PENANDA GOLONGAN KATA BAHASA MELAYU UNTUK TEKS MEDIA SOSIAL

Siti Noor Allia Binti Noor Ariffin
Sabrina Tiun

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Penanda Golongan Kata (GK) adalah langkah penting yang perlu di ambil kira di dalam algoritma Komputeran Bahasa Tabii (NLP). Antara cabaran pemprosesan Bahasa Tabii adalah untuk memproses makna perkataan di dalam teks media sosial seperti *tweets*. Hal ini dikatakan demikian kerana, penulisan *tweets* adalah lebih pendek iaitu hanya 140 patah perkataan sahaja yang dibenarkan, penulisan *tweets* bersifat bebas iaitu tidak mengekalkan tatabahasa formal, ejaan yang betul dan singkatan perkataan sering digunakan untuk mengatasi had terpanjang yang dikenakan. *Tweets* Bahasa Melayu tidak terkecuali menunjukkan fenomena linguistik berbeza seperti penggunaan dialek, pinjaman kata bahasa asing dalam konteks Bahasa Melayu, penggunaan campuran bahasa, penggunaan singkatan Bahasa Melayu dan kesilapan ejaan di dalam struktur ayat. Dengan bercirikan gaya tersendiri, bunyi dan kesilapan linguistik, teks media sosial *Twitter*, menjadi sukar untuk penandaan penanda GK. Kebanyakkan GK Bahasa Melayu yang sedia ada, hanya tertumpu kepada Bahasa Melayu yang formal dan tidak sesuai untuk digunakan dalam teks *tweets*. Oleh kerana itu, perlunya dibangunkan sebuah sistem GK yang khusus untuk Bahasa Melayu tidak formal seperti *tweets*. Tujuan projek ini dibangunkan adalah untuk mereka bentuk dan melaksanakan model penandaan Bahasa Melayu tidak formal *tweets*, dengan membuat penilaian menggunakan data ujian teks twitter tidak bertanda GK. Cadangan penyelesaian projek ini berdasarkan kepada pendekatan korpus dan pembelajaran mesin berselia QTAG. Dalam projek ini, berdasarkan pendekatan yang dinyatakan sebelum ini, kerja-kerja yang perlu dilakukan ialah menyediakan korpus latihan dengan mengumpul teks media sosial Bahasa Melayu *tweets*, menjalankan pra-pemprosesan ke atas korpus, melatih korpus dengan menggunakan pembelajaran mesin berselia QTAG dan penilaian dijalankan dengan menggunakan data ujian teks *Twitter* tidak bertanda GK. Hasil daripada penilaian yang telah dijalankan mendapati kedua-dua set korpus iaitu set korpus mentah dan set korpus normalisasi berjaya mendapat tahap purata ketepatan yang tinggi melebih 85%. Kesimpulannya, set penanda GK Bahasa Melayu tidak rasmi yang dibangunkan ini berjaya mencapai objektif kajian dan set korpus yang telah dinormalisasi adalah lebih sesuai digunakan sebagai set korpus ujian berbanding set korpus mentah *tweets*.

1 PENGENALAN

Golongan Kata (GK) adalah langkah penting yang perlu di ambil kira di dalam algoritma Komputeran Bahasa Tabii (NLP). Antara cabaran pemprosesan Bahasa Tabii adalah untuk memproses makna perkataan di dalam teks media sosial seperti *tweets*. Hal ini dikatakan demikian kerana, penulisan *tweets* adalah lebih pendek iaitu hanya 140 patah perkataan sahaja dibenarkan, penulisan tweets lebih bebas iaitu tidak mengekalkan tatabahasa formal, ejaan yang betul dan singkatan perkataan sering digunakan untuk mengatasi had terpanjang yang dikenakan (Java et al., 2007). GK adalah sebahagian daripada perisian yang membaca teks

dalam beberapa bahasa dan memberikan penanda GK kepada setiap perkataan (dan token lain), seperti kata nama, kata kerja, kata sifat, dan sebagainya, walaupun secara umumnya aplikasi komputasi menggunakan GK lebih halus seperti ‘kata-majmuk’.

