NORMALISASI TEKS DI MEDIA SOSIAL DENGAN MENGGUNAKAN KAEDAH PERATURAN

Nur Shameen Aina Abdul Rahim
Nazlia Omar

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Sebahagian besar pengguna internet di Malaysia menggunakan media sosial untuk mendapatkan informasi secara rutin. Namun, frekuensi penggunaan media sosial yang tinggi, tidak setanding dengan penggunaan bahasa yang tidak standard yang digunakan dalam mengisi kandungan media sosial dengan maksud untuk memudahkan komunikasi. Ejaan bahasa yang digunakan tidak hanya mengganggu pengguna media sosial namun juga mempengaruhi pemprosesan terhadap data media sosial yang biasa disebut Pemprosesan Bahasa Tabii (PBT). Oleh yang demikian, pra pemprosesan bagi data dari media sosial seperti Twitter agak mencabar. Kajian ini mencadangkan proses menormalisasikan data Twitter bahasa Melayu yang hingar kepada bahasa piawai bahasa Melayu dengan menggunakan kaedah peraturan. Kajian ini akan melibatkan pemprosesan teks yang merangkumi proses pembersihan data diikuti dengan pengecaman perkataan di luar perbendaharaan kata (*out of vocubulary*) dan penggantian perkataan. Data yang hingar dan mempunyai unsur loghat kedah telah berjaya ditukarkan kepada bahasa Melayu piawai setelah melalui kesemua proses dalam kajian ini. Kajian ini menggunakan pendekatan baru untuk meningkatkan normalisasi teks bahasa Melayu. Dengan itu, penyelidik lain dapat menggunakan kajian ini sebagai rujukan untuk terus meningkatkan normalisasi teks bahasa Melayu.

1 PENGENALAN

Media sosial adalah sejenis media yang digunakan terus secara bertalian dengan membolehkan para pengguna dengan mudah untuk menyertai, berkongsi, dan mencipta isi kandungan tersendiri (Ahlqvist Toni, 2008). Media sosial juga didefinisikan sebagai sebuah kumpulan aplikasi berasaskan internet yang membangun atas dasar ideologi dan teknologi Web 2.0, dan yang membolehkan penciptaan dan pertukaran (Andreas Kaplan dan Michael Haenlein, 2010). Ia datang dalam pelbagai format termasuk blog, rangkaian sosial, wiki, forum dan dunia maya di mana ia membolehkan pemuatnaikan dan perkongsian karangan, gambar, video, audio dan grafik sesama rakan sepengunaan yang boleh memberi maklum balas secara terbuka dalam masa yang lebih cepat dan tidak terhad ruang berbanding media tradisional (termasuk percetakan dan penyiaran) (H. Kietzmann, Jan, 2011).

Twitter merupakan aplikasi dalam talian yang percuma yang mempunyai lebih 500 juta pengikut. Namun, Twitter memiliki keterbatasan karakter penulisan iaitu hanya 280 patah perkataan yang menyebabkan pengguna Twitter sering melakukan penyingkatan. Singkatan tersebut mengakibatkan kata menjadi tidak piawai (Wahyuningtyas 2016). Salah satu tujuan pengguna Twitter melakukan penyingkatan kata tersebut adalah untuk memanfaatkan ruang penulisan bagi pengguna yang ingin mengutarakan informasi lebih dari 280 karakter.

Dalam melakukan pemprosesan data teks yang tidak berstruktur ini, para peneliti menggunakan kaedah yang disebut dengan Pemprosesan Bahasa Tabii atau yang biasa disingkat PBT iaitu satu kaedah pembentukan model komputasi bahasa yang melakukan interaksi antara manusia dan komputer dengan perantaraan bahasa tabii. PBT berupaya untuk memahami bahasa tabii manusia dengan segala aturan gramatik dan semantiknya, serta mengubah bahasa tersebut menjadi repersentasi formal yang dapat diproses oleh komputer.

