



JURNAL BAHASA, SASTERA DAN BUDAYA

<https://e-journal.uum.edu.my/index.php/rentas>

How to cite this article:

Nomoto, H., D. Moeljadi, & Farhan Athirah Abdul Razak (2024). Masalah teknologi dan isu sosial berkaitan penggunaan ChatGPT dalam bahasa Melayu. *RENTAS: Jurnal Bahasa, Sastra dan Budaya*, 3, Julai, 1-23. <https://doi.org/10.32890/rentas2024.3.1>

MASALAH TEKNOLOGI DAN ISU SOSIAL BERKAITAN PENGUNAAN CHATGPT DALAM BAHASA MELAYU

*(Technological and Social Issues
Related to Using ChatGPT in Malay)*

¹Hiroki Nomoto ²David Moeljadi &
³Farhan Athirah Abdul Razak

^{1&3}Tokyo University of Foreign Studies

²Kanda University of International Studies

¹Corresponding author: nomoto@tufs.ac.jp

Received: 14/1/2024 Revised: 29/4/2024 Accepted: 1/5/2024 Published: 16/7/2024

ABSTRAK

Kajian ini menganalisis masalah yang dihadapi oleh ChatGPT apabila seseorang menggunakannya dalam bahasa yang serupa, dengan menjadikan bahasa Melayu dan bahasa Indonesia sebagai contoh, yakni: ChatGPT sering menjana respons kepada prom bahasa Melayu (bahasa yang memiliki bilangan penutur yang sedikit) dalam bahasa Indonesia (bahasa yang mempunyai jumlah penutur yang lebih ramai). Pemerhatian terhadap keupayaan identifikasi bahasa (LangID) ChatGPT dibuat untuk mengenal pasti punca timbulnya masalah pemilihan bahasa tersebut, sama ada diakibatkan oleh ralat LangID ataupun sebab selainnya. Hasil pemerhatian mendapati ralat LangID tidak dapat menjelaskan keparahan masalah tersebut. Berdasarkan perbandingan pola respons kepada prom bahasa Melayu dan prom bahasa Jawa, kajian ini menyimpulkan bahawa sebab utama berlaku

masalah tersebut adalah kerana ChatGPT tidak menganggap bahasa Melayu dan bahasa Indonesia secara sama rata sebagai bahasa yang berlainan. ChatGPT sebaliknya menganggap bahasa Melayu adalah ragam tidak standard kepada bahasa Indonesia. Selain itu, kajian ini juga membincangkan isu sosial yang ditimbulkan oleh masalah pemilihan bahasa dan cara penyelesaian yang wajar dipertimbangkan bagi mengatasinya.

Kata kunci: ChatGPT, bahasa Melayu, bahasa Indonesia, bahasa dan teknologi, pemprosesan bahasa tabii.

ABSTRACT

We report a problem that one faces when using ChatGPT in similar languages, taking Malay and Indonesian as examples: ChatGPT often responds to prompts in Malay (the language with fewer speakers) in Indonesian (the language with more speakers). We examined ChatGPT's language identification (LangID) ability to find out whether this language choice problem arises from LangID errors. The results show that LangID errors alone cannot explain the problem's severity. By comparing the patterns of responses to Malay prompts and those to Javanese prompts, we conclude that the problem happens mainly because ChatGPT does not treat Malay and Indonesian equally as distinct languages. Rather, it behaves as if Malay were a non-standard variety of Indonesian. We also discuss social issues the language choice problem causes and possible solutions to them.

Keywords: ChatGPT, Malay, Indonesian, language and technology, natural language processing.

PENGENALAN¹

Apabila ChatGPT dan kecerdasan buatan (*artificial intelligence*, AI) generatif lain yang seangkatan dengannya menjadi semakin popular, maka masalah dan keterbatasan teknologi tersebut terbuka untuk diketahui dan dikongsi secara meluas oleh masyarakat. Walaupun sesetengah masalah dan keterbatasan itu bersifat umum dan mengenai

¹ Kebanyakan kandungan makalah ini adalah berdasarkan Nomoto (2023). Namun begitu, penambahbaikan telah dilakukan supaya perbincangannya sesuai dengan pembaca di Malaysia.

semua pengguna (Ray, 2023), namun terdapat juga masalah dan keterbatasan yang sangat khusus hanya kepada kumpulan pengguna tertentu. Hal yang kedua ini boleh kekal tanpa disedari, dan mungkin diabaikan sahaja disebabkan oleh kekhususannya itu. Akan tetapi, perkara tersebut bakal mengundang kesulitan kepada para pengguna jikalau tidak dikenal pasti punca dan kaedah untuk mengatasinya. Maka, kajian ini bertujuan membincangkan isu yang dinyatakan. Secara spesifik, kajian ini berhasrat menampilkan analisis tentang masalah dan keterbatasan yang dialami oleh ChatGPT dalam soal pemilihan bahasa yang salah melibatkan dua bahasa yang serupa dan sekeluarga, iaitu bahasa Melayu dan bahasa Indonesia.

Dengan lebih terperinci, hal yang terkait dengan masalah ini ialah ChatGPT cenderung memberikan respons kepada prom (*prompt*) berbahasa Melayu dalam bahasa Indonesia, iaitu bahasa yang lebih ramai penutur justeru lebih banyak data. Pengkaji berpendapat bahawa masalah teknologi seperti ini berpotensi untuk menjelmakan isu sosial lain yang berkaitan. Sehubungan dengan itu, dalam makalah ini, pengkaji berhasrat mengemukakan beberapa cadangan konkrit untuk mengurangkan masalah tersebut justeru berdaya pula menangani isu sosial yang terimplikasi daripadanya.

Dari segi cakupan keseluruhan perbincangan, makalah ini disusun seperti berikut: bahagian “BAHASA MELAYU DAN BAHASA INDONESIA” menerangkan secara ringkas hubungan antara bahasa Melayu dan bahasa Indonesia. Pengkaji mengemukakan contoh konkrit masalah pemilihan bahasa yang disebutkan di atas dalam bahagian “MASALAH PEMILIHAN BAHASA”. Kemudian, pengkaji meneliti kemungkinan kegagalan identifikasi bahasa (*language identification*; LangID) yang menjadi punca kepada masalah yang dialami oleh ChatGPT dalam bahagian “KEUPAYAAN LANGID CHATGPT”. Bahagian “ISU SOSIAL DAN KEMUNGKINAN PENYELESAIAN” membincangkan isu sosial yang berpunca daripada masalah pemilihan bahasa. Akhir sekali, bahagian “KESIMPULAN” menyimpulkan keseluruhan perbincangan makalah ini.