Penandaan GK pada asalnya dibangunkan oleh Kristina Toutanova untuk penanda GK Bahasa Inggeris serta telah dibuat penambahbaikan dengan meningkatkan kelajuan, prestasi, kebolehgunaan, dan sokongan untuk bahasa-bahasa lain (Toutanova, et., al., 2003). Resolusi ketaksaan leksikal merupakan tugas utama dalam pemprosesan bahasa tabii (Baayen & Sproat, 1996). Ketaksaan leksikal boleh dianggap sebagai masalah klasifikasi. Hal ini adalah demikian kerana, ketaksaan leksikal ini boleh diklasifikasikan secara berasingan dan konteks yang diberikan akan ditentukan kelas mereka oleh penyahkodan atau pengelas. Ciri-ciri yang berkaitan dengan tugas pengelasan dikodkan ke dalam penanda. Kajian ini memerlukan set penanda GK dan korpus latihan yang bersesuaian untuk mendapatkan tahap purata ketepatan penandaan yang tinggi dalam statistik pengelasan golongan kata. Di dalam Elworthy, (1995), telah dibincangkan bahawa saiz set penanda GK memberi kesan kepada prestasi penandaan. Korpus latihan yang bersesuaian bermaksud korpus tersebut mempunyai jumlah patah perkataan di antara 100,000 sehingga lebih daripada satu juta perkataan di dalam satu korpus yang sama.

Walaupun sesetengah penandaan telah diprogramkan untuk mempelajari model bahasa daripada teks mentah (tanpa anotasi), penanda tersebut masih memerlukan pengesahan selepas output dikeluarkan dan prosedur *bootstrapping* akan dijalankan untuk memastikan penandaan yang dilakukan mencapai tahap kadar ralat yang minima. Semakin besar set penanda, semakin besar saiz korpus latihan yang diperlukan (Berger, Pietra & Pietra, 1996), dengan syarat bahawa korpus latihan yang cukup tersedia ini, tidak boleh terlalu bermasalah kerana masa tindak balas yang panjang akan memberikan kesan yang serius.

Penandaan GK Bahasa Melayu untuk teks media sosial perlu dibangunkan kerana terdapat permintaan yang tinggi untuk menyokong pemprosesan teks media sosial Bahasa Melayu tidak formal.

2 PENYATAAN MASALAH

Kebanyakkan penandaan GK dan algoritma penyahtaksaan adalah sama ada berasaskan peraturan, stokastik atau pembelajaran mesin. Teks sosial media, seperti *tweets*, *messenger*,

dan komentar di dalam Facebook merupakan cabaran di dalam bidang NLP. Tidak seperti bahasa formal (English, German, Arabic and Italy), yang mana instrumen NLP konvensional telah dibangunkan, teks perbualan mengandungi banyak item leksikal yang tidak formal dan berbentuk sintaktik. Ini adalah disebabkan hasil daripada kesilapan yang tidak disengajakan, variasi dialek, perbualan yang meninggalkan atau membuang perkataan, kepelbagaian topik, dan penggunaan bahasa yang kreatif dan ortografi (Eisenstein, 2013).

Kajian ini mengambil kira masalah penandaan GK untuk *tweets* bahasa Melayu tidak formal iaitu teks media sosial yang diberi fokus di dalam kajian ini. Teks daripada akaun *Twitter* adalah sukar untuk ditanda GK, kerana sifatnya yang hingar, terdapat kesilapan linguistik dan gaya idiosinkratik. Di samping itu, secara linguistiknya kebanyakan *tweets* telah terbentuk dengan baik, namun begitu perkatan di dalam teks *tweets* adalah tidak mengikut tatabahasa yang sebenar.

Seterusnya, kebanyakan penanda GK yang telah dibangunkan adalah khusus untuk teks berbentuk bahasa formal dan penanda GK tersebut dilatih menggunakan set korpus latihan yang besar iaitu melebihi 100,000 perkataan di dalam korpus. Bagaimanapun, kajian sebelum ini mendapati, jika teks media sosial iaitu teks yang menggunakan bahasa tidak formal, diimplementasikan dengan penandaan GK seperti ini akan menyebabkan penurunan prestasi penandaan secara ketara (Derczynski et al., 2013; Gimpel et al., 2011; Owoputi et al., 2012; Lynn et al.,2015).

Hal ini telah mecipta suatu cabaran baru untuk penyelidik memulakan kajian NLP bahasa Melayu di dalam teks media sosial iaitu *tweets,* kerana *tweets* bahasa Melayu ditulis di dalam beberapa bahasa dialek di Malaysia, penggunaan linguistik bebas dan *tweets* menunjukkan variasi signifikasi yang jelas.

3 OBJEKTIF KAJIAN

Objektif utama kajian ini adalah untuk mereka bentuk dan membangunkan sebuah model penanda GK Bahasa Melayu tidak formal untuk teks media sosial menggunakan QTAG yang berselia. Untuk memenuhi objektif kajian yang telah dinyatakan sebelum ini, beberapa kerja-kerja kajian telah dikenal pasti seperti pembinaan model penanda GK, penilaian model penanda GK dan antara muka pengguna ringkas.