2 PENYATAAN MASALAH

Media sosial semakin berkembang dari semasa ke semasa mengikut perkembangan arus. Teks di media sosial mempunyai tatabahasa dan cara penggunaannya tersendiri yang menyukarkan teks tersebut untuk diproses menggunakan pemprosesan bahasa tabii (PBT) atau dalam bahasa inggerisnya dikenali sebagai natural language processing (NLP).

Walaubagaimanapun, cabaran utama dalam pemprosesan teks dari mikroblog ialah bentuk ayat yang digunakan oleh pengguna. Kebanyakan ayat dari mikroblog menggunakan teks gaya bebas, mengandungi banyak singkatan, kesalahan tipografi serta mengandungi emotikon. Ini adalah disebakan hasil daripada kesilapan yang tidak disengajakan, variasi dialek, perbualan yang meninggalkan atau membuang perkataan, kepelbagaian topik dan penggunaan bahsa yang kreatif dan ortografi (Eisenstein, 2013). Oleh itu, normalisasi teks perlu dilakukan supaya ayat yang tidak berstruktur boleh difahami oleh mesin.

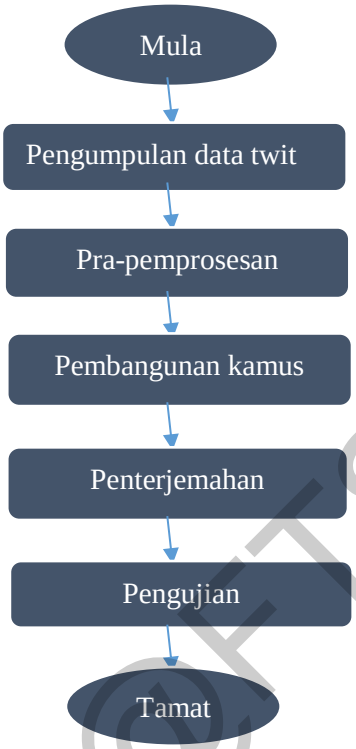
Berdasarkan analisis mesej Twitter dari pengguna Malaysia, beberapa masalah ditemui dan dikenal pasti. Contoh masalah ini adalah penggunaan bahasa Melayu tempatan yang memberi makna yang sama tetapi menggunakan ejaan yang berbeza dari kata- kata standard, penggunaan singkatan yang tidak tetap atau teks yang dihasilkan oleh pengguna. Munculnya bahasa trend bahasa Melayu tempatan yang mempunyai perkataan yang serupa tetapi berbeza makna dari perkataan asal yang terdapat di dalam kamus dan campuran bahasa yang kebanyakannya terdiri daripada bahasa Inggeris dan bahasa Melayu. Menurut Baldwin and Yunyao (2015), kebanyakan alat PBT digunakan dalam teks formal.

3 OBJEKTIF KAJIAN

Matlamat utama projek ini adalah untuk mengenal pasti dan menormalisasikan variasi leksikal perkataan dari data Twitter. Projek ini mengeksploit data dari mikroblog seperti Twitter yang penuh dengan data hingar (noisy data). Data yang diambil merangkumi bahasa Melayu dan mempunyai unsur dialek. Seterusnya, kajian ini juga akan membangunkan peraturan dan algoritma bagi menyelesaikan normalisasi teks seperti pembangunan kamus dialek dan kamus hingar. Untuk memenuhi objektif kajian yang telah dinyatakan, beberapa langkah pemprosesan teks telah dilakukan yang merangkumi proses pembersihan data diikuti dengan pengecaman perkataan LDKK dan penggantian perkataan.

