala apabila nilai k pada model KNN bertambah. Graf ini juga memaparkan bahawa nilai kadar ralat adalah paling rendah adalah apabila nilai kluster k = 20. Nilai kadar ralat ini boleh berubah setiap kali model ini berjalan dalam kod program. Berdasarkan graf ini, model ini mempunyai nilai kadar ralat yang berbeza pada setiap nilai k dan tidak memaparkan sebarang tren menaik atau menurun bagi mengurangkan kadar ralat dalam projek.



Rajah 5: Graf nilai kadar ralat bagi setiap pertambahan nilai k KNN

6 KESIMPULAN

Secara keseluruhannya, sistem pengecaman entiti nama Bahasa Melayu berjaya dibangunkan mengikut objektif yang digariskan. Terdapat beberapa kelemahan sistem seperti nilai ketepatan sistem dalam mengekstrak entiti nama dalam teks dokumen Bahasa Melayu yang menjejaskan keberkesanan sistem. Sistem ini akan dapat membantu orang awam atau pakar sejarawan bagi melihat bagaimana padanan entiti nama pada setiap kata dijalankan oleh sistem secara automatik. Walaupun terdapat beberapa kekurangan, diharapkan sistem ini dapat dijadikan titik kajian untuk kajian pada masa hadapan.

7 RUJUKAN

(Pedamkar, P. 2020. *Incremental Model*)https://www.educba.com/incremental-model/

(Clark, Kevin, Khandelwal, Urvashi. Levy, Omer. Manning, Christopher D. 2019. What does BERT look at? An Analysis of BERT's attention) https://aclanthology.org/W19-4828.pdf

(Wikipedia. 2021. *Bahasa Melayu.*) https://ms.wikipedia.org/wiki/Bahasa_Melayu

(GOYAL, C. 2021. *Part 10: Step by Step Guide to Master NLP – Named Entity Recognition.*

) https://www.analyticsvidhya.com/blog/2021/06/part-10-step-by-step-guide-to-master-nlp-named-entity-recognition/#h2_3

(Willyawan, A. 2018. - *NAMED ENTITY RECOGNITION (NER) BAHASA INDONESIA MENGGUNAKAN CONDITIONAL RANDOM FIELD DAN POS-TAGGING*) https://repositori.usu.ac.id/bitstream/handle/123456789/6868/131402089.pdf?sequence=1&isAllowed=y

(Kalyanathaya, K.P, D. Akila and P. Rajesh. 2019. *Advances in Natural Language Processing – A Survey of Current Research Trends, Development Tools and Industry Applications. International Journal of Recent Technology and Engineering (IJRTE). Judul 7. Isu 5C. 2019*) https://www.ijrte.org/wp-content/uploads/papers/v7i5c/E10480275C19.pdf

(Dewan Bahasa dan Pustaka. 2021. *Pusat Rujukan Persuratan Melayu (PRPM)*) https://prpm.dbp.gov.my/

(Jurafsky, D, Martin, H. J. 2021. *Information Extraction. Speech and Language Processing. Stanford University*) https://web.stanford.edu/~jurafsky/slp3/17.pdf

(Jurgita, K. D, Anders, N, Johannessen, J.B, Algis K. 2013. *Exploring Features for Named Entity Recognition in Lithuanian Text Corpus*) https://aclanthology.org/W13-5611.pdf

(Olle Bridal, 2021. *Named-entity recognition with BERT for anonymization of medical records*) https://liu.diva-portal.org/smash/get/diva2:1566701/FULLTEXT01.pdf

(Wikipedia Bahasa Melayu, 2021. Sejarah Malaysia - Wikipedia Bahasa Melayu) https://ms.wikipedia.org/wiki/Sejarah_Malaysia

(Morsidi, F., Sulaiman, S. & Abdul, R. 2017. Feature Extraction using Regular Expression in Detecting Proper Noun for Malay News Articles based on KNN Algorithm (June), 0–23.) Webpage: https://www.ukm.my/apjitm/article/2019/0801/04.pdf doi:10.4314/jfas.v9i5s.16

(M. N Mansor, G.K Khairul. 2015. Keberkesanan Mata Pelajaran Sejarah dalam Membina Etos Bangsa Generasi Muda di Malaysia. Jurnal Komunikasi Borneo Edisi Khas (Konvokesyen ke-17 UMS) 2015) https://www.researchgate.net/profile/Khairul-Kaspin/publication/330184929_Keberkesanan_Mata_Pelajaran_Sejarah_dalam_Membina_Etos_Bangsa_Generasi_Muda_di_Malaysia/links/600edd6892851c13fe3616ed/Keberkesanan-Mata-Pelajaran-Sejarah-dalam-Membina-Etos-Bangsa-Generasi-Muda-di-Malaysia.pdf

(Gabriella, Carrington, McMahon, Sydney, Michel, Adente. 2021. *Python*) https://www.bu.edu/lernet/artemis/years/2021/projects/FinalPresentations/Python.pdf

(Suyal, Manish, Goyal, Parul. A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms based on Supervised Learning. 2022) https://ijettjournal.org/Volume-70/Issue-7/IJETT-V70I7P205.pdf

(Zolkepli, Husein. 2018. *Malaya.*) https://malaya.readthedocs.io/en/latest/Api.html?highlight=api#module-malaya

(Ting, K.M. (2011). Error Rate. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA.) https://doi.org/10.1007/978-0-387-30164-8_262

Muhammad Arif Syamil bin Mohd Rahimi (A177313)
Saidah Saad
Fakulti Teknologi & Sains Maklumat,
Universiti Kebangsaan Malaysia