4 METOD KAJIAN

Kajian ini membangunkan model penanda GK mengikut gambaran rekabentuk metodologi dan perlaksanaan model penandaan GK untuk bahasa Melayu berdasarkan kajian terhadap model pembelajaran mesin (ML) iaitu model kebarangkalian kebebasan penandaan bahasa (QTAG). Dalam konteks ini, kaedah pembelajaran mesin telah digunakan untuk membina dan membangunkan model ML untuk penandaan GK bahasa Melayu yang memerlukan beberapa langkah, termasuk perancangan dan penyusun sumber bahasa, permodelan QTAG, dan penilaian model QTAG.

Kajian ini juga membincangkan struktur sistem pengelasan ML yang dicadangkan. Fasa pembinaan model merupakan fasa pengumpulan korpus *tweets* Bahasa Melayu bertanda GK dan fasa penilaian model adalah fasa menguji purata ketepatan model dengan menggunakan beberapa jenis set korpus ujian berlainan saiz.

4.1 Fasa Perancangan

Fasa ini melibatkan proses pengenalpastian masalah, objektif, persoalan kajian dan menentukan skop kajian. Langkah seterusnya adalah sorotan susastera yang melibatkan pengumpulan, pencarian dan pembacaan jurnal serta kajian lepas bagi mencetus idea dan inspirasi. Contoh topik yang berkaitan dikaji terutama yang berkaitan dengan penandaan golongan kata untuk teks media sosial *Twitter* yang sedia ada. Penggunaan internet untuk mencapai maklumat dan bahan berkaitan di laman sesawang carian sumber artikel dan jurnal turut dilakukan. Maklumat tersebut dikumpul, distruktur dan disentesis serta dipersembahkan secara kritis dan kreatif di dalam fasa analisis.

4.2 Fasa Analisis

Fasa ini melibatkan analisis dan tafsiran maklumat yang dikumpul di dalam fasa perancangan. Analisa tentang kesesuaian topik dan menilai kepentingan untuk menjalankan kajian ini

dilakukan. Selain daripada itu, analisis terhadap set data korpus dan perisian yang sedia ada adalah sesuai untuk membangunkan projek ini.

4.3 Fasa Reka Bentuk

Kajian ini membangunkan model ML untuk penandaan GK *tweets* bahasa Melayu. Di dalam bahagian ini, gambaran seni bina bahagian penandaan GK dan penyahtaksaan GK sistem *tweets* Bahasa Melayu dibentangkan dan fungsi setiap komponen dalam penandaan GK diterangkan.

Pertama sekali, sebelum membina atau membangunkan model penanda GK QTAG ini, perlulah menyediakan satu set korpus latihan. Set korpus latihan ini merupakan set korpus yang telah ditoken dan ditanda GK secara manual oleh penyelidik. Data untuk set korpus latihan ini diambil daripada laman *Twitter* Bahasa Melayu. Hanya *tweets* bahasa Melayu sahaja yang dipilih sebagai korpus latihan. Set korpus latihan ini menjalani proses normalisasi kerana fokus utama kajian ini adalah untuk membangunkan set penanda GK yang mampu memberi penandaan GK kepada bahasa Melayu tidak formal media sosial *Twitter*. Penanda GK yang digunakan adalah set penanda GK yang telah dipilih dan ditambah komponen daripada kamus dwibahasa Melayu - Bahasa Inggeris Oxford Fajar (Edisi Kelima – 2014).

Kajian diteruskan di dalam fasa pembinaan model penanda GK. Tugas fasa pembinaan model ini adalah direka untuk menerima data yang tidak sempurna, hingar dan data sporadik. Model dibangunkan dengan mengumpul data teks media sosial *tweets* yang sudah ditanda GK. Data teks tersebut dipanggil sebagai korpus latihan. Korpus latihan kemudiannya di tokenisasikan dan di anotasi untuk di proses bersama-sama dengan QTAG.jar untuk menghasilkan dua fail berbeza iaitu fail lexicon dan fail matriks.

Seterusnya, kajian ini bersambung di dalam fasa penilaian model penanda GK. Tugas bagi fasa penilaian model ini untuk menginput set korpus *tweets* baharu iaitu set korpus ujian, ke dalam model QTAG dan output penandaan *tweets* Bahasa Melayu tidak formal ini akan dinilai dengan menghitung purata ketepatan penandaan GK di dalam set korpus ujian tersebut. Fasa ini dijalankan bertujuan untuk mendapatkan ketepatan penandaan GK model QTAG yang dibangunkan. Hasil daripada penilaian akan dianalisa purata ketepatan penandaan GK Seni bina keseluruhan model ML ini melibatkan fasa-fasa berikut:

4.4 Fasa Pengujian

Penilaian model QTAG merupakan fasa penilaian yang dirancang untuk menjalankan hanya satu jenis analisis percubaan sahaja iaitu menilai prestasi pendekatan penandaan untuk teks Melayu tidak formal *tweets*.