4 METOD KAJIAN

Fasa pertama kajian ini akan dimulakan dengan mengumpul data teks yang terdapat di Twitter.Fasa kedua merupakan fasa pra-pemprosesan iaitu fasa yang akan mengeluarkan beberapa data hingar dalam data twit sekaligus melakukan proses pembersihan *'cleaning'*. Terdapat tujuh proses yang dilakukan dalam fasa kedua ini. Seterusnya, fasa ketiga merupakan fasa pembangunan kamus. Kesemua perkataan yang telah melalui fasa kedua akan diuji dengan Kamus dialek dan Kamus hingar yang telah dibangunkan. Pada fasa keempat data yang telah diproses akan melalui proses penterjemahan bahasa bagi menterjemahkan data kepada bahasa Melayu. Selepas itu, fasa terakhir iaitu fasa normalisasi akan diteruskan dengan proses mengenalpasti perkataan yang telah melalui kesemua fasa sebelumnya. Fasa ini juga akan mengeluarkan 'output' bagi data twit yang hingar pada permulaannya. Jangkaan hasil kajian ini akan meneutralkan perkataan yang tidak mencapai piawai kepada perkataan bahasa Melayu yang tepat. Rajah 4.1 menerangkan proses untuk menormalisasikan data teks kepada bahasa Melayu.



Rajah 4.1 Carta aliran normalisasi leksikal

Mikroblog ialah medium siaran dalam talian yang wujud sebagai bentuk blog tertentu. Blog mikro berbeza dengan blog tradisional kerana kandungannya biasanya lebih kecil di kedua-dua saiz fail sebenar dan agregat. Blog mikro membolehkan pengguna menukar unsur kecil kandungan seperti kalimat pendek, imej individu atau pautan video, yang mungkin menjadi sebab utama popularitinya (Aichner, T.; Jacob, F. (Mac 2015). Pada permulaan proses ini, data akan dikumpul daripada Twitter. Twitter mengandungi bilangan yang besar bagi mesej yang pendek yang diwujudkan oleh pengguna Twitter itu sendiri. Anggaran korpus Twitter yang akan dikumpul adalah 5000 mesej pendek. Berdasarkan korpus tersebut, ia mungkin mengandungi beberapa jenis data hingar utama seperti di bawah :

- "Perkataan asing"- Perkataan asing merujuk kepada bahasa selain bahasa Melayu, seperti bahasa Inggeris atau bahasa Arab. Perkataan asing sering digunakan oleh pengguna Twitter. Contohnya, perkataan 'so' yang berasal dari bahasa Inggeris. Perkataan ini akan terus diasingkan sebagai perkataan asing. Jika pengasingan tidak dibuat, ia akan menyukarkan proses normalisasi.

- "Bahasa Melayu tidak piawai" – Bahasa ini merujuk kepada bahasa Melayu yang masih boleh difahami oleh pengguna, tetapi perkataan ini tidak menepati piawai bahasa Melayu. Berikut merupakan contoh perkataan bahasa Melayu yang tidak menepati piawai:

➢ Kesalahan ejaan. Misalnya pemendekan perkataan *aku* dieja sebagai *aq.* Selain itu, pengulangan huruf dalam perkataan seperti perkataan *keeee* yang menggunakan pengulangan huruf 'e'.

➢ "Dialek" digunakan secara meluas di Twitter. Contoh penggunaan dialek Utara seperti perkataan *mai* yang berasal dari perkataan *datang.*

➢ "Huruf dan nombor" merujuk kepada perkataan yang mengandungi kedua huruf dan ayat. Sebagai contoh, perkataan *jalan-jalan* dipendekkan kepada *jalan2*.

➢ "Lain-lain" merujuk kepada penggunaan simbol seperti @, #, / dan sebagainya. Misalnya, #waktusolat

Data di atas merupakan contoh-contoh jenis data hingar. Proses normalisasi ini akan dimulakan dengan fasa pertama iaitu fasa pengumpulan data daripada Twitter. Melalui fasa ini, langkah-langkah dalam mendapatkan data tersebut akan dijelaskan.