BAHASA MELAYU DAN BAHASA INDONESIA

Dari segi linguistik, bahasa Melayu (ISO693-3 zsm) dan bahasa Indonesia (ISO693-3 ind) merupakan dua kelainan standard bagi bahasa yang sama. Kedua-dua bahasa ini dituturkan di dua kawasan

berbeza. Berpandukan fakta linguistik juga, bahasa Melayu makro (ISO693-3 msa) merupakan pangkal kedua-dua kelainan standard tersebut, maka bahasa ini merangkumi semua kelainan bahasa Melayu yang tersebar di seluruh Kepulauan Melayu. Jika direnungkan dengan lebih mendalam, nomenklatur “bahasa Melayu” adalah taksa. Hal ini demikian kerana maksudnya boleh merujuk kepada bahasa makro, yakni bahasa Melayu (msa), juga salah satu kelainannya (zsm). Oleh sebab itulah dalam makalah ini, pengkaji menggunakan nomenklatur “bahasa Melayu” bagi merujuk kepada denotasinya yang kedua. Maka, yang pengkaji maksudkan dengan bahasa Melayu ialah bahasa rasmi di negara Malaysia dan Brunei serta merupakan salah satu daripada empat bahasa rasmi di Singapura. Bahasa Indonesia pula ialah bahasa rasmi bagi negara Indonesia. Bilangan penutur bagi kedua-dua bahasa ini ialah kira-kira 32 juta untuk bahasa Melayu dan 270 juta untuk bahasa Indonesia. Fakta umumnya ialah bahasa Melayu dan bahasa Indonesia berkongsi sebahagian besar perbendaharaan kata dan tatabahasa. Asmah Omar (2001) menganggarkan perbezaan leksikal antara kedua-dua bahasa ini hanya sekitar 10%. Perbezaan sintaksis pula sememangnya wujud antara bahasa Melayu dengan bahasa Indonesia, namun tidak begitu jelas berbanding dengan perbezaan leksikal; lihat contoh berkaitan dengan perkara ini dalam Nomoto dan Kartini Abd. Wahab (2011).

MASALAH PEMILIHAN BAHASA

Masalah pemilihan bahasa telah lama wujud dalam aplikasi pemprosesan bahasa tabii atau *natural language processing* (NLP) yang melibatkan bahasa Melayu dan bahasa Indonesia. Secara yang lebih spesifik, masalah pemilihan bahasa yang dimaksudkan ini merujuk kepada keadaan aplikasi NLP yang gagal membuat pemisahan mutlak antara bahasa Melayu dengan bahasa Indonesia. Masalah seperti ini tetap wujud dalam terjemahan mesin meskipun terdapat kemajuan yang belum pernah terjadi sebelumnya seperti mana yang dapat dikesan dalam beberapa tahun kebelakangan ini dalam bidang NLP. Contoh konkrit tentang masalah ini ialah terjemahan bahasa Inggeris ke dalam bahasa Melayu kadangkala mengandungi ungkapan daripada bahasa Indonesia. Masalah tersebut sangat ketara berlaku dalam terjemahan leksikal, tetapi jarang sekali pada peringkat ayat. Hal ini bermakna terjemahan mesin tidak mungkin memberikan terjemahan sepenuhnya dan setepatnya dalam bahasa Melayu apabila inputnya ialah ungkapan yang sangat pendek yang hanya terdiri daripada beberapa perkataan.

Masalah pemilihan bahasa dengan ChatGPT jauh lebih berat berbanding dengan terjemahan mesin. Sewaktu pengoperasian ChatGPT, respons kepada prom yang diberikan dalam bahasa Melayu kerap dikemukakan dalam bahasa Indonesia. Oleh sebab prom adalah dalam bentuk ayat lengkap, maka hal ini boleh dikatakan menimbulkan masalah pada peringkat ayat. Dalam konteks bahasa Melayu-bahasa Indonesia, masalah seperti ini banyak berlaku dengan prom yang ditulis menggunakan bahasa Melayu. Kesannya, pengguna Indonesia pastilah tidak menyedari kewujudan masalah tersebut kerana responsnya sentiasa tampil dalam bahasa Indonesia jika promnya dalam bahasa mereka sendiri, iaitu bahasa Indonesia.

Untuk menunjukkan betapa beratnya masalah yang dimaksudkan, pengkaji menguji 100 prom diberikan dalam bahasa Melayu secara rawak (*random*).² Prom tersebut sedapat mungkin ditulis dengan jelas dalam bahasa Melayu. Topik yang disentuh oleh prom melibatkan pelbagai domain, termasuk sains, kesihatan, sejarah, politik, bahasa, sastra dan agama. Pengkaji menggunakan versi percuma ChatGPT (3 Mei). Daripada 100 prom yang diberikan, terdapat sebanyak 66 respons yang dinukilkan dalam bahasa Indonesia, seperti dalam (1), manakala 31 jawapan pula diberikan dalam bahasa Melayu. Terdapat juga tiga respons yang mengandungi unsur kedua-dua bahasa, seperti dalam (2). Item yang dicondongkan dengan subskrip M dan I adalah masing-masing unik untuk bahasa Melayu dan bahasa Indonesia.

(1) Pengguna: Apakah *perbezaan_M* bahasa Melayu dengan bahasa Indonesia?

ChatGPT: Bahasa Melayu dan bahasa Indonesia memiliki banyak persamaan *karena_I*, *keduanya_I* berasal dari bahasa Melayu Polinesia dan memiliki sejarah yang *terkait_I*. Namun, ada beberapa *perbedaan_I* dalam kosa kata, *ejaan_I*, *tata bahasa_P*, dan *pengucapan_I* antara *keduanya_P*.

Secara umum, bahasa Indonesia lebih dipengaruhi oleh bahasa Belanda, sedangkan bahasa Melayu lebih dipengaruhi oleh bahasa Arab dan Sanskrit. Selain itu, ada *perbedaan_I*, *regional_I*, antara *kedua_I* bahasa yang menghasilkan variasi dalam kosa kata dan *tata bahasa_P*.

Misalnya, dalam bahasa Melayu Malaysia, kata “kereta” digunakan untuk merujuk pada *mobil_P*, sementara dalam bahasa Indonesia, kata

² Ujian yang sama tidak dijalankan untuk bahasa Indonesia kerana rakyat Indonesia nampaknya tidak mengalami masalah pemilihan bahasa yang berkaitan.

yang lebih sering digunakan adalah “mobil”. Ada juga beberapa *perbedaan*₁ dalam *pengucapan*₁ dan ejaan, seperti *pengucapan*₁ “s” yang lebih sering digunakan di bahasa Melayu daripada di bahasa Indonesia.

Namun, meskipun ada *perbedaan*₁ antara *keduanya*_p, bahasa Melayu dan bahasa Indonesia masih sangat mirip satu sama lain dan dapat saling *dimengerti*₁ oleh penutur *kedua*₁ bahasa.

(2) Pengguna: Bagaimana *Jepun*_M menjajah Malaysia?

ChatGPT: *Jepun*_M tidak pernah menjajah Malaysia dalam *arti*₁ tradisional. Namun, pada masa Perang Dunia II, *Jepun*_M memasuki Malaysia dan menguasai wilayah tersebut selama *tiga setengah tahun*_p, dari 1941 hingga 1945. Ini adalah *periode*₁ yang penuh dengan penderitaan dan kesengsaraan bagi penduduk Malaysia, terutama yang hidup di bawah pemerintahan *Jepun*_M.

Pada *saat*₁ itu, Malaya (nama lama untuk Malaysia) adalah koloni British yang penting dan kaya akan sumber daya alam seperti timah dan *karet*_r. *Jepun*_M menyatakan perang terhadap *Inggris*₁ pada tahun 1941 dan memulai serangan ke seluruh wilayah Asia Tenggara, termasuk Malaya.

Pada awalnya, pasukan *Inggris*₁ mempertahankan Malaya, tetapi pada bulan Januari 1942, pasukan *Jepun*_M *berhasil*₁ merebut Kuala Lumpur, ibu kota Malaya, dan memaksa *tentara*₁ *Inggris*₁ untuk mundur ke Singapura. Pada bulan Februari 1942, Singapura juga jatuh ke tangan *Jepun*_M, dan lebih dari 80.000 *tentara*₁ *Inggris*₁ dan sekutunya ditawan sebagai tawanan perang.

Di bawah pemerintahan *Jepun*_M, penduduk Malaysia dikenakan hukuman yang sangat keras dan dipaksa untuk bekerja sebagai buruh paksa. *Banyak*₁ orang Malaysia yang ditangkap, *disiksa*_p, atau *dieksekusi*₁ oleh pasukan *Jepun*_M. *Banyak*₁ juga wanita Malaysia yang diperkosa dan *disiksa*₁ oleh *tentara*₁ *Jepang*₁ selama pendudukan tersebut.

Setelah Perang Dunia II berakhir, *Jepun*_M menyerah pada Sekutu dan meninggalkan Malaysia pada tahun 1945. Pemerintahan kolonial *Inggris*₁ kemudian dipulihkan di Malaysia, tetapi semangat kemerdekaan mulai tumbuh di kalangan rakyat Malaysia, dan pada tahun 1957, Malaysia memperoleh kemerdekaannya dari *Inggris*_r.

Perkara yang harus diberikan perhatian ialah terdapat juga sedikit percampuran kod melibatkan bahasa Melayu dan bahasa Indonesia

seperti yang ditunjukkan dalam (2), namun respons ChatGPT yang muncul dalam keadaan yang sangat semula jadi. Hal ini cukup menakjubkan memandangkan kecilnya jumlah data bahasa Melayu dan bahasa Indonesia dalam set data latihan ChatGPT (Jadual 1). Tambahan pula, percampuran kod, yang dianggap berpunca daripada pengetahuan bahasa sasaran yang tidak mencukupi, berlaku hanya dalam tiga respon daripada 100 respon yang diberikan oleh ChatGPT.

Jadual 1

Saiz Set Data Latihan ChatGPT Mengikut Kiraan Perkataan (Brown et al., 2020)³

Bahasa	Jumlah perkataan	Peratusan jumlah perkataan
Inggeris	181,014,683,608	92.64708%
Perancis	3,553,061,536	1.81853%
Belanda	669,055,061	0.34244%
Jepun	217,047,918	0.11109%
Cina (Mandarin)	193,517,396	0.09905%
Indonesia	116,930,321	0.05985%
Melayu	13,389,340	0.00685%
Tamil	5,163,171	0.00264%
Jawa	1,448	0.00000%

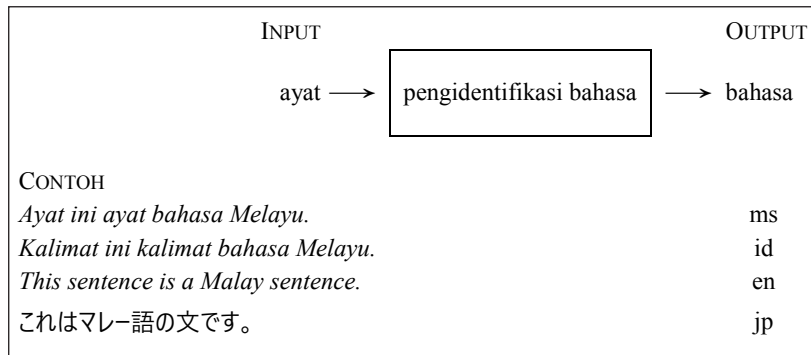
KEUPAYAAN LANGID CHATGPT

Mengapakah masalah pilihan bahasa yang dibincangkan di atas berlaku? Hipotesis pengkaji bagi menjawab persoalan ini terletak pada identifikasi bahasa (*language identification*; LangID). Dalam hal ini LangID ialah salah satu tugas NLP yang bertujuan untuk mengidentifikasikan bahasa bagi sesuatu input dalam bentuk teks. Dalam konteks ChatGPT, input yang diberikan biasanya merupakan ayat berbentuk soalan atau ayat tanya. Proses mengidentifikasi bahasa berlaku sebelum ChatGPT menjana output dalam bentuk respons kepada pengguna (lihat Rajah 1).

³ https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv, diakses pada 16 September 2023.

Rajah 1

Proses Mengidentifikasi Bahasa



Secara lebih terperinci, hipotesis pengkaji boleh diungkapkan seperti ini: apabila ChatGPT memberi respons kepada prom bahasa Melayu dalam bahasa Indonesia, maka ChatGPT tersalah mengenal pasti bahasa prom itu sebagai bahasa Indonesia dan memutuskan untuk membalasnya dengan menggunakan bahasa Indonesia. Untuk menguji hipotesis itu, pengkaji memeriksa keupayaan LangID ChatGPT. Pada masa yang sama, pengkaji juga memeriksa keupayaan LangID penutur asli dan Terjemahan Google (*Google Translate*) untuk perbandingan.

Metodologi

Pengkaji meminta ChatGPT mengenal pasti bahasa berdasarkan 600 ayat (300 ayat untuk setiap bahasa).

Data Ujian

Ayat-ayat yang terlibat diambil daripada data ujian yang diatur oleh Nomoto et al. (2018). Data tersebut terdiri daripada tiga komponen, iaitu berita, wiki dan fiksyen. Setiap komponen mengandungi 100 ayat bagi bahasa yang terlibat. Komponen berita terdiri daripada artikel dua akhbar tempatan versi dalam talian, iaitu *Sinar Harian*⁴ untuk bahasa Melayu dan *Kompas*⁵ untuk bahasa Indonesia.

Komponen wiki dibuat berdasarkan *Asian Language Treebank*

⁴ <https://www.sinarharian.com.my/>

⁵ <https://www.kompas.com/>

Parallel Corpus (Riza et al., 2016),⁶ yang dibina dengan menterjemahkan artikel *Wikinews* berbahasa Inggeris. Komponen fiksiyen pula melibatkan cerpen yang dikumpul daripada koleksi cerpen dalam talian, iaitu *Penulisan2u*⁷ untuk bahasa Melayu dan *Cerpenmu*⁸ bagi bahasa Indonesia. Untuk setiap komponen, pengkaji mengambil dua ayat pertama daripada 50 fail pertama. Sekiranya ayat tersebut didapati lebih pendek daripada empat perkataan atau terdapat ayat bahasa Inggeris hasil daripada penukaran kod, maka dalam kes seperti ini pengkaji mengambil ayat seterusnya. 600 ayat yang telah dikumpulkan itu disusun dan dinomborkan secara rawak.

Prom

Pengkaji menggunakan empat prom dalam bahasa Inggeris, dengan mengandaikan bahawa ChatGPT beroperasi dengan baik dalam bahasa Inggeris.

Prom 1-3 berbeza dalam ungkapan yang merujuk kepada bahasa. Prom ini bermula dengan arahan dalam (3), diikuti dengan ayat ujian. Prom 4 tidak menyatakan pilihan bahasa sama sekali, seperti dalam (4).

(3) **Prom 1-3:** What languages are the following sentences written in,

{	1. Melayu atau Indonesia	}	?
	2. “id” atau “ms”		
	3. Malaysia atau Indonesia		

For each sentence, choose one answer. No explanation is necessary.

(4) **Prom 4:** Identify the languages of the following sentences. No explanation is necessary.

Prom 2 menggunakan kod bahasa ISO639-3 kerana pengkaji mengandaikan bahawa data yang ada pada ChatGPT dilatih untuk mengandungi kod bahasa ini. Bahasa Melayu dirujuk sebagai “Malaysian” dalam Prom 3 untuk mengelakkan kekeliruan yang mungkin berlaku antara “bahasa Melayu” sebagai bahasa makro (msa) dan “bahasa Melayu” sebagai kelainan standard yang digunakan di Malaysia (zsm) (lihat bahagian “BAHASA MELAYU DAN BAHASA INDONESIA” untuk kesamaran ini).

⁶ <https://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/index.html>, dimuat turun pada 19 Julai 2017

⁷ <http://www.penulisan2u.my/>

⁸ <https://cerpenmu.com/>

Tetapan Ujian

Versi percuma ChatGPT telah digunakan. Versi ChatGPT tersebut ialah “3 Mei” semasa ujian dengan Prom 1–2 dijalankan. Versi telah dinaik taraf kepada “12 Mei” apabila ujian dengan Prom 3–4 dijalankan. Disebabkan terdapat had panjang prom, ayat ujian terpaksa dipecahkan kepada bahagian-bahagian kecil yang terdiri daripada kira-kira 25 ayat.

Bagi ujian perbandingan, penutur asli (penulis kedua dan ketiga makalah ini, yang masing-masing merupakan penutur asli bahasa Indonesia dan bahasa Melayu) digunakan bagi mengklasifikasikan ayat ujian secara manual sama ada sebagai bahasa Melayu ataupun bahasa Indonesia. Untuk ujian dengan Terjemahan Google, fungsi *DETECTLANGUAGE* telah digunakan.

Dapatan Kajian dan Perbincangan

Jadual 2 menunjukkan dapatan keseluruhan. Dapatan yang lebih terperinci ditunjukkan dalam LAMPIRAN. Pengkaji mengira metrik penilaian dengan bahasa Melayu sebagai kelas positif.

Jadual 2

Keupayaan LangID ChatGPT, Penutur Asli (Manusia) dan Terjemahan Google

		Bahasa Melayu	Bahasa Indonesia	Ketepatan (precision)	Perolehan (recall)	F1
ChatGPT	Prom 1	414	176	0.67	0.93	0.78
ChatGPT	Prom 2	285	296	0.75	0.71	0.73
ChatGPT	Prom 3	215	381	0.89	0.64	0.74
ChatGPT	Prom 4	332	266	0.76	0.84	0.79
Manusia	Moeljadi	269	331	0.96	0.86	0.91
Manusia	Farhan	313	287	0.90	0.94	0.92
Terjemahan Google		290	309	0.95	0.92	0.94

Walaupun pengkaji secara jelas mengarahkan ChatGPT untuk pilih antara bahasa Melayu dan bahasa Indonesia, tetapi ternyata ia merespons dengan bahasa lain juga, termasuk bahasa Inggeris (Prom 1: 7; Prom 2: 16; Prom 3: 4; Prom 4: 2), bahasa Melayu/bahasa Inggeris (Prom 1: 3), bahasa Indonesia/bahasa Inggeris (Prom 2: 1) dan bahasa Indonesia/bahasa Melayu (Prom 2: 2). Terjemahan

Google pula mengenal pasti satu ayat sebagai bahasa Hawaii, dan ini menyebabkan nombor dalam Jadual 2 tidak ditambah sehingga 600 untuk ChatGPT dan Terjemahan Google.

Maklumat dalam jadual ini menunjukkan pencapaian yang tinggi bagi penutur asli kedua-dua bahasa. Dalam hal ini terdapat beberapa ayat yang boleh diterima dalam kedua-dua bahasa, dan dianggap mustahil untuk mengklasifikasikannya ke dalam salah satu daripada dua bahasa tersebut. Faktanya ialah kesilapan pengenalan bagi ayat sedemikian bukanlah kekhilafan yang sebenar. Akan tetapi, ia timbul disebabkan bentuk tugas, yakni, pilihan “bahasa Melayu/Indonesia” tidak disediakan selain pilihan “bahasa Melayu” dan “bahasa Indonesia”. Walau bagaimanapun, maklumat dalam Jadual 1 juga menunjukkan pencapaian ChatGPT jauh lebih rendah berbanding dengan kemampuan manusia.

Hipotesis pengkaji yang mengandaikan bahawa ChatGPT merespons prom bahasa Melayu dalam bahasa Indonesia kerana ralat LangID hanya betul sebahagian, tetapi salah bagi bahagian yang lain. Hal ini dibuktikan dengan perolehan yang sangat tinggi bagi Prom 1 yang bertentangan dengan hakikat bahawa masalah pemilihan bahasa agak lazim dalam bahasa Melayu tetapi tidak dalam bahasa Indonesia. Perolehan Prom 2–4 pun masih tinggi memandangkan dapatan ujian yang menggunakan 100 prom Melayu yang dibentangkan dalam bahagian “MASALAH PEMILIHAN BAHASA”, di mana ChatGPT merespons sebanyak 66 daripada 100 prom dalam bahasa Indonesia. Oleh itu, pengkaji menyimpulkan bahawa masalah utama bukan terletak pada LangID prom tetapi pada pilihan bahasa dalam penjaan respons. Kesimpulan ini dapat disahkan apabila ChatGPT berupaya mengenal pasti bahasa kesemua 100 prom bahasa Melayu dengan betul dalam ujian yang dilakukan oleh pengkaji dengan menggunakan Prom 2. Hal ini menunjukkan ChatGPT boleh mengidentifikasi bahasa yang digunakan sebagai prom.

Selain itu, bagi mendukung kesimpulan yang ditampilkan, pengkaji menyertakan bukti daripada bahasa Jawa, yakni bahasa yang dituturkan di Indonesia justeru berkaitan erat dengan bahasa Melayu dan bahasa Indonesia. Dalam hal ini ChatGPT lazimnya tidak menjawab prom bahasa Jawa secara sepenuhnya dalam bahasa ini. ChatGPT sebaliknya hanya menjawab beberapa ayat pertama dalam bahasa Jawa, manakala selebihnya adalah dalam bahasa Indonesia.⁹

⁹ Perkara yang sama terjadi kepada dialek bukan standard bahasa Melayu. Sebagai contoh, ia terjadi kepada dialek Kedah. ChatGPT pada awalnya cuba menyesuaikan dan memberi respons dalam dialek Kedah tetapi pada pertengahan respons ia menjawab dengan dialek standard bahasa Melayu.

Sebagai contoh, apabila pengkaji bertanya soalan “Bagaimana anda membuat baju kebaya?” dalam bahasa Jawa (*kepriye carane nggawe baju kebaya?*), ChatGPT hanya menyediakan tiga ayat pertama sebagai jawapan dalam bahasa Jawa, kemudian penukaran kod yang tidak diinginkan berlaku apabila baki sembilan ayat semuanya diungkapkan dalam bahasa Indonesia. Oleh itu, pengkaji meneruskan dengan mengarahkan ChatGPT untuk menjawab dalam bahasa Jawa (*jawaben nganggo basa jawa*). Ternyata jawapan ChatGPT tidak mengandungi ayat bahasa Indonesia, tetapi pengkaji menemui perkataan dan frasa bahasa Indonesia dalam hampir keseluruhan ayat. Berkait dengan perkara ini juga ChatGPT kelihatan menganggap bahawa bahasa Jawa adalah bahasa yang berlainan daripada bahasa Indonesia. Walau bagaimanapun, ChatGPT gagal bertindak balas sepenuhnya dalam bahasa Jawa. Hal ini mungkin disebabkan oleh data latihan yang tidak mencukupi (lihat Jadual 1) justeru mengakibatkan penukaran kod dan percampuran kod.

Dalam kes bahasa Melayu, penukaran kod di tengah-tengah respons kelihatan tidak berlaku. Respons yang bermasalah adalah seperti ini: sama ada sepenuhnya dalam bahasa Indonesia (1) atau dalam bahasa Melayu bercampur dengan perkataan dan frasa bahasa Indonesia (2).

Perbezaan ini menunjukkan bahawa ChatGPT tidak menganggap bahasa Melayu sebagai bahasa yang berlainan daripada bahasa Indonesia dengan cara yang sama seperti bahasa Jawa. Sebaliknya, ia menganggap bahasa Melayu seolah-olah adalah dialek bukan standard atau laras bukan formal bagi bahasa Indonesia. Adalah diketahui bahawa ChatGPT bertindak balas dalam bahasa standard yang formal sesuatu bahasa, tanpa mengira dialek dan laras prom, seperti mana yang dipaparkan dalam contoh (5) di bawah.

(5) Pengguna: Macam mana nak ke UUM dari KL?

ChatGPT: Untuk pergi ke Universiti Utara Malaysia (UUM) dari Kuala Lumpur, terdapat beberapa pilihan pengangkutan yang boleh anda pertimbangkan. Berikut adalah beberapa cara untuk mencapai ke destinasi tersebut: ...

Tidak dinafikan bahawa dari segi ciri linguistik, bahasa Melayu dan bahasa Indonesia adalah dua kelainan bahasa yang sama (lihat bahagian “BAHASA MELAYU DAN BAHASA INDONESIA”). Namun begitu, adalah tidak wajar untuk menganggapnya satu sebagai standard dan satu lagi sebagai bukan standard. Kedua-duanya adalah kelainan standard, dan oleh itu mesti dibezakan dengan jelas dalam aplikasi praktikal seperti ChatGPT.

ISU SOSIAL DAN KEMUNGKINAN PENYELESAIAN

Masalah pemilihan bahasa yang dibincangkan bukan sekadar masalah teknologi tetapi membawa kepada isu sosial. Dalam bahagian ini, pengkaji membincangkan beberapa daripada masalah sosial dan mengusulkan penyelesaian untuk mengatasinya.

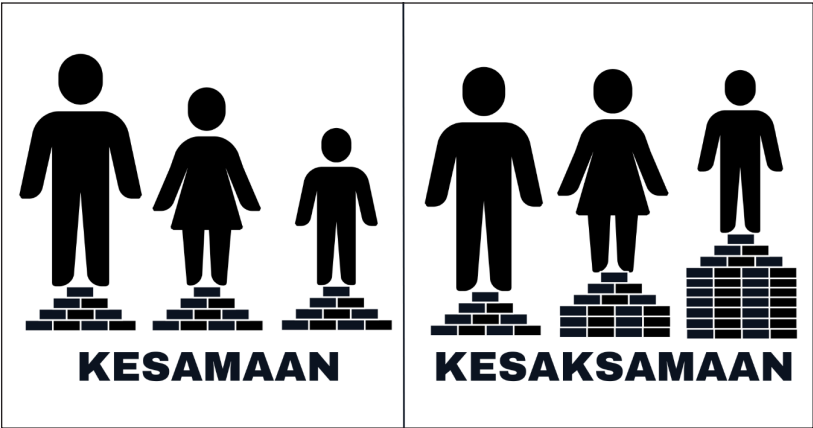
Isu Sosial Disebabkan oleh ChatGPT

Ketidaksamaan dan Ketaksaksamaan Linguistik

Ketidaksamaan (*inequality*) dan ketaksaksamaan (*inequity*) merupakan konsep yang serupa tetapi sebenarnya berbeza (Rajah 2). *Ketidaksamaan* ialah kata negatif daripada perkataan *kesamaan*, yang bermaksud semua pihak diberi benda yang sama. Tahap yang boleh dicapai adalah berbeza-beza mengikut situasi yang dialami oleh setiap pihak. *Ketaksaksamaan* pula ialah kata negatif daripada perkataan *kesaksamaan*. Kata ini bermaksud semua pihak boleh mencapai matlamat yang sama. Demi tujuan tersebut, apa-apa sahaja yang diberikan boleh berbeza-beza mengikut situasi yang dialami oleh setiap pihak yang terlibat.

Rajah 2

Konsep Kesamaan dan Kesaksamaan



Berbalik kepada masalah ChatGPT yang menjadi mauduk perbincangan makalah ini, penutur bahasa Melayu selalunya tidak dapat menerima respons dalam bahasa mereka sedangkan penutur bahasa Indonesia selalu mendapatnya (ketidaksamaan). Akibatnya

ialah penutur bahasa Melayu tidak boleh menerima jumlah faedah yang sama seperti penutur bahasa Indonesia terima daripada ChatGPT (ketaksaksamaan). Penutur bahasa Melayu boleh mengelakkan respons bahasa Indonesia dengan kejuruteraan prom (*prompt engineering*). Kejuruteraan prom merujuk kepada proses merancang penyusunan perkataan yang baik untuk memperoleh jawapan yang dikehendaki semasa berinteraksi dengan program AI (contoh: “Jawab dalam bahasa Melayu, dan bukannya bahasa Indonesia”). Masalah dengan penyelesaian ini ialah usaha tambahan sedemikian tidak perlu dilakukan oleh penutur bahasa Indonesia. Maka, adalah menjadi tanggungjawab sosial penyedia perkhidmatan untuk memastikan kesamaan dan kesaksamaan dalam kalangan penutur.

Peralihan Bahasa

Jika ChatGPT terus merespons dalam bahasa yang berbeza daripada bahasa yang digunakan dalam prom, maka penutur bahasa yang kedua akan kecewa. Oleh sebab kebanyakan masyarakat di dunia adalah berdwibahasa atau berbilang bahasa, ramai yang berkemungkinan akan berhenti menggunakan bahasa pertama (B1) mereka dan memihak kepada bahasa kedua (B2). Hal ini akan mengurangkan input B1 kepada ChatGPT dan perbezaan prestasi akan merosot antara B1 dan B2, yang seterusnya boleh mendorong sebahagian penutur untuk beralih daripada B1 ke B2, sekurang-kurangnya dalam domain tertentu, termasuk dalam perkhidmatan IT. Dalam kes bahasa Melayu, kemungkinan besar penutur akan beralih kepada bahasa Inggeris, lebih-lebih lagi kerana kurangnya kesedaran tentang peranan bahasa Melayu dalam kalangan rakyat Malaysia.