Set korpus latihan merupakan koleksi korpus *tweets* Bahasa Melayu mentah iaitu korpus yang tidak bertanda GK. Set ini dipastikan mengandungi sekurang-kurangnya 80 peratus teks Bahasa Melayu tidak formal daripada *tweets* kerana bertepatan dengan tujuan asal penandaan ini dijalankan iaitu untuk teks *tweets* Bahasa Melayu tidak formal, set korpus ujian yang telah dikumpul daripada akaun *Twitter* Bahasa Melayu mengandungi peratusan teks yang sama. Set data ini dihimpun bagi bertujuan untuk digunakan sebagai input baru yang akan menilai ketepatan penandaan GK model yang telah dibangunkan. Rajah 1 menunjukkan contoh set korpus *tweets* Bahasa Melayu mentah.

Proses penilaian model dilaksanakan dengan menjalankan pra-pemprosesan ke atas set korpus. Pra-pemprosesan bermula dengan proses normalisasi teks di dalam korpus. Normalisasi teks ini merupakan antara langkah terpenting yang perlu dilaksanakan sebelum penilaian dijalankan. Korpus yang diambil secara terus daripada twitter akan menjalani normalisasi dengan membuang semua tanda bacaan atau simbol yang berkemungkinan akan mengganggu atau mengubah makna penanda GK yang akan ditanda semasa proses anotasi dijalankan. Rajah 2 menunjukkan contoh set korpus *tweets* Bahasa Melayu yang telah menjalani proses normalisasi.

Proses kedua di dalam pra-pemprosesan ialah tokenisasi. Tokenisasi ini dijalankan secara automatik oleh model QTAG. Teks ditoken kepada serpihan-serpihan perkataan untuk memudahkan proses penandaan GK. Selepas proses tokenisasi selesai dijalankan, perkataan daripada korpus *tweets* Bahasa Melayu tersebut akan ditanda GK menggunakan model penanda GK QTAG. Model ini merupakan set penanda GK Bahasa Melayu yang telah dibangunkan di dalam fasa pembinaan model dan set penanda ini akan digunakan untuk keseluruhan penandaan yang akan dijalankan di dalam kajian ini. Rajah 3 dan Rajah 4 menunjukkan contoh penulisan kod bagi model penanda GK QTAG.

Setelah selesai proses tokenisasi, terhasil output korpus *tweets* Bahasa Melayu yang telah bertanda GK Bahasa Melayu. Output ini kemudian dinilai purata ketepatan penandaan GK

Bahasa Melayu dengan menggunakan kaedah yang dicadangkan oleh Tufis dan Mason, (1998). Output ini juga dijangka memiliki tahap ketepatan penandaan GK yang tinggi. Rajah 5 menunjukkan contoh output korpus *tweets* yang telah bertanda GK Bahasa Melayu.

```
Den dah Koba'an, jangan poei kek topi sungai tu. Kan dah Kono.
Sapo punyo goba ini, dah koto?
Sambal udang tu podeh bona.
Poei ambik buah kelapo tu, den nak masak gulai ni.
Cubolah jalan copek cikit, kot tetingga bas nanti.
bahan kuat sket..bior mampuih.
mak suka benor ngan si tipah tu..beralus budaknya
```

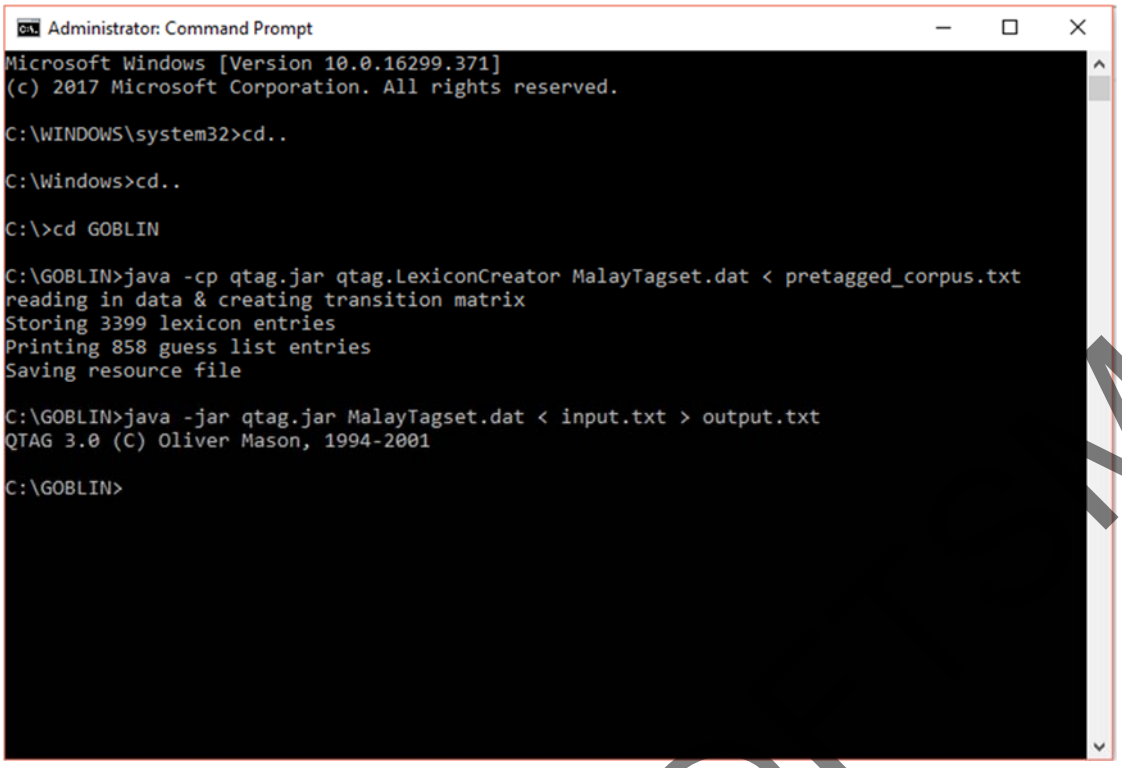
Rajah 1 Contoh set korpus *tweets* Bahasa Melayu mentah

```
Den dah Koba an jangan poei kek topi sungai tu Kan dah Kono
Sapo punyo goba ini dah koto
Sambal udang tu podeh bona
Poei ambik buah kelapo tu den nak masak gulai ni
Cubolah jalan copek cikit kot tetingga bas nanti
bahan kuat sket bior mampuih
mak suka benor ngan si tipah tu beralus budaknya
```

Rajah 2 Contoh set korpus *tweets* Bahasa Melayu yang telah dinormalisasikan

```
C:\WINDOWS\system32>cd..
C:\Windows>cd..
C:\>cd GOBLIN
C:\GOBLIN>java -cp qtag.jar qtag.LexiconCreator
MalayTagset.dat < pretagged_corpus.txt
C:\GOBLIN>java -jar qtag.jar MalayTagset.dat < input.txt
> output.txt
```

Rajah 3 Contoh penulisan kod proses model QTAG



Rajah 4 Contoh proses model QTAG menggunakan platform *Command Prompt* atau *Windows PowerShell*

```
<w pos=”KGDP”>Den</w>
<w pos=”ADJ”>dah</w>
<w pos=”KK”>Koba</w>
<w pos=”KKIK”>an</w>
<w pos=”KNF”>jangan</w>
<w pos=”KK”>poei</w>
<w pos=”KSN”>kek</w>
<w pos=”ADJ”>topi</w>
<w pos=”KN”>sungai</w>
<w pos=”KGNT”>tu</w>
<w pos=”KPB”>Kan</w>
<w pos=”ADJ”>dah</w>
<w pos=”KK”>Kono</w>
<w pos=”KT”>Sapo</w>
<w pos=”KK”>punyo</w>
<w pos=”KN”>goba</w>
<w pos=”KN”>ini</w>
<w pos=”ADJ”>dah</w>
<w pos=”ADJ”>koto</w>
<w pos=”KN”>Sambal</w>
<w pos=”KN”>udang</w>
<w pos=”KGNT”>tu</w>
<w pos=”ADJ”>podeh</w>
<w pos=”ADJ”>bona</w>
```

```
<w pos=”KK”>Poei</w>
<w pos=”KK”>ambik</w>
<w pos=”KNM”>buah</w>
<w pos=”KN”>kelapo</w>
<w pos=”KGNT”>tu</w>
<w pos=”KGDP”>den</w>
<w pos=”KK”>nak</w>
<w pos=”KK”>masak</w>
<w pos=”KN”>gulai</w>
<w pos=”KN”>ni</w>
<w pos=”#E”>Cubolah</w>
<w pos=”KN”>jalan</w>
<w pos=”ADJ”>copek</w>
<w pos=”ADJ”>cikit</w>
<w pos=”UNG”>kot</w>
<w pos=”KKIW”>tetingga</w>
<w pos=”KN”>bas</w>
<w pos=”KK”>nanti</w>
<w pos=”KN”>bahan</w>
<w pos=”ADJ”>kuat</w>
<w pos=”ADJ”>sket</w>
<w pos=”KH”>bior</w>
<w pos=”ADJ”>mampuih</w>
<w pos=”KN”>mak</w>
<w pos=”ADJ”>suka</w>
<w pos=”ADJ”>benor</w>
<w pos=”KSN”>ngan</w>
<w pos=”UNG”>si</w>
<w pos=”KN”>tipah</w>
<w pos=”KGNT”>tu</w>
<w pos=”KKIW”>beralus</w>
<w pos=”@KG”>budaknya</w>
```