FASA 1: PENGUMPULAN DATA

Data akan diekstrak daripada Twitter. Seterusnya, data tersebut akan dipisahkan kepada dua bahagian data percubaan (*training set*) dan data pengujian (*test set*). Data percubaan adalah data set yang diuji untuk mencapai objektif kajian. Data ini selalunya mempunya dua jenis input iaitu vektor dan skala. Data pengujian adalah data set yang memberi penilaian berdasarkan data percubaan. Selain itu, data tambahan juga dibenarkan untuk dimasukkan. Pengambilan data diambil daripada Twitter perlulah mengikuti beberapa langkah untuk memastikan data boleh didapati dengan mudah. Twit random yang dikumpul disediakan dalam kajiannya oleh Twitter API semasa tahun 2012 sehingga tahun 2018 (van der Goot, 2019). Berikut langkah untuk mendapatkan data daripada Twitter dengan menggunakan "Streaming Api".

a. Mendapatkan kekunci 'API' Twitter

Langkah pertama bermula dengan pengguna perlulah mengakses masuk ke dalam akaun Twitter terlebih dahulu. Seterusnya, pengguna harus membuat aplikasi baru dan mengisi borang, persetujuan dengan bersyarat. Pengguna perlu menyalin kunci 'API' dan rahsia 'API' dan mencipta akses kepada token.

b. Menghubungkan 'Twitter Streaming API' dan memuat turun data

Proses ini akan menghubungkan 'Twitter Streaming API' dan sekali gus memuatturun data. Langkah ini akan menggunakan perisian berasaskan bahasa pengaturcaraan 'Python'. API yang digunakan haruslah mempunyai kata kunci yang hendak dicari yang akan digunakan bagi proses normalisasi. Data yang akan diperuntukkan merupakan data yang berfokuskan tajuk kajian. Ini akan memudahkan proses normalisasi sekali gus mencapai objektif kajian.

c. Membaca dan memahami data

Data akan disimpan dalam bentuk fail pembacaan seperti *.csv*. Format yang digunakan dalam data ini mudah difahami oleh manusia. Berikut merupakan ayat contoh data yang diambil daripada Twitter.

Sebelum pra-pemprosesan:

" RT @jnmalaysia: Nayaaaa Pak Karimmmmmm.. "

Ayat di atas mempunyai gabungan jenis-jenis kata hingar dan juga terdapat unsur dialek Utara digunakan dalam ayat ini. Fasa seterusnya akan melakukan pra-pemprosesan bagi ayat di atas.

FASA 2: PRA-PEMPROSESAN

Melalui fasa ini, beberapa teknik pra-pemprosesan akan dilakukan bagi membersihkan (‘cleaning’) daripada perkataan yang akan menganggu proses normalisasi.

a. Tokenisasi

Fasa ini merupakan fasa pra-pemprosesan awal bagi kajian ini. Proses ini melibatkan tokenisasi iaitu langkah dalam mengasingkan perkataan satu persatu daripada ayat penuh sesuatu data. Tokenisasi juga boleh diklasifikasikan sebagai segmentasi ayat atau analisis leksikal. Output bagi proses tokenisasi adalah pecahan ayat kepada bentuk perkataan. Segmentasi merupakan asas yang penting dalam suatu pemprosesan bahasa tabii. Tujuan utama segmentasi adalah untuk mengenalpasti perkataan yang betul dalam sesuatu ayat. (Choochart Haruechaiyasak & Alisa Kongthon, 2013). Selain itu, fasa ini juga akan mengasingkan yang tiada makna seperti tanda baca, simbol, dan lain-lain. Berikut merupakan contoh fasa tokenisasi dalam ayat tersebut.

Sebelum tokenisasi:

" RT @jnmalaysia: Nayaaaa Pak Karimmmmmm.. "

Selepas tokenisasi:

RT	*@jnmalaysia*	*:*	*Nayaaaa*	*Pak*	*Karimmmmmm*	.	.