Pengkaji percaya bahawa perkhidmatan IT ialah salah satu domain terpenting yang mempengaruhi daya hidup sesuatu bahasa. Slogan “bahasa dan sastera” mungkin sudah kurang relevan mengikut peredaran masa terkini. Sehubungan itu, slogan yang terkini perlulah beralih daripada “bahasa dan sastera” kepada “bahasa dan teknologi”. “Bahasa dan teknologi” yang akan menjadi pemangkin masa depan Malaysia ke arah kemajuan teknologi dan pembangunan ekonomi negara.

Ketidakseimbangan Kuasa Linguistik

Pada akhir bahagian “KEUPAYAAN LANGID CHATGPT,” pengkaji menunjukkan kemungkinan ChatGPT tidak memperlakukan bahasa Melayu dan bahasa Indonesia secara sama rata. Walaupun kedua-dua bahasa Melayu dan bahasa Indonesia ialah bahasa rasmi khususnya

bagi negara Malaysia dan Indonesia, tetapi ChatGPT berkelakuan seolah-olah bahasa Melayu ialah ragam yang tidak standard kepada bahasa Indonesia. Oleh itu, ChatGPT mewujudkan ketidakseimbangan kuasa antara dua bahasa yang tidak sepatutnya wujud. Seperti yang digambarkan pada Rajah 3, bahasa Melayu standard dan dialek-dialek bahasa Melayu di Malaysia dan Indonesia semuanya dikategorikan oleh ChatGPT sebagai kelainan bahasa Indonesia. Klasifikasi bahasa yang sebenar adalah seperti yang digambarkan pada Rajah 4, di mana kedua-dua bahasa Melayu di Malaysia dan bahasa Indonesia ialah kelainan bahasa Melayu sebagai bahasa makro.

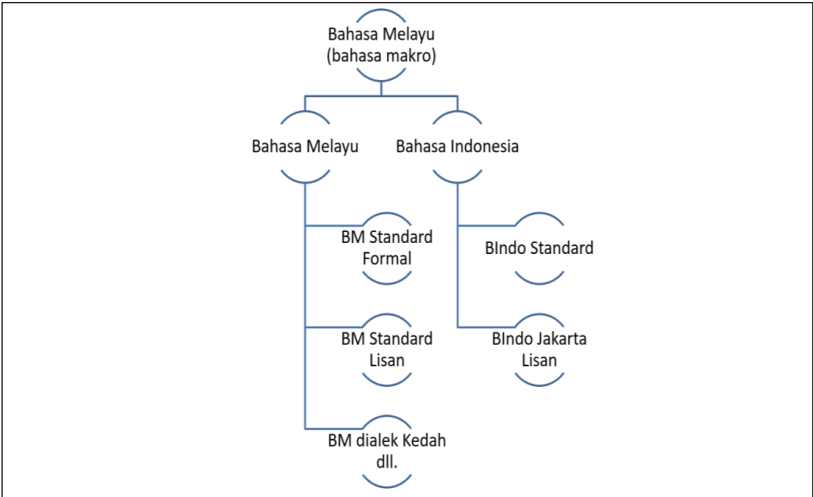
Rajah 3

Klasifikasi Bahasa oleh ChatGPT Yang Menunjukkan Bahasa Indonesia Lebih Tinggi daripada Bahasa Melayu



Rajah 4

Klasifikasi Bahasa Sebenar Yang Menunjukkan Kesamarataan Bahasa Melayu dan Bahasa Indonesia



Adalah mudah untuk membayangkan pewajaran ChatGPT mempamerkan kelakuan sedemikian, iaitu bahasa Indonesia mempunyai lebih ramai penutur berbanding bahasa Melayu (lihat bahagian “BAHASA MELAYU DAN BAHASA INDONESIA”), justeru lebih banyak data latihan tersedia (lihat Jadual 1). Oleh itu, tanpa campur tangan manusia yang sewajarnya, ChatGPT akan terus meluaskan jurang perbezaan.

Penyelesaian Masalah

Pertama, memperkenalkan tetapan bahasa. Dalam hal ini tetapan bahasa membolehkan pengguna menentukan bahasa yang berkeupayaan menghalang respons dalam bahasa yang tidak diinginkan. Walau bagaimanapun, ini bukanlah penyelesaian yang ideal jika hanya penutur bahasa “dominan” (bahasa Indonesia dalam konteks makalah ini) boleh menikmati fungsi identifikasi bahasa automatik. Oleh itu, keupayaan LangID mesti dipertingkatkan pada masa yang sama. Modul LangID khusus boleh dimasukkan ke dalam ChatGPT. Pengidentifikasian bahasa sedia ada adalah lebih baik daripada ChatGPT dan, dari segi bahasa Melayu dan Indonesia, boleh mencapai prestasi peringkat manusia (lihat Jadual 2) walaupun masih banyak ruang untuk penambahbaikan (Caswell et al., 2020).

Kedua, berkaitan dengan perkara pertama, perlu ada senarai bahasa yang perlu diperlakukan secara berasingan. Jika kedua-dua bahasa Melayu dan bahasa Indonesia disenaraikan secara bersama, maka ia akan dianggap sama rata sebagai bahasa yang berlainan, dan bukannya satu bahasa standard dan satu lagi tidak standard. Senarai sedemikian sudah tersedia dalam perkhidmatan terjemahan mesin seperti Terjemahan Bing dan Terjemahan Google. Walau bagaimanapun, tidak selalunya jelas mengapa sesetengah bahasa disertakan (dan ditawarkan dengan ciri tambahan tertentu) tetapi yang lain tidak, yang mana mungkin menimbulkan pelbagai spekulasi mengenai sikap penyedia perkhidmatan terhadap bahasa-bahasa yang berlainan dan penuturnya.

Akhir sekali, kerajaan juga boleh mengambil tindakan supaya rakyat mereka boleh mendapat manfaat daripada ChatGPT. Tindakan yang perlu berbeza dari negara ke negara. Sebagai contoh, kerajaan Iceland bekerjasama dengan OpenAI untuk meningkatkan keupayaan ChatGPT bagi mengendalikan bahasa Iceland. Berikut ialah petikan

langsung mengenai inisiatif yang diambil oleh Presiden Iceland sebagai usaha memelihara bahasa Iceland.

On the initiative of the country's President, HE Guðni Th. Jóhannesson, and with the help of private industry, Iceland has partnered with OpenAI to use GPT-4 in the preservation effort of the Icelandic language—and to turn a defensive position into an opportunity to innovate.¹⁰

Oleh sebab ChatGPT sudah dapat mengendalikan bahasa Melayu dengan baik, strategi Iceland tidak relevan kepada kerajaan Malaysia. Sebaliknya, ia boleh menggalakkan rakyat dan syarikatnya untuk menggunakan lebih banyak bahasa Melayu di internet untuk memperluaskan jumlah data web dalam bahasa Melayu.

Pada masa ini, banyak laman web korporat hanya tersedia dalam bahasa Inggeris sahaja, walaupun bahasa Inggeris bukanlah bahasa rasmi mahupun bahasa kebangsaan negara. Contohnya termasuk syarikat-syarikat besar seperti Toyota¹¹ dan Perodua.¹² Namun begitu, terdapat juga syarikat yang menggunakan bahasa Melayu seperti syarikat Mydin,¹³ Petronas,¹⁴ dan Proton.¹⁵ Malahan terdapat juga syarikat Malaysia yang hanya menggunakan bahasa Cina Mandarin dan bahasa Inggeris sahaja seperti Kopi Ah Huat,¹⁶ *Sing Sing Coffee*¹⁷ dan *Julie's Biscuit*.¹⁸ Keadaan ini tidak sepatutnya berlaku di Malaysia lebih-lebih lagi apabila usaha memartabatkan bahasa Melayu sedang hangat diperkatakan.

Di samping itu, kerajaan Malaysia, mungkin boleh bekerjasama dengan kerajaan Brunei dan Singapura, meminta OpenAI dan Common Crawl, iaitu sumber utama set data latihan ChatGPT, untuk menjadikan bahasa Melayu diwakili sama seperti bahasa Indonesia dan bahasa rasmi negara lain. Situasi pada masa ini seperti yang ditunjukkan pada Jadual 1 adalah terlalu bersifat Eurosentrik. Tidak jelas berdasarkan kriteria apa (selain daripada bahasa Eropah) saiz

¹⁰ <https://openai.com/customer-stories/government-of-iceland>, diakses pada 23 Mei 2023

¹¹ <https://www.toyota.com.my/en.html>, diakses pada 23 Oktober 2023

¹² <https://www.perodua.com.my/>, diakses pada 23 Oktober 2023

¹³ <https://www.mydin.my/>, diakses pada 29 Disember 2023

¹⁴ <https://www.mymesra.com.my/bm/>, diakses pada 23 Oktober 2023

¹⁵ <https://proton.my/>, diakses pada 23 Oktober 2023

¹⁶ <https://ahhuat.com/en/home2/>, diakses pada 29 Disember 2023

¹⁷ <https://www.singsingcoffee.com/>, diakses pada 29 Disember 2023

¹⁸ <https://www.julies.com.my/>, diakses pada 29 Disember 2023

data pelbagai bahasa ditentukan. Tambahan pula, tidak ada satu pun laman web berdomain .my yang disenaraikan dalam 500 laman web teratas yang berdaftar di Common Crawl.¹⁹ Untuk makluman, laman web berdomain .sg dan .id dapat ditemui dalam senarai tersebut. Jadual 3 menunjukkan tiga penunjuk sosioekonomi yang berkemungkinan relevan dengan pelbagai saiz data latihan ChatGPT, iaitu populasi, KDNK dan KDNK per kapita. Walau bagaimanapun, tiada satu pun yang menerangkan perbezaan saiz data sebenar.

Jadual 3

*Statistik Sosioekonomi Negara Yang Menggunakan Bahasa dalam Jadual 1 pada 2022. Unit bagi KDNK dan KDNK per kapita ialah (malar 2015) USD.*²⁰

Negara	Populasi	KDNK (bilion)	KDNK per kapita
Belanda	17,703,090	884	49,979
Jepun	125,124,989	4,508	36,032
Indonesia	275,501,339	1,122	4,073
Malaysia	33,938,221	385	11,372

Tanggungjawab mempromosi bahasa Melayu dan pengurusan data tersebut terletak pada negara Malaysia sendiri. Hal ini berbeza daripada dua bahasa utama yang dituturkan di Malaysia seperti bahasa Inggeris yang menjadi tumpuan seluruh dunia dan bahasa Cina yang sumber bahasanya dipelihara oleh negara Taiwan dan China.

KESIMPULAN

Makalah ini melaporkan masalah penggunaan ChatGPT dalam bahasa Melayu dan Indonesia, iaitu ChatGPT sering bertindak balas terhadap prom Melayu dalam bahasa Indonesia. Masalah ini berlaku sebahagiannya disebabkan oleh ralat LangID, tetapi punca utamanya ialah layanan yang tidak sama rata terhadap kedua-dua bahasa. Secara khusus, bahasa Melayu diperlakukan seolah-olah ia adalah laras bahasa Indonesia yang tidak standard. Masalah ini bukan sahaja masalah teknologi tetapi juga mempunyai kesan sosial yang negatif, yang boleh dikurangkan dari segi teknologi dan sosiopolitik.

¹⁹ <https://commoncrawl.github.io/cc-crawl-statistics/plots/domains>, diakses pada 12 Julai 2023

²⁰ <https://data.worldbank.org>, diakses pada 16 September 2023

Kajian ini dengan itu menyumbang kepada perdebatan berterusan mengenai pembangunan AI yang bertanggungjawab. Walaupun makalah ini memfokuskan pada bahasa Melayu dan Indonesia, isu dan penyelesaian yang dibincangkan di sini boleh diterapkan kepada set bahasa lain yang serupa seperti bahasa Bosnia, Croatia dan Serbia serta bahasa Portugis Brazil dan Eropah.

Kajian ini mempunyai sekurang-kurangnya dua batasan. Pertama, pengkaji memeriksa keupayaan LangID ChatGPT dengan memintanya mengenal pasti bahasa menerusi ayat-ayatnya. Walau bagaimanapun, kaedah ini tidak menyasarkan bahasa yang dikenal pasti oleh ChatGPT secara langsung tetapi menekanya secara tidak langsung berdasarkan andaian bahawa ia akan ditunjukkan dalam respons. Andaian ini berkemungkinan salah. Kedua, pengkaji menggunakan versi percuma ChatGPT (GPT-3.5). Beberapa isu yang dibincangkan dalam kajian ini mungkin tidak muncul dalam versi berbayar (GPT-4).

Sebagai penutup, pengkaji ingin menegaskan bahawa pada masa ini jumlah data bahasa Melayu dalam ChatGPT adalah tidak mencukupi. Sekiranya data bahasa Melayu jauh lebih banyak dan meliputi semua domain kehidupan manusia, termasuk perniagaan, sains dan teknologi, maka ChatGPT pun tidak mungkin salah menganggap bahasa Melayu sebagai kelainan bahasa Indonesia. Layanan ChatGPT terhadap bahasa Melayu menjadi sedemikian boleh dikatakan berpunca daripada sikap sesetengah penutur bahasa Melayu di Malaysia yang menganggapnya tidak penting sehinggalah banyak laman web hanya menggunakan bahasa Inggeris sebagai mediumnya. Bahasa Inggeris dilihat telah mempelopori segenap pemikiran masyarakat sebagai bahasa utama, terutamanya di dalam bidang ekonomi. Tanggapan sedemikian menyebabkan bahasa Melayu tidak lagi berada di tempat yang sepatutnya sebagai bahasa rasmi negara. Penggunaan bahasa Inggeris dalam bidang tertentu mengakibatkan bahasa Melayu tidak lagi dapat berkembang dalam bidang tersebut dan AI berbahasa Melayu terbatas dari segi domain. Nilainya akan terus turun dan penutur bahasa Melayu akan mengalami kerugian dalam bidang teknologi. Maka, daya saing negara pun turut turun.

Justeru itu, rakyat perlulah merenung kembali dan beralih kepada syarikat yang menyumbang kepada manfaat penutur bahasa Melayu. Selain itu, “bahasa dan teknologi” memainkan peranan penting dalam bidang pendidikan bagi melahirkan sumber manusia yang mahir dalam

teknologi bahasa, yang semakin ramai di negara lain, termasuklah Jepun, Korea, Indonesia dan China. Zaman sedang beralih daripada “bahasa dan sastra” kepada “bahasa dan teknologi” dengan pesat seiring dengan perkembangan AI generatif.

PENGHARGAAN

Penulis ingin merakamkan penghargaan terima kasih kepada mereka yang terlibat secara langsung atau tidak langsung dalam penghasilan dan proses penerbitan makalah ini.

PENDANAAN

Kajian yang dilaporkan dalam makalah ini disokong oleh geran JSPS KAKENHI bernombor JP23H00639.

SUMBANGAN PENGARANG

Kesemua bahagian dalam penulisan ini dihasilkan secara bersama oleh kesemua penulis makalah.

PERISYTIHARAN KONFLIK KEPENTINGAN

Penulis mengisytiharkan bahawa tiada konflik kepentingan bagi penyelidikan ini dan tiada potensi konflik kepentingan berkaitan dengan penyelidikan, penulisan, atau penerbitan artikel ini.

PERNYATAAN KETERSEDIAAN DATA

Data yang menyokong kajian tersedia dalam makalah ini.

RUJUKAN

- Asmah Omar. (2001). The Malay language in Malaysia and Indonesia: From lingua franca to national language. *The Aseanists ASIA*, 2.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, C., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. Dlm. Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., & Lin, H. (Eds.) *Advances in Neural Information Processing Systems*, 33, 1877–1901. <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>
- Caswell, I., Breiner, T., Van Esch, D., & Bapna, A. (2020). Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. Dlm. Scott, D., Bel, N., & Zong, C. (Eeds.) *Proceedings of the 28th International Conference on Computational Linguistics*, 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics. <https://doi.org/10.18653/v1/2020.coling-main.579>
- Nomoto, H. (2011). Konstruksi *kena* dalam bahasa Indonesia: Perbandingan dengan bahasa Melayu. *Linguistik Indonesia*, 29(2), 111–131. https://www.tufs.ac.jp/ts/personal/nomoto/kena_Indonesia.pdf
- Nomoto, H. (2023). Issues surrounding the use of ChatGPT in similar languages: The case of Malay and Indonesian. Dlm. Park, J. C., Arase, Y., Hu, B., Lu, W., Wijaya, D., Purwarianti, A., & Krisnadhi, A. A. (Eeds.) *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, 76–82. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.ijcnlp-short.9>
- Nomoto, H., & Kartini Abd. Wahab. (2011). Konstruksi *kena* dalam bahasa Indonesia: Perbandingan dengan bahasa Melayu. *Linguistik Indonesia*, 29(2), 111–131.
- Nomoto, H., Akasegawa, S., & Shiohara, A. (2018). Reclassification of the Leipzig corpora collection for Malay and Indonesian. *NUSA*, 65, 47–66. <https://doi.org/10.15026/92899>

- Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3, 121–154. <https://doi.org/10.1016/j.iotcps.2023.04.003>
- Riza, H., Purwoadi, M., Gunarso, Uliniansyah, T., Ti, A. A., Aljuneid, S. M., Mai, L. C., Thang, V. T., Thai, N. P., Chea, V., Sun, R., Sam, S., Seng, S., Soe, K. M., Nwet, K. T., Utiyama, M., & Ding, C. (2016). Introduction of the Asian language treebank. Dlm. *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*. <https://doi.org/10.1109/ICSDA.2016.7918974>

LAMPIRAN

Versi Jadual 2 Yang Lebih Terperinci

Jadual 2 melaporkan dapatan keseluruhan ujian LangID, di mana tiga komponen data ujian digabungkan. Jadual 4–6 melaporkan dapatan bagi ketiga-tiga komponen secara berasingan. Jadual 4 menunjukkan bahawa keupayaan LangID ChatGPT sebenarnya tidak terlalu buruk untuk berita, yang mana menunjukkan bahawa ia dilatih terutamanya pada data berita. Prestasi ChatGPT jauh lebih rendah daripada penutur asli manusia dan Terjemahan Google untuk wiki (Jadual 5) dan fiksyen (Jadual 6). Dalam Jadual 6, markah biasanya lebih rendah daripada yang terdapat dalam jadual lain. Oleh itu, boleh dikatakan bahawa tugas LangID adalah lebih sukar dalam genre fiksyen berbanding dalam genre kewartawanan. Hakikat bahawa trend yang sama juga diperhatikan dengan penutur asli manusia bermakna perbezaan yang wujud antara bahasa Melayu dan Indonesia lebih sedikit dalam bahasa cerita fiksyen.

Jadual 4

Keupayaan LangID ChatGPT, Penutur Asli (Manusia) dan Terjemahan Google: Berita

		Bahasa Melayu	Bahasa Indonesia	Ketepatan (precision)	Perolehan (recall)	F1
ChatGPT	Prom 1	118	80	0.84	0.99	0.91
ChatGPT	Prom 2	112	85	0.84	0.94	0.89
ChatGPT	Prom 3	94	106	0.96	0.9	0.93
ChatGPT	Prom 4	109	91	0.86	0.94	0.90
Manusia	Moeljadi	98	102	0.98	0.96	0.97
Manusia	Farhan	106	94	0.92	0.98	0.95
Terjemahan Google		101	99	0.97	0.98	0.98

Jadual 5

Keupayaan LangID ChatGPT, Penutur Asli (Manusia) dan Terjemahan Google: Wiki

		Bahasa Melayu	Bahasa Indonesia	Ketepatan (precision)	Perolehan (recall)	F1
ChatGPT	Prom 1	145	47	0.62	0.90	0.73
ChatGPT	Prom 2	99	86	0.71	0.70	0.70
ChatGPT	Prom 3	57	139	0.81	0.46	0.67
ChatGPT	Prom 4	116	82	0.70	0.81	0.75
Manusia	Moeljadi	100	100	0.95	0.95	0.95
Manusia	Farhan	93	107	0.97	0.90	0.93
Terjemahan Google		100	100	0.97	0.97	0.97

Jadual 6

Keupayaan LangID ChatGPT, Penutur Asli (Manusia) dan Terjemahan Google: Fiksiyen

		Bahasa Melayu	Bahasa Indonesia	Ketepatan (precision)	Perolehan (recall)	F1
ChatGPT	Prom 1	151	49	0.66	0.90	0.72
ChatGPT	Prom 2	74	125	0.68	0.50	0.57
ChatGPT	Prom 3	64	136	0.86	0.55	0.67
ChatGPT	Prom 4	107	93	0.72	0.77	0.74
Manusia	Moeljadi	71	129	0.96	0.68	0.80
Manusia	Farhan	114	86	0.82	0.93	0.87
Terjemahan Google		89	110	0.91	0.81	0.86