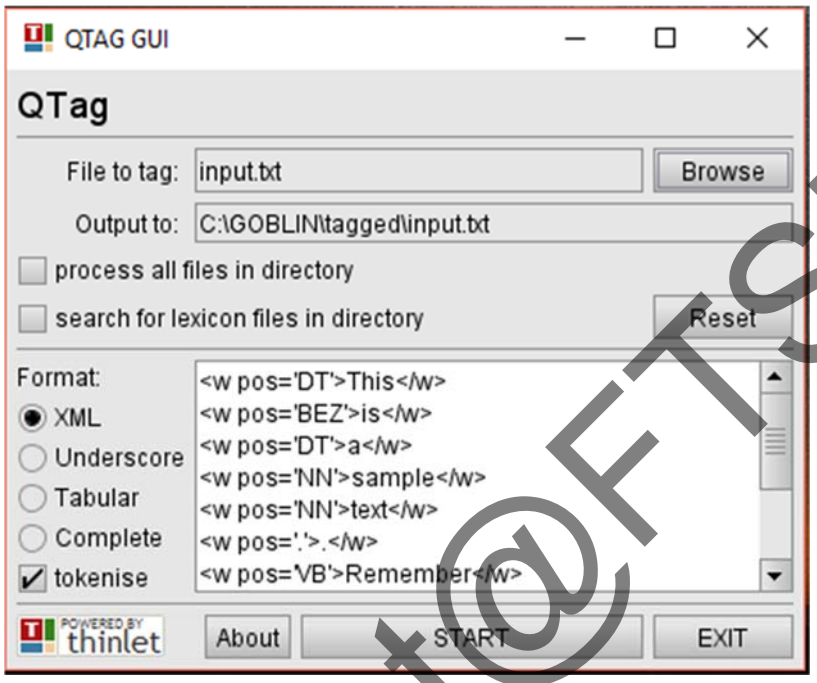
Rajah 5 Contoh output korpus bertanda GK Bahasa Melayu

3.6 ANTARA MUKA RINGKAS

Antara objektif kajian ini adalah untuk memastikan model penanda GK QTAG yang dibangunkan ini dapat diimplementasikan ke dalam aplikasi ringkas QTAG GUI yang dicipta oleh Tufis dan Mason, (1998).

Antara muka QTAG GUI ini khusus dibangunkan untuk penandaan GK menggunakan model QTAG. Antara muka ini lebih ringkas dan mudah dikendalikan berbanding kaedah manual iaitu menggunakan platform *Command Prompt* atau *Windows PowerShell.* Pengguna hanya perlu menyediakn set korpus teks media sosial *tweets* sebagai input data, proses penandaan GK

seperti pra-pemprosesan akan dilakukan secara automatik oleh model QTAG dan akhir sekali menghasilkan output korpus di dalam satu fail yang diberi nama *tagged* yang disimpan di dalam C.



Rajah 6 Contoh antara muka ringkas QTAG GUI

5 HASIL KAJIAN

Di dalam kajian ini, model penandaan GK QTAG telah digunakan untuk menguji set data yang mengandungi dua korpus berbeza iaitu korpus mentah dan korpus yang telah dinormalisasikan. Penilaian ini dijalankan mengikut kaedah penilaian yang dicadangkan oleh Tufis dan Mason, (1998). Penilaian ini dilaksanakan sebanyak 15 kali ke atas kedua-dua korpus tersebut bagi memastikan purata ketepatan penandaan GK bahasa Melayu tidak formal ini mendapat peratusan yang tinggi.

Set korpus ujian yang pertama merupakan set korpus teks media sosial *tweets* mentah. Set korpus mentah bermaksud set data ini diambil terus daripada *Twitter* Bahasa Melayu tanpa mengubah walau sedikitpun susunan tatabahasanya atau membuang simbol di dalam ayat. Set data ini menjadi set korpus ujian pertama kerana objektif model penandaan GK QTAG bahasa Melayu ini adalah untuk menanda GK perkataan di dalam bahasa Melayu tidak formal iaitu

bahasa Melayu yang mempunyai dialek, kesalahan ejaan & penambahan bahasa asing di dalam struktur ayat. Keputusan penilaian set data ujian pertama ini boleh dilihat di dalam Jadual 1.

Set korpus ujian yang kedua merupakan set korpus yang telah dinormalisasikan. Set korpus yang telah melalui proses normalisasi ini merupakan korpus mentah yang telah dibuang tanda baca atau simbol di dalam perkataan. Set data ini digunakan untuk menilai tahap ketepatan penandaan GK bahasa Melayu. Keputusan penilaian set korpus ujian kedua ini boleh dilihat di dalam Jadual 2. Ringkasan keseluruhan penilaian boleh dilihat di dalam Jadual 3.