Jadual 4.1 Tokenisasi bagi contoh ayat di atas.

b. Pembuangan data hingar, URL, tanda pagar '#' dan sebutan pengguna '@'

Perkataan dan aksara yang tidak diperlukan diklasifikasikan sebagai data hingar. Selain itu, penggunaan 'URL' sebagai tambahan informasi dalam data twit juga akan dibuang. Twitter merupakan antara media sosial yang menggunakan tanda pagar bagi mengaitkan twit mereka dengan sesuatu perkara. Tanda pagar atau popular disebut sebagai *'hashtag'* merupakan sejenis tag metadata yang digunakan pada rangkaian sosial yang membolehkan pengguna memohon penandaan dinamik, yang dihasilkan oleh pengguna yang membolehkan orang lain mudah mencari mesej dengan tema atau kandungan teretentu. Seterusnya, sebutan pengguna ataupun *'username'* juga kerap digunakan dalam mikroblog Twitter. Sebutan pengguna digunakan untuk membalas twit dan juga mennyebut nama pengguna tersebut dalam Twitter.

Sebelum pra pemprosesan : *RT @jnmalaysia: Nayaaaa Pak Karimmmmmm..*

Selepas pra pemprosesan : *Nayaaaa Pak Karimmmmmm..*

c. Mengeluarkan emotikon dan emoji

Pengguna Twitter sering menggunakan emotikon dan emoji dalam twit mereka bagi mengeksperesikan perasaan mereka ataupun menggambarkan sesuatu. Sebagai contoh emotikon ":)" membawa ekspresi senyum dalam twit mereka. Contoh ayat dibawah merupakan

Sebelum pra pemprosesan : *Tak orang Utagha lah kalau tak cakap Utagha :D*
Selepas pra pemprosesan : *Tak orang Utagha lah kalau tak cakap Utagha*

d. Menggantikan aksara yang memanjang

Selain itu, pengguna Twitter juga sering mengeja perkataan dengan tambahan huruf secara berulang dalam satu perkataan. Contohnya, perkataan 'sedapnya' dieja sebagai 'sedaapnyaaa'. Oleh itu, amat penting proses untuk menggantikan perkataan yang mempunyai lebihan aksara dalam ejaan ini kepada ejaan asal.

Sebelum pra pemprosesan : *Nayaaaa Pak Karimmmmmm..*
Selepas pra pemprosesan : *Naya Pak Karim..*

e. Mengeluarkan tanda baca

Dalam data twit mengandungi tanda baca. Tanda baca merupakan simbol atau tanda yang melebihi satu dan digunakan untuk memberi isyarat kepada pembaca supaya melakukan sesuatu dalam bacaan. Ia diletakkan di tempat-tempat tertentu dalam ayat berdasarkan tujuan dan kesesuaiannya. Berikut merupakan pra-pemprosesan bagi mengeluarkan tanda baca . Selepas pra- pemprosesan dilakukan, tiada tanda baca yang dipaparkan dalam data twit tersebut. Walaupun begitu, pengeluaran tanda baca ini tidak akan mengubah maksud ayat.

Sebelum pra pemprosesan : *Nayaaaa Pak Karimmmmmm..*
Selepas pra pemprosesan : *Naya Pak Karim*

FASA 3: PEMBANGUNAN KAMUS

Dalam fasa ini pembangunan Kamus dialek akan dibangunkan bagi membantu proses normalisasi. Perkara utama yang dilakukan terlebih dahulu adalah mengumpul perkataan dialek. Selain itu, Kamus hingar juga akan dibangunkan dalam fasa ini. Kamus hingar mengandungi singkatan di mana corak teksnya tidak tetap atau teks dari pengguna yang ejaannya sedikit berbeza daripada ejaan standard tetapi mempunyai maksud yang sama. Selain itu, ia juga mengandungi singkatan perkataan yang mengandungi kata nama dan perkataