PENGECAMAN ENTITI NAMA PADA TEKS MEDIA SOSIAL TWITTER BAHASA MELAYU MENGGUNAKAN KAEDAH BiLSTM-CRF

PUTRI MAYA SYAKILLA BINTI HAIROL AKHMA

SABRINA TIUN

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Era teknologi diakui kian pesat membangun di seluruh dunia termasuk negara Malaysia. Namun, penggunaan teknologi serta pelbagai aplikasi media sosial tidak pernah dilupakan kerana ia merupakan platform bagi setiap individu tidak mengira lapisan masyarakat untuk berkongsi mahupun melihat segala informasi yang berada di internet. Hal ini demikian kerana, informasi di internet lebih laju dan mudah dicapai berbanding pembelian surat khabar seperti dahulu kerana segalanya hanya dihujung jari. Media sosial yang sering dijadikan platform utama bagi menyampaikan maklumat ialah *Twitter*, *Facebook*, *Whatsapp*, *Telegram* dan *Instagram*. Sebagai contoh, ketika negara dilanda musim pandemik Covid-19 yang lalu, platform *Twitter* dijadikan medium utama dalam menyampaikan informasi penting mengenai Covid-19 kerana ia lebih inovatif dan pantas. Meskipun begitu, setiap informasi yang disampaikan adalah penting untuk memahami apa yang ingin disampaikan. Oleh itu, bagi memahami sesuatu perkara dalam teks, penggunaan alatan Pemprosesan Bahasa Tabii (PBT) seperti Pengecaman Entiti Nama (PEN) adalah sangat membantu. PEN digunakan bertujuan untuk mencari dan mengesan entiti nama di dalam sesebuah teks dan mengklasifikasikannya mengikut kategori yang telah ditetapkan. Sebagai contoh, entiti individu, tempat, kewangan, peratus, masa, tarikh dan ukuran. Permasalahan kajian ialah wujudnya teks yang tidak berstruktur dan ditulis dengan gaya bebas iaitu berbeza dari teks formal. Oleh itu, kajian ini akan membangunkan sistem PEN dengan menggunakan Dua Arah Ingatan Jangka Pendek Jangka Panjang bersama lapisan medan rawak bersyarat (BiLSTM-CRF). Metodologi kajian yang digunakan ketika membangun sistem PEN ini ialah berdasarkan Kitaran Hidup Pembelajaran Mesin. Jangkaan hasil projek ini ialah dapat menghasilkan sistem PEN bagi teks media sosial Twitter Bahasa Melayu dan ia dibangunkan dengan menggunakan bahasa pengaturcaraan *Python* dan *HTML*. Akhir kata, sistem PEN ini berjaya dihasilkan dan ini telah dibuktikan melalui bab 4 mengenai Pembangunan dan Pengujian Sistem. Meskipun keputusan prestasi yang diperoleh tidak setanding dengan PEN bahasa Inggeris, namun sistem ini mempunyai ruang untuk ditambah baik di masa hadapan.

1 PENGENALAN

Lapisan masyarakat pasti tahu akan pelbagai platform media sosial yang kini menjadi pilihan setiap individu untuk berkomunikasi mahupun menyampaikan informasi. Contohnya, media *Twitter* telah menjadi pilihan segelintir masyarakat sebagai medium untuk menyampaikan maklumat penting mahupun mengenai kehidupan seharian. Perkara ini dapat dilihat penggunaanya ketika musim pandemik Covid-19 yang lalu di mana kebanyakkan informasi mengenai Covid-19 akan diberitahu di media Twitter sebelum ia ditayangkan di layar televisyen. Inisiatif ini memudahkan rakyat untuk menerima informasi dengan lebih pantas. Namun, segelintir masyarakat selesa menggunakan Twitter untuk meluahkan perasaan mahupun ayat-ayat rawak mengenai kehidupan seharian. Setiap ayat yang ditulis di Twitter, pasti memiliki makna dan entiti tersendiri dan oleh itu, alatan Pemprosesan Bahasa Tabii (PBT) memainkan peranan utama dalam mengesan entiti nama. PBT dapat memberi komputer keupayaan untuk memahami teks dan perkataan yang dituturkan oleh masyarakat dalam norma seharian. Terdapat tiga jenis teks data iaitu data berstruktur, data tidak berstruktur dan data separa berstruktur. Data berstruktur ialah data berbentuk standard, mempunyai struktur yang jelas, dan mempunyai lajur serta baris seperti pangkalan data. Data tidak berstruktur ialah data yang tidak berformat pangkalan data seperti PowerPoint, fail audio dan video, imej serta lain-lain. Kebanyakkan data di media sosial adalah berbentuk data tidak berstruktur kerana teks yang digunakan berbentuk gaya bebas dan berbeza dengan teks formal.

Melalui PBT, setiap perkataan mahupun ayat dapat dikategorikan mengikut entiti dan kategori yang tertentu. Sebagai contoh, entiti individu, lokasi, organisasi, masa, alamat, acara, harga dan nombor. Oleh itu entiti ini dapat ditentukan melalui kaedah PEN kerana ia digunakan bertujuan untuk mencari dan mengesan entiti nama di dalam sesebuah teks dan mengasingkannya mengikut kategori yang telah ditentukan.

2 PERNYATAAN MASALAH

PEN adalah tugas pelabelan urutan yang mencabar yang memerlukan pemahaman mendalam mengenai ortografik dan perwakilan pengedaran perkataan (Rrubaa Panchendrarajan & Aravindh Amaresan 2018). Dalam projek ini, pembangunan PEN mempunyai dua permasalahan yang besar: (1) Penyediaan set data untuk melatih model PEN. Menyediakan set data tweet bahasa Melayu untuk model PN adalah permasalahan yang rumit dan renyah. Hal ini kerana, selain teks Bahasa Melayu di *Twitter* adalah data yang mempunyai pelbagai masalah bahasa, ianya perlu dianotasi dengan tag yang membantu kejituan PEN, seperti tag golongan kata (GK). (2) Terdapat beberapa kajian mengenai PEN bagi teks tertentu, tetapi kebanyakkan kaedah yang digunakan ialah pendekatan lama iaitu berasaskan peraturan, set data formal dan dalam bahasa lain. Pendekatan tersebut menghasilkan keputusan berdasarkan peraturan tertentu yang telah disenaraikan oleh pengaturcara dan ia memerlukan dua komponen utama iaitu satu set peraturan dan satu set fakta. Selain itu, pembelajaran mesin adalah ditujukan untuk menangani isu yang kompleks dengan persekitaran yang bervariasi (Jason M. 2020). Rayner et al. (2014) memperoleh kadar ketepatan sebanyak 85% dengan menggunakan pendekatan berasaskan peraturan dan ia ternyata masih rendah. Oleh itu, kadar ketepatan ini masih boleh ditingkatkan dengan menggunakan pendekatan berlainan. Model PEN bagi Bahasa Melayu perlu dibangunkan dengan menggunakan pendekatan terkini dalam domain tweet Bahasa Melayu.

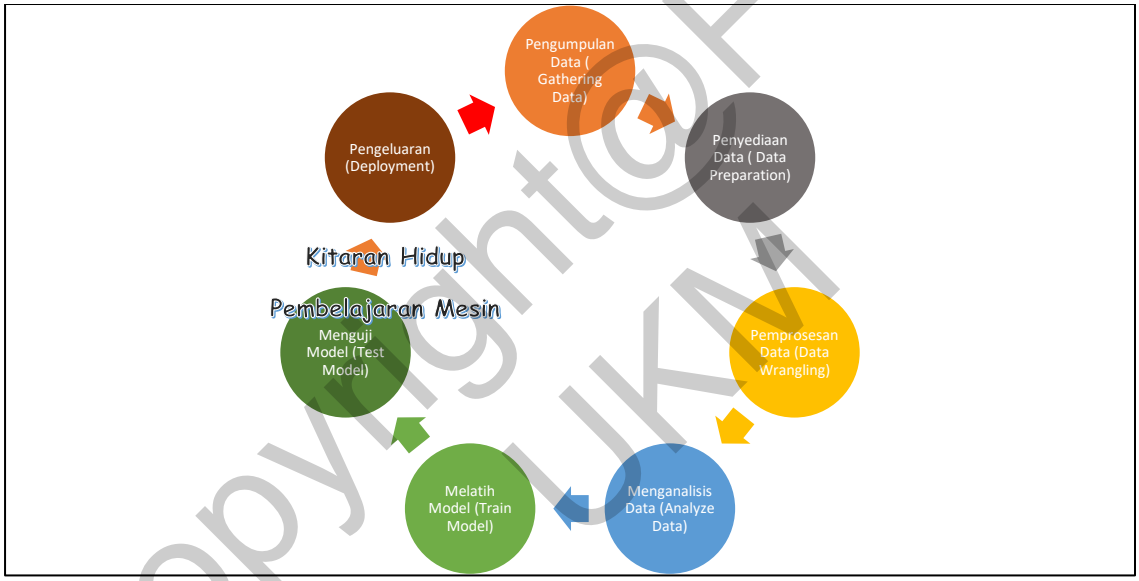
3 OBJEKTIF KAJIAN

Berdasarkan cadangan penyelesaian yang dinyatakan, beberapa objektif kajian yang bakal dicapai ialah:

I. Menyediakan set data untuk membina model PEN teks media sosial Bahasa Melayu.
II. Membangunkan model PEN bagi teks media sosial Bahasa Melayu dengan menggunakan kaedah BiLSTM-CRF.

4 METOD KAJIAN

Kajian ini menggunakan Kitaran Hidup Pembelajaran Mesin, *Machine Learning Life Cycle (*MLLC*).* Ia merupakan satu kitaran yang berulang antara penambahbaikan data, model dan penilaian yang tidak pernah selesai. Kitaran ini penting dalam membangunkan model pembelajaran mesin kerana ia memfokuskan kepada penggunaan hasil model dan penilaian untuk memperbaiki sesebuah set data. Set data yang berkualiti tinggi ialah cara efektif untuk melatih model agar ia berkualiti (Eric Hofesmann 2021). Berdasarkan rajah 4.1, terdapat 7 fasa yang terdapat di dalam kitaran MLLC ini, iaitu pengumpulan data, penyediaan data, pemprosesan data, menganalisis data, melatih model, menguji model dan pengeluaran.



Rajah 4.1 Kitaran Hidup Pembelajaran Mesin

MLLC memberikan manfaat kuasa, kelajuan serta kecekapan melalui pembelajaran tanpa memprogramkannya secara eksplisit ke dalam aplikasi. MLLC juga digunakan oleh rata-rata saintis data dan jurutera data untuk membangunkan, melatih serta menyediakan model yang menggunakan sejumlah besar data supaya organisasi boleh memanfaatkan kecerdasan buatan dan algoritma pembelajaran mesin untuk memperoleh nilai perniagaan praktikal (Priya Pedamkar 2020).

4.1 FASA PENGUMPULAN DATA

Dalam pembangunan sebuah sistem yang memiliki kadar ketepatan tinggi pasti mempunyai data yang berskala besar serta data diperoleh dari sumber yang betul. Oleh itu, kajian ini akan melaksanakan pembangunan sistem bagi pengecaman entiti nama Bahasa melayu dalam teks media sosial *Twitter*. Set data diperoleh dari dua sumber iaitu data dari kajian Chew (2021) dan data dari laman sesawang *Kaggle* di mana data ini telah ditandakan dengan PGK secara manual menggunakan karya kajian sebelumnya.

4.2 FASA PENYEDIAAN DATA

Di fasa ini data akan ditempatkan atau disimpan di fail yang bersesuaian serta data yang diperoleh akan diletak secara rawak tanpa mengikut susunan data. Set data yang diperoleh dari kajian lepas serta dari laman sesawang *Kaggle* adalah bertaraf .txt di mana ia agak sukar sekiranya ingin digunakan dalam model algoritma. Oleh itu, fasa ini akan menukar taraf dokumen asal iaitu .txt kepada .csv bagi memudahkan proses pemprosesan data.

Namun, data ini masih lagi tidak lengkap dengan entiti nama dan ia akan ditandakan secara manual di dalam kajian ini. Penandaan entiti nama ini mengikut format *Inside-Outside-Begin Tagging* (IOB) iaitu jika entiti terdiri daripada berbilang (sub) perkataan, seperti nama ‘ahmad abu’, maka perkataan pertama ‘ahmad’ diwakili oleh B-ind, dan perkataan kedua ‘abu’ diwakili oleh I-ind yang menunjukkan bahawa entiti tersebut merupakan sebahagian daripada entiti sebelumnya. Perkataan yang tidak memiliki sebarang entiti dianotasi dengan O. Format ini berguna untuk membezakan antara entiti yang berturutan dalam sesuatu ayat (Sybren Jansen 2021). Jadual 4.1 menunjukkan senarai entiti nama yang terlibat dalam sistem PEN ini manakala jadual 4.2 menunjukkan contoh format IOB dalam data korpus.

Jadual 4.1 Senarai Entiti PEN

Bil.	Tag PEN	Entiti Nama	Penerangan	Contoh
1.	B-ind	Individu	Nama individu / gelaran	Ahmad, Sarah, Datuk Khairi, pakcik, kakak, nenek
2.	B-tmp	Tempat	Lokasi, sesuatu yang menerangkan tempat	Langkawi, Jalan Sultan Ismail, Masjid Kota Damansara
3.	B-tarikh	Tarikh	Tarikh hari, bulan dan tahun	31 Disember 2000, 31 Dis
4.	B-msa	Masa	Jam waktu dan kejadian	Jam 11.30 pagi, 9 am, 3 pm
5.	B-kew	Kewangan	Nilai duit atau mata wang	RM23.00, 15 sen, $20
6.	B-ukrn	Ukuran	Sesuatu yang nilainya boleh diukur	35 cm, 27g
7.	B-prts	Peratusan	Mempunyai simbol %	97%, 17.7%

Jadual 4.2 Contoh Format IOB dalam Data Korpus

Ayat	Perkataan	PGK	Tag PEN
1	nama	KN	O
1	saya	GN1	O
1	maria	KN	B-ind
1	dan	KH-KEP	O
1	saya	GN1	O
1	bekerja	KK	O
1	dekat	KA	O
1	kuala	KN	B-tmp
1	lumpur	KN	I-tmp

4.3 FASA PEMPROSESAN DATA

Data Wrangling atau pemprosesan data merupakan proses pembersihan data dan menukar data mentah kepada format yang boleh digunakan. Ia adalah proses membersihkan data dan memilih pemboleh ubah untuk menjadikannya lebih sesuai pada fasa berikutnya iaitu fasa menganalisis data. Pembersihan data diperlukan untuk menangani isu data tidak berkualiti. Realitinya, setiap data yang diperoleh pasti mempunyai pelbagai isu antaranya, nilai yang hilang, data pendua, data tidak sah, dan *noise*. Oleh itu, dalam fasa ini, akan digunakan beberapa kaedah untuk pembersihan data.

Set data yang telah ditandakan dengan Penandaan Golongan Kata (PGK) serta Pengecaman Entiti Nama (NER) akan ditunjukkan dalam dataframe. Seterusnya, data akan dikenalpasti sama ada setiap baris atau lajur mempunyai data dengan menggunakan isnull( ).any( ) . Bilangan tag PEN unik akan dikira kekerapan muncul di dalam data. Setelah data telah dikenalpasti dan dibersihkan, *Data Cleaning*, maka data akan divisualisasikan dengan graf histogram. Seterusnya data akan dibahagikan kepada 2 set data iaitu X_train yang digunakan untuk melatih model dan X_test untuk menguji model. Proses yang terlibat dalam fasa pemprosesan data ialah:

1. Memuatkan set data.
2. Menyemak setiap baris dan lajur sama ada mempunyai data atau tidak.
3. Token dan label tag PEN diberi indeks.
4. Setiap token digabungkan dengan label tag PEN ke dalam sebuah urutan array yang sama berdasarkan indeks yang diberikan. Sebagai contoh, “kenapa ahmad tinggal di damansara” akan menjadi [(‘kenapa’ , ‘O’), (‘ahmad’ , ‘B-ind’), (‘tinggal’ , ‘O’), (‘di’ , ‘O’), (‘damansara’, ‘B-tmp’)] .
5. Pra-urutan *padding* dijalankan untuk menyamakan panjang setiap urutan perkataan. Sebagai contoh, maksimum token dalam satu ayat ditetapkan sebagai 140 token. Sekiranya token dalam sesebuah ayat hanyalah 48 maka ‘ENDPAD’ akan ditambah bagi urutan berikutnya kerana *padding* yang digunakan dalam model ini ialah *post padding*.
6. Pembolehubah kategori akan diubah kepada vektor binari yang unik menggunakan *One-hot encoding*.
7. Set data akan dibahagi kepada 2 set iaitu data latihan X_train dan data ujian X_test.

4.4 FASA MENGANALISIS DATA

Seterusnya, data yang telah dibersihkan akan digunakan di fasa menganalisis data. Di fasa ini, pemilihan teknik analisis, membina model dan semak keputusan akan dilakukan. Objektif langkah ini adalah untuk membina model pembelajaran mesin untuk menganalisis data menggunakan pelbagai teknik dan diakhiri dengan penyemakan hasilnya. Pembangunan model merupakan fasa penting merujuk kepada proses menggunakan kecerdasan buatan (AI) untuk membuat model yang berbeza, memilih dan menunjukkan data dari kumpulan yang lebih besar. Fasa ini juga penting untuk meningkatkan kecekapan proses pengeluaran, *Release,* dan membantu dalam pengurangan risiko ralat. Pembangunan model bagi PEN boleh dilatih dengan banyak kaedah seperti *Rule-Based Approach, Neural Network Approach* dan *Data Driven Approach.*

Kajian ini menggunakan pendekatan rangkaian neural iaitu BiLSTM-CRF. Z.Huang et al. (2015) memperkenalkan model LSTM-CRF dua arah (BiLSTM-CRF) untuk mencapai ketepatan tinggi pada tugasan PGK, *chunking* dan PEN. Semua model rangkaian saraf yang disebutkan memerlukan sejumlah besar korpus berlabel untuk pengekstrakan ciri. Liu et al. (2018) mencadangkan karya saraf untuk mengekstrak maklumat dari teks mentah tanpa sebarang tambahan. Ia menggabungkan model Bahasa BiLSTM-CRF untuk menyelesaikan tugas pelabelan urutan dan ia terbukti berkesan. Pembinaan model melibatkan Pustaka Tensorflow dan Keras serta melibatkan 3 lapisan iaitu lapisan pembenaman, BiLSTM dan lapisan CRF.

LAPISAN PEMBENAMAN

Lapisan pembenaman merupakan lapisan pertama dalam Keras dan ia mengira perwakilan vektor bagi setiap perkataan mengikut gabungan empat ciri berikut :

1. Perwakilan vektor berasaskan aksara
2. Pembenaman perkataan pra-terlatih
3. PGK setiap perkataan
4. *Casing features* setiap perkataan

Perwakilan berasakan watak dan pembenaman perkataan digabungkan dengan tag PGK dan *casing features* untuk mendapatkan pembenaman terakhir sesuatu perkataan (Rrubaa

dan Aravindh 2018). Natasha (2019) ada menyatakan bahawa pemetaan perkataan ke *integer*, *word2idx*, perkataan boleh diwakili sebagai vektor nombor seperti berikut :

1. Setiap perkataan akan diwakili oleh vektor n-dimensi, di mana n ialah saiz perbendaharaan kata.
2. Perwakilan vektor setiap perkataan akan kebanyakkannya "O", kecuali terdapat entiti dalam kedudukan yang sepadan dengan indeks perkataan dalam perbendaharan kata.

BiLSTM

Model PEN ini akan menggunakan ciri-ciri masa lalu dan hadapan dan ia akan menjalankan urutan vektor perkataan yang baru. Tugas pelabelan jujukan seperti PEN, maklumat konteks masa lalu dan akan datang amat penting, dan oleh itu model LSTM dua arah sesuai untuk tugas PEN (Graves et al. 2013). LSTM dua arah ialah model yang terdiri daripada dua LSTM, satu mengambil input ke arah hadapan dan satu lagi ke arah belakang. BiLSTM berkesan dalam meningkatkan jumlah maklumat yang tersedia untuk rangkaian, meningkatkan konteks yang tersedia untuk algoritma. Sebagai contoh, ia mengetahui perkataan apa yang akan menyusuli dan mendahului perkataan dalam ayat. Selepas itu, output dari lapisan BiLSTM akan ke dalam lapisan CRF.

CRF

Seperti yang dinyatakan, lapisan BiLSTM digunakan untuk menangkap maklumat masa lalu dan masa hadapan manakala lapisan CRF digunakan untuk meramalkan tag keseluruhan ayat dengan mempertimbangkan kebergantungan tag output (Yangzeng et al. 2018). Oleh itu, kajian ini akan membina rangkaian saraf dengan menggunakan lapisan tersembunyi BiLSTM sebagai jujukan input lapisan CRF. CRF akan meramalkan Tag PEN yang sesuai terhadap sesuatu perkataan.

4.5 FASA MELATIH MODEL

Seterusnya, fasa melatih model akan dijalankan. Melalui LSTM, ia dilatih menggunakan *Backpropagation Through Time,* (BPTT). Rangkaian LSTM mempunyai blok memori yang disambungkan ke dalam lapisan Terdapat 3 jenis pintu atau *gate* dalam satu unit ingatan iaitu *Forget Gate*, *Input Gate* dan *Output Gate*. Setiap unit mempunyai pemberat yang dipelajari semasa prosedur latihan. Parameter yang akan digunakan dalam fasa melatih model ini ialah saiz *epoch, batch size, loss function* dan *optimizer*. Nilai *epoch* akan dipilih dengan jumlah paling maksimum tetapi akan berubah nilainya sekiranya nilai *loss function* semakin tinggi atau tidak berubah. *Loss function* merupakan satu fungsi yang ingin diminimumkan atau maksimumkan dan apabila ia ingin diminimumkan, ia dipanggil fungsi kos (*Cost function*), fungsi kehilangan (*Loss function*) atau fungsi ralat (Jason Brownlee 2019). Fungsi ini digunakan untuk menentukan ralat antara pengeluaran algoritma dan nilai sasaran yang diberikan. Model ini akan dilatih dengan nilai parameter yang berbeza dan akan dicatat di akhir proses pelatihan model ini. Parameter yang menghasilkan nilai *loss function* terendah dan kadar ketepatan yang tinggi akan digunakan dalam fasa berikutnya iaitu fasa pengujian model.

4.6 FASA MENGUJI MODEL

Akhir sekali, fasa pengujian terhadap model dan pengeluaran akan dijalankan. Berdasarkan hasil dari fasa pelatihan model, parameter terbaik telah dipilih dan akan diuji dengan menggunakan data ujian, *Testing Data.* Fasa ini penting dalam usaha untuk mengelakkan daripada berlakunya *overfitting* iaitu berlaku apabila model mahir dalam mengklasifikasikan atau meramalkan pada data latihan tetapi tidak mahir dalam meramalkan data ujian (Deeplizard 2017). Secara amnya, model seperti ini melebihkan data dalam set latihan menyebabkan ia *overfitting*. Setelah selesai fasa pengujian, fasa pemilihan model terbaik akan dijalankan. Manakala fasa pengeluaran merupakan fasa menggunakan model di dalam sistem dunia sebenar.

Penilaian model akan diuji dengan menggunakan formula Sadman (2020) yang menerangkan tentang ketepatan model mengikut 6 jenis parameter iaitu TP (*True Positive*), FP (*False Positive*), TN (*True Negative)* dan FN (*False Negative*). TP berlaku apabila model meramalkan label sebagai positif dan tag sebenar ialah positif. Manakal bagi FP, ia berlaku apabila model meramalkan label sebagai positif dan tag sebenar ialah negatif. TN pula berlaku

apabila model meramalkan label sebagai negatif dan tag sebenar ialah negatif. FN ialah model meramalkan label sebagai negatif dan tag sebenar ialah positif (Sadman 2020).

Formula Sadman(2020) untuk mengira ketepatan Tag PEN adalah seperti berikut:

$$\text{Ketepatan} = \frac{TP}{TP+FP}$$

Formula Sadman(2020) untuk mengira sensitiviti Tag PEN adalah seperti berikut:

$$\text{sensitiviti} = \frac{TP}{TP+FN}$$

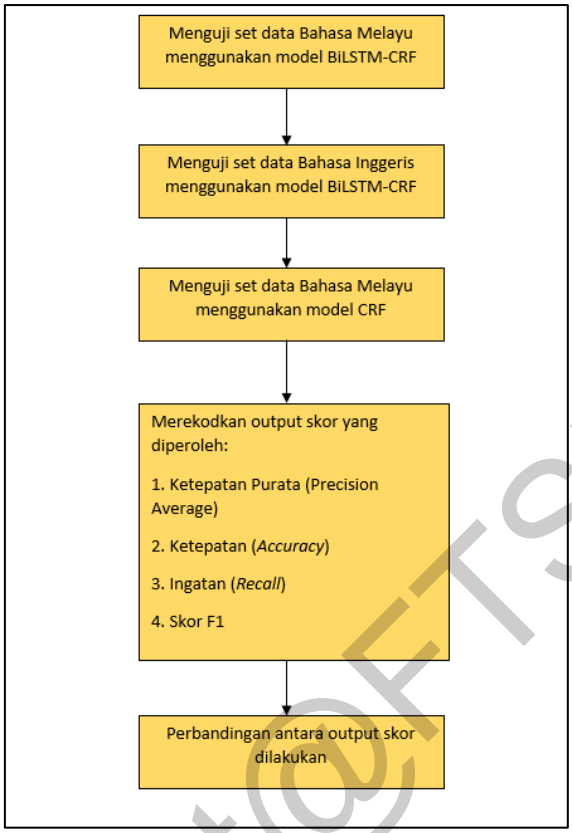
Formula Sadman(2020) untuk mengira skor f1 Tag PEN adalah seperti berikut:

$$\text{skor f1} = \frac{ketepatan * sensitiviti}{2 * (ketepatan + sensitiviti)}$$

Setelah selesai pengiraan ketepatan model di atas, pemilihan model terbaik akan dijalankan iaitu model yang mempunyai ketepatan, skor f1 tertinggi akan dipilih dalam penghasilan sistem PEN ini.

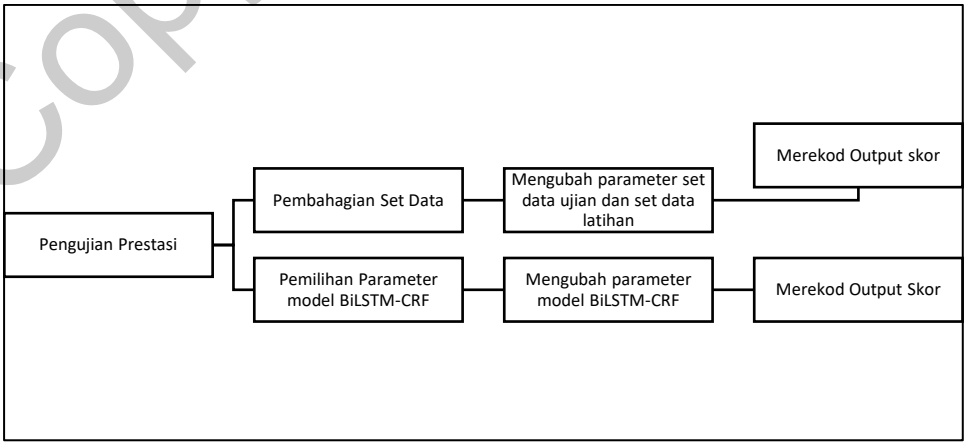
5 HASIL KAJIAN

Sistem PEN ini akan menggunakan pengujian penilaian pra-pemprosesan ataupun lebih mudah difahami dengan analisis data. Secara amnya, sistem ini akan diuji dan akan dibandingkan keputusan ketepatan antara kaedah BiLSTM-CRF Bahasa Inggeris, BiLSTM-CRF Bahasa Melayu dan juga CRF Bahasa Melayu. Oleh itu, rajah 5.1 di bawah menunjukkan pelan pengujian objektif yang akan digunakan bagi fasa pengujian ini.



Rajah 5.1 Peta Alir Pelan Pengujian Objektif

Pengujian bagi pembangunan sistem PEN ini akan diteruskan dengan menggunakan jenis pengujian tidak berfungsi iaitu pengujian prestasi. Prestasi yang dimaksudkan ialah output skor yang diperoleh ketika proses pengujian dijalankan seperti yang ditunjukkan dalam rajah 5.2.



Rajah 5.2 Peta Alir Pelan Pengujian Prestasi

Rajah 5.2 menunjukkan proses pengujian secara berperingkat bagi pengujian prestasi yang dibahagikan kepada dua bahagian iaitu Pembahagian set data dan Pemilihan parameter model BiLSTM-CRF. Proses ini penting bagi memastikan parameter yang dipilih mampu menghasilkan output skor terbaik dan menghasilkan model yang tepat serta jitu. Berdasarkan rajah 5.1 dan 5.2, pengujian akan dilakukan bagi dua kategori iaitu pengujian objektif dan pengujian prestasi. Tujuan pengujian objektif dijalankan adalah untuk memastikan bahawa objektif sistem PEN dibangunkan tercapai, manakala pengujian prestasi dijalankan untuk menguji keberkesanan parameter yang dipilih.

5.1 PENGUJIAN OBJEKTIF

Objektif kajian ialah membangunkan model PEN bagi teks media sosial Bahasa Melayu dengan menggunakan kaedah BiLSTM-CRF. Oleh itu, bagi menguji sama ada objektif ini tercapai mahupun tidak, berdasarkan peta alir dalam Rajah 5.1, pengujian akan dimulakan dengan menguji set data Bahasa Melayu menggunakan model BiLSTM-CRF dan diikuti dengan menguji set data Bahasa Inggeris menggunakan model BiLSTM-CRF. Perbandingan ini dilakukan untuk menguji sama ada model BiLSTM-CRF dapat menghasilkan output sesuai mahupun tidak.

Jadual 5.1 Perbandingan Keputusan bagi Bahasa Berbeza

No.	**Keputusan** / **Kaedah**	**Ketepatan Purata (Precision average)**	**Ketepatan *(Accuracy)***	**Ingatan *(Recall)***	**Skor F1 *(F1 Score)***
1	**BiLSTM-CRF (Bahasa Inggeris)**	**99%**	**99%**	**99%**	**99%**
2	**BiLSTM-CRF (Bahasa Melayu)**	**6%**	**97%**	**7%**	**7%**

Berdasarkan jadual 5.1, dapat dilihat bahawa sistem PEN ini diuji dari segi perbezaan keputusan bagi penggunaan bahasa berbeza di mana nombor 1 menggunakan Bahasa Inggeris manakala nombor 2 menggunakan Bahasa Melayu. Secara amnya, dapat dirumuskan bahawa,

algoritma yang dipilih tidak mempunyai masalah kerana bagi set data Bahasa Inggeris telah mencapai ketepatan 99% namun bagi set data Bahasa Melayu pula telah memperoleh ketepatan sebanyak 97%.

Walaupun begitu, ketepatan bagi setiap entiti menunjukkan perbezaan ketara di mana Bahasa Melayu hanya memperoleh purata sebanyak 6% dan ini menunjukkan bahawa set data Bahasa Melayu yang digunakan bagi membina model BiLSTM-CRF PEN ini masih lagi mempunyai kekurangan. Kekurangan yang dinyatakan adalah dari segi saiz set data kerana set data Bahasa Inggeris memiliki 1050794 data manakala set data Bahasa Melayu sangat kecil dengan hanya memiliki 15576 data.

Objektif kajian yang seterusnya ialah menyediakan set data berlabel entiti nama untuk membina model PEN bagi teks media sosial Bahasa Melayu. Oleh itu, bagi menguji sama ada objektif ini tercapai mahupun tidak, berdasarkan peta alir dalam Rajah 5.1, pengujian akan diteruskan dengan menguji set data Bahasa Melayu menggunakan dua kaedah yang berbeza iaitu model BiLSTM-CRF dan model CRF.

Jadual 5.2 Perbandingan Keputusan bagi Kaedah Berbeza

No.	**Keputusan** / **Kaedah**	**Ketepatan Purata (Precision Average)**	**Ketepatan *(Accuracy)***	**Ingatan *(Recall)***	**Skor F1 *(F1 Score)***
1	**BiLSTM-CRF (Bahasa Melayu)**	**6%**	**97%**	**7%**	**7%**
2	**CRF (Bahasa Melayu)**	**92%**	**98%**	**91%**	**91%**

Jadual 5.2 menunjukkan perbandingan keputusan bagi kaedah berbeza di mana nombor 1 menggunakan kaedah BiLSTM-CRF dan nombor 2 menggunakan kaedah CRF. Kaedah CRF ini dipilih kerana ia merupakan kaedah tradisional yang popular digunakan bagi Pengecaman Entiti Nama (Arnaud 2021). Kedua-dua kaedah ini menggunakan set data yang sama dan dapat dilihat bahawa kaedah CRF mempamerkan keputusan yang lebih bagus berbanding kaedah BiLSTM-CRF. Oleh hal yang demikian, dapat dibuktikan bahawa set data Bahasa Melayu yang

dihasilkan tidak mempunyai sebarang masalah kerana mampu menghasilkan ketepatan tinggi ketika menggunakan kaedah CRF.

Oleh itu, dapat disimpulkan disini, model BiLSTM-CRF PEN perlu dibina menggunakan data yang besar agar model BiLSTM-CRF PEN Bahasa Melayu yang dibina setanding dengan BiLSTM-CRF PEN Bahasa Inggeris. Namun, objektif kajian iaitu membangun sistem PEN menggunakan kaedah BiLTSM-CRF telah mencapai sasaran, cuma perlu penambahbaikan seperti penambahan data. Hal ini demikian kerana algoritma pembelajaran mesin akan belajar daripada data dan skeiranya lebih besar data latihan, lebih bagus prestasi model yang dihasilkan (Appen 2020).

5.2 PENGUJIAN PRESTASI

Bahagian ini akan menunjukkan pengujian terhadap ketepatan apabila nilai parameter bagi pembahagian set data berlabel berubah mengikut jadual 5.3.

Jadual 5.3 Perubahan Parameter bagi Pembahagian Set Data

Data Latihan (*Train Data*)	**Data Ujian (*Test Data*)**	**Ketepatan (*Precision*)**
80%	20%	6%
70%	30%	1%
50%	50%	0%

Berdasarkan jadual 5.3, dapat dilihat variasi ketepatan yang dihasilkan dalam fasa pengujian model PEN ini. Walaupun begitu, ketepatan bagi 80% data latihan dan 20% data ujian mempamerkan ketepatan tertinggi iaitu 6%. Selain itu, jadual 5.4 merupakan nilai-nilai parameter yang akan diubah di dalam model BiLSTM-CRF ini.

Jadual 5.4 Perubahan Parameter bagi Model PEN

Parameter / No.	Saiz Pembenaman Embedding size	*Dropout*	*Recurrent Dropout*	*Activation*	Ketepatan Purata (*Precision Average*)	Ketepatan (Accuracy)
1	**270**	**0.5**	**0.2**	**relu**	**0%**	**0%**
2	**270**	**0.5**	**0.3**	**relu**	**6%**	**0%**
3	**300**	**0.5**	**0.2**	**relu**	**6%**	**97%**
4	**300**	**0.5**	**0.3**	**relu**	**6%**	**0%**

Berdasarkan jadual 5.4, dapat dilihat bahawa ketepatan, *Accuracy* bagi nombor 3 mencatat ketepatan tertinggi walaupun nilai *Precision Average* rendah. Nilai *Dropout* berfungsi untuk melatih nod tertentu dalam lapisan model, di mana 1.0 bermaksud tiada dropout, dan 0.0 bermaksud tiada output dari lapisan model. Nilai yang bagus bagi dropout ialah antara 0.5 sehingga 0.8 (Jason 2019). Nombor 3 memperoleh ketepatan 97% namun ketepatan ini masih lagi tidak jitu kerana ia hanya fokus kepada entiti 'O'. Ketepatan purata bagi model PEN ini memperoleh hanya 6% sahaja.

5.3 KEPUTUSAN PENGUJIAN

Kesimpulannya, pengujian yang dijalankan pada bahagian 5.4 menunjukkan bahawa parameter di jadual 5.5 memberikan keputusan ketepatan yang lebih tinggi dan akan digunakan di dalam model PEN ini.

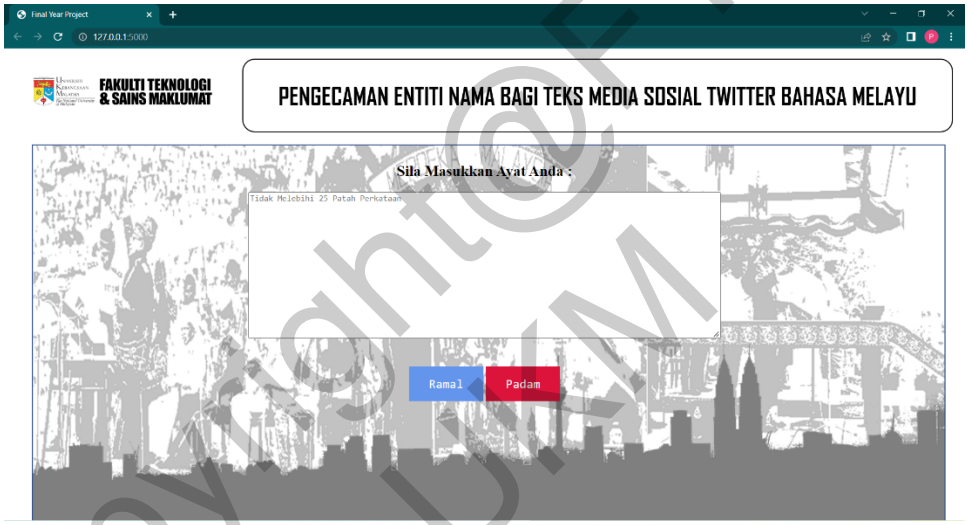
Jadual 5.5 Parameter Model PEN

Data Latihan	**Data Ujian**	**Saiz Pembenaman**	***Dropout***	***Recurrent Dropout***	***Activation***
80%	20%	300	0.5	0.2	relu

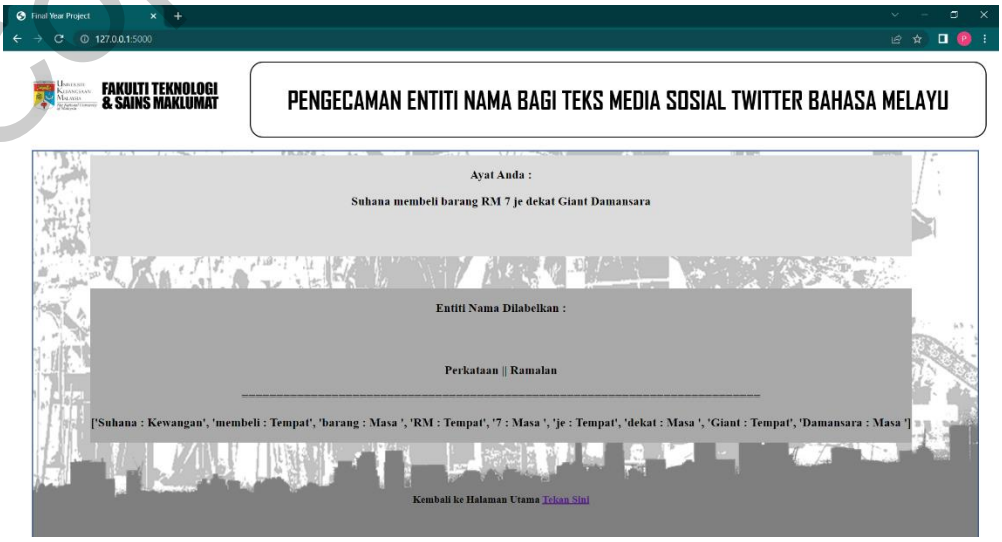
Gabungan kesemua parameter ini akan menghasilkan ketepatan yang lebih tinggi walaupun output yang dihasilkan masih lagi tidak begitu tepat.

5.4 ANTARA MUKA PENGGUNA

Antara muka pengguna memainkan peranan penting dalam usaha untuk menarik minat pengguna serta memudahkan penggunaan sistem. Oleh itu, sistem PEN ini dibangunkan dengan menggunakan perisian *Visual Studio Code* serta bahasa pengaturcaraan *Python*, *HyperText Markup Language* (HTML) dan *Casacading Style Sheets* (CSS). Rajah 5.3 menunjukkan bentuk antara muka input bagi pengguna untuk memasukkan ayat dan apabila pengguna klik pada butang 'Padam', maka input yang dimasukkan akan hilang dari ruangan input. Sekiranya pengguna klik pada butang 'Ramal', maka bentuk antara muka dalam rajah 5.4 akan muncul dengan memiliki ruangan output bagi entiti nama yang telah diramalkan bagi setiap perkataan dalam ayat input.



Rajah 5.3 Antara Muka Input



Rajah 5.4 Antara Muka Output

6 KESIMPULAN

Sistem Pengecaman Entiti Nama atau PEN diakui tentang kepentingannya terutama dalam sesebuah ayat mahupun karangan. Hal ini demikian kerana, setiap ayat yang ditulis mahupun dikarang pasti memiliki entiti seperti Individu, Tempat, Kewangan, Ukuran, Peratus dan lain-lain. Kewujudan entiti seperti ini mampu memudahkan si pembaca untuk lebih memahami tentang isi ayat tersebut. Selain itu, PEN mampu membantu golongan kanak-kanak untuk belajar mengenai entiti nama dalam sesuatu ayat. Namun, bagi penghasilan sistem yang memiliki ketepatan tinggi haruslah bermula dari penyediaan set data yang berskala besar agar algoritma mampu menghasilkan ramalan dengan lebih tepat dan jitu. Walaupun begitu, ketiadaan pustaka bagi Bahasa Melayu khasnya untuk PEN, menyukarkan proses pembangunan sistem PEN ini. Sistem ini dibangunkan bertujuan untuk membina sebuah sistem PEN bagi teks media sosial Twitter Bahasa Melayu dengan menggunakan model BiLSTM-CRF. Berdasarkan kajian susatera yang telah dinyatakan di Bab 2, dapat dilihat bahawa sistem BiLSTM-CRF sering digunakan oleh pengkaji sebelum ini kerana ia mampu menghasilkan ketepatan tinggi. Walaupun begitu, model BiLSTM-CRF ini masih belum pernah digunakan bagi teks Bahasa Melayu. PEN bagi teks Bahasa Melayu sering dibangunkan menggunakan pendekatan berasaskan peraturan. Namun, sistem ini dibangunkan dengan 2 objektif utama iaitu, menyediakan set data untuk membina model PEN teks media sosial Bahasa Melayu dan Membangunkan model PEN bagi teks media sosial Bahasa Melayu dengan menggunakan kaedah BiLSTM-CRF. Kedua-dua objektif ini tercapai dengan jayanya dan dibuktikan melalui pengujian sistem di bahagian Hasil Kajian.

Kesimpulannya, PEN bagi teks media sosial Twitter Bahasa Melayu telah berjaya dibangunkan meskipun menghadapi beberapa kekangan. Sistem PEN menggunakan model BiLSTM-CRF memperoleh ketepatan rendah, namun telah dibuktikan bahawa kaedah BiLSTM-CRF dapat ditingkatkan sekiranya set data berskala besar digunakan.

7 RUJUKAN

Appen. 2020. What is Training Data?. https://appen.com/blog/training-data/ [27 Mei 2022]

Arnaud Stiegler. 2021. Exploring Conditional Random Fields for NLP Applications.

https://hyperscience.com/tech-blog/exploring-crfs-for-nlp-applications/#:~:text=CRFs%20are%20used%20for%20a,implementations%20on%20major%20benchmark%20datasets. [27 Mei 2022]

Chew. 2021. Model Rangkaian Neural Bagi Penandaan Golongan Kata Pada Teks Media Sosial Bahasa Melayu.

Deeplizard. 2017. Machine Learning & Deep Learning Fundamentals

https://deeplizard.com/learn/video/DEMmkFC6IGM [22 Oktober 2021]

Eric Hofesmann. 2021. The Machine Learning Lifecycle in 2021

https://towardsdatascience.com/the-machine-learning-lifecycle-in-2021-473717c633bc [29 Oktober 2021]

Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. *Speech recognition with deep recurrent neural networks*. In: 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, pp. 6645–6649.

Jason Brownlee. 2019. Loss and Loss Functions for Training Deep Learning Neural Networks.

https://machinelearningmastery.com/loss-and-loss-functions-for-training-deep-learning-neural-networks/ [21 Oktober 2021]

Jason Brownlee. 2019. How to Visualize a Deep Learning Neural Network Model in Keras.

https://machinelearningmastery.com/visualize-deep-learning-neural-network-model-keras/ [24 Mei 2022]

Jason M. 2020. How to choose between rules-based vs. machine learning system

https://searchenterpriseai.techtarget.com/feature/How-to-choose-between-a-rules-based-vs-machine-learning-system [20 Oktober 2021]

Liu, Liyuan, Jingbo Shang, Xiang Ren, Frank Fangzheng Xu, Huan Gui, Jian Peng, and Jiawei Han. 2018. *Empower sequence labeling with task-aware neural language model*. In: Thirty-Second AAAI Conference on Artificial Intelligence.

Priya Pedamkar. 2020. Machine Learning Life Cycle

https://www.educba.com/machine-learning-life-cycle/ [29 Oktober 2021]

Rayner Alfred, Leow Chin Leong, Chin Kim On, and Patricia Anthony. 2014. *Malay Named Entity Recognition Based on Rule-Based Approach*. In: International Journal of Machine Learning and Computing (Volume 4: No 3).

Rrubaa Pachendrarajan & Aravindh Amaresan. 2018. Bidirectional LSTM-CRF for Named Entity Recognition.

Sadman Kabir Soumik. 2020. How to Calculate Confusion Matrix Manually https://medium.com/analytics-vidhya/how-to-calculate-confusion-matrix-manually-14292c802f52 [17 Disember 2021]

Sybren Jansen. 2021. Who's Who and What's What: Advances in Biomedical Named Entity Recognition(BioNER) https://towardsdatascience.com/whos-who-and-what-s-what-advances-in-biomedical-named-entity-recognition-bioner-c42a3f63334c [17 Disember 2021]

Yangzeng Li, Tingwen Liu, Diying Li, Quangang Li, Jinqiao Shi. 2018. *Character-based BiLSTM-CRF Incorporating POS and Dictionaries for Chinese Opinion Target Extraction*. In: Proceedings of Machine Learning Research. pp. 518-533.

Zhang, Boliang, Hongzhao Huang, Xiaoman Pan, Sujian Li, Chin-Yew Lin, Heng Ji, Kevin Knight, Zhen Wen, Yizhou Sun, Jiawei Han, et al. 2015. *Contextaware entity morph decoding*. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International 31 Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 586–595.

Putri Maya Syakilla Binti Hairol Akhma (A175210)

Sabrina Tiun

Fakulti Teknologi & Sains Maklumat,

Universiti Kebangsaan Malaysia