Jadual 1 Hasil keputusan penilaian korpus ujian pertama.

Set Korpus Ujian Pertama (Korpus Mentah)	Jumlah Patah Perkataan	Jumlah Kesalahan Penandaan GK	Ketepatan
CWS_1	11	3	72.7 %
CWS_2	141	5	96.4 %
CWS_3	305	18	94.1 %
CWS_4	134	18	86.6 %
CWS_5	287	33	88.5 %
CWS_6	29	6	79.3 %
CWS_7	130	5	96.2 %
CWS_8	123	5	96.0 %
CWS_9	155	5	88.4 %
CWS_10	53	11	79.2 %
CWS_11	140	21	85.0 %
CWS_12	90	15	83.3 %
CWS_13	177	46	74.0 %
CWS_14	163	20	87.7 %
CWS_15	174	25	85.6 %

Jadual 2 Hasil keputusan penilaian korpus ujian kedua.

Set Korpus Ujian Kedua (Korpus Normalisasi)	Jumlah Patah Perkataan	Jumlah Kesalahan Penandaan GK	Ketepatan
COS_1	11	0	100.0 %
COS_2	124	5	96.0 %
COS_3	301	5	98.7 %
COS_4	133	4	97.0 %
COS_5	289	11	96.2 %
COS_6	29	1	96.6 %

COS_7	127	3	97.6 %
COS_8	124	3	97.6 %
COS_9	157	4	97.5 %
COS_10	56	2	96.4 %
COS_11	138	15	89.1 %
COS_12	88	13	85.2 %
COS_13	176	36	80.1 %
COS_14	164	2	98.8 %
COS_15	172	18	89.5 %

Jadual 3 Ringkasan keputusan penilaian kedua-dua set korpus.

Set Korpus Ujian	Jumlah Patah Perkataan	Jumlah Kesalahan Penandaan GK	Purata Ketepatan
Korpus mentah	2112	236	88.8 %
Korpus normalisasi	2089	122	94.6 %

Hasil daripada penilaian kedua-dua set korpus ujian ini iaitu set korpus mentah dan set korpus yang telah melalui proses normalisasi ini menunjukkan tahap purata ketepatan yang tinggi iaitu melebihi 85% (88.8 peratus bagi korpus mentah dan 94.6 peratus bagi korpus yang telah dinormalisasi). Set korpus ujian kedua iaitu set korpus normalisasi ini mendapat purata peratusan ketepatan yang tinggi kerana set ini diuji dengan membuang tanda baca dan simbol di dalam struktur ayat. Pembuangan tanda bacaan dan simbol ini dilaksanakan kerana *tweets* Bahasa Melayu yang ditulis oleh rakyat Malaysia, sering kali mempunyai tanda yang berlebihan. Hal ini kerana, penutur dan penulis bahasa Melayu sering menulis mengikut gaya percakapan mereka. Oleh kerana itu, proses penilaian ini menggunakan set korpus yang telah dinormalisasikan bagi mendapatkan tahap purata ketepatan penandaan GK yang tinggi.

6 KESIMPULAN

Disertasi ini dapat disimpulkan dengan pembangunan model penanda GK Bahasa Melayu ini sudah berjaya dibangunkan dengan set korpus latihan melebihi 5000 patah perkataan dan dengan purata ketepatan penandaan melebihi 90% ketepatan. Pembangunan model penandaan GK Bahasa Melayu ini tidak akan terhenti setakat ini kerana kajian yang lebih mendalam akan dilaksanakan di dalam kajian yang akan datang di mana saiz korpus latihan akan ditambah bagi

meningkatkan tahap purata ketepatan penandaan GK model QTAG Bahasa Melayu bagi Bahasa Melayu tidak formal teks media sosial *tweets.*

7 RUJUKAN

Abdullah, H., S. Rohani, L. J., Ayob, R. dan Osman, Z. 2006. Sintaksis siri pengajaran dan pembelajaran Bahasa Melayu. Kuala Lumpur: PTS Professional.

Albogamy, F., & Ramsay, A. (2015): POS Tagging for Arabic Tweets, School of Computer Science, University of Manchester, Manchester, M13 9PL, UK.

Al-Sabbagh, R., & Girju, R. (2012). A supervised POS tagger for written Arabic social networking corpora.

A Omar, Ensiklopedia Bahasa Melayu, KL, Malaysia: Dewan Bahasa dan Pustaka, 2008

Arbak, O. 2005. *Kamus komprehensif Bahasa Melayu*. Shah Alam: Oxford Fajar.

Baayen, H., Sproat, R. (1996): Estimating Lexical Priors for Low-Frequency Morphologically Ambiguous Forms in *Computational Linguistics*, vol. 22, no. 2 (pp. 155-166), June 1996.

Berger, A., L., Della Pietra, SA., Della Pietra, V., J. (1996): A Maximum Entropy Approach to Natural Language Processing in Computational Linguistics, vol. 22, no. 1 (pp. 39-72), March 1996.

Derczynski, L., Ritter, A., Clark, S., & Bontcheva, K. (2013). Twitter Part of Speech Tagging for All: Overcoming Sparse and Noisy Data. Paper presented at the RANLP.

Elworthy, D. (1995): Tagset Design and Inflected Languages, *Proceedings of the ACL SIGDAT Workshop*, Dublin, (also available as cmp-lg archive 9504002).

Feldman, A. (2006). Portable language technology: a resource-Light approach to morpho-Syntactic tagging. (Ph.D.), The Ohio State University.

Gonda, J. (1949). Prolegomena tot een theorie der woordsoorten in Indonesische talen. Bijdragen tot de Taal-, Land-en Volkenkunde, (2/3de Afl), 275-331.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., . . . Smith, N. A. (2011). Part of speech tagging for twitter: Annotation, features, and experiments. Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2.

Gui, T., Zhang, Q., Huang, H., Peng, M., & Huang, X. (2017). Part-of-Speech Tagging for Twitter with Adversarial Neural Networks. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2401-2410).

Hasan, A. (1974) The Morphology of Malay. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Hassan, A. (1987). Penerbitan kata dalam bahasa Malaysia. Penerbit Fajar Bakti.

Hawkins, J. M. (2008). Kamus dwibahasa Bahasa Inggeris–Bahasa Malaysia. *Selangor: Oxford Fajar*.

Hock, O. Y. 2009. *Kamus Dwibahasa*. Petaling Jaya: Pearson Longman.

Hussien, M.A., (2016) "Part of Speech Tagging Model for Arabic Tweet Based on Machine Learning", 2016, Master Dissertation, UKM, Malaysia.

Hussien, M.A., (2017): Comparative Analysis of Ml OSs on Arabic Tweets, Journal of Theoretical and Applied Information Technology 31st January 2017. Vol.95. No.2.

Java, A., Song, X., Finin, T., & Tseng, B. (2007, August). Why we twitter: understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (pp. 56-65). ACM.

Knowles, G. and Zuraida. M. D. 2006. *World Class in Malay: A Corpus-based Approach*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Lynn, T., Scannell, K., & Maguire, E. (2015). Minority Language Twitter: Part of Speech Tagging and Analysis of Irish Tweets. ACL-IJCNLP 2015, 1.

Mohamed, H., Omar, N., & A.A, M.J. (2015): Malay Part of Speech: A comparative Study on Tagging Tools", 2015, http://www.ftsm.ukm.my/apjitm , Asia-Pacific Journal of Information Technology and Multimedia, Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik, Vol. 4 No. 1, June 2015: 11 – 23, e-ISSN: 2289-2192.

Nik Safiah, K., Farid, M. O., Hashim, M. dan Abdul Hamid, M. 2010. *Tatabahasa dewan edisi ketiga*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

N.S. Karim, F.M. Onn, H. Musa and A.H. Mahmood, Tatabahasa Dewan - Edisi Baham, KL, Malaysia: Dewan Bahasa dan Pustaka, 2006

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters.

Palshikar, G. K., Apte, M., & Pandita, D. (2017). Weakly Supervised Classification of Tweets for Disaster Management. In SMERP@ ECIR (pp. 4-13).

Ranaivo-Malançon, B. 2008. Issues in building a Malay part of speech tag-set. *Proceeding of the 2nd International MALINDO Workshop*. Cyberjaya: Multimedia University, 104-108.

Samuel, D. (2017). On the use of vector representation for improved accuracy and currency of Twitter POS Tagging (Doctoral dissertation).

Schmid, H., & Laws, F. (2008). Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. Paper presented at the Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, UK.

Teeuw, A. (1962). Some problems in the study of word-classes in Bahasa Indonesia. Lingua, 11, 409-421.

Toutanova, K., Klein, D., Manning, C., & Singer, Y., (2003) Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.

Tufis, D., & Mason, O. (1998). Tagging romanian texts: a case study for qtag, a language independent probabilistic tagger. In Proceedings of the First International Conference on Language Resources and Evaluation (LREC) (Vol. 1, pp. 589-596).

Wahab, A. H. A. (1978). Beberapa Prinsip Dasar Untuk Tatabahasa Melayu (1) Penggolongan Kata dan Sintaksis" dlm. Jurnal Dewan Bahasa, 22(8), 2.

van der Goot, R., Plank, B., & Nissim, M. (2017). To normalize, or not to normalize: The impact of normalization on part-of-speech tagging. arXiv preprint arXiv:1707.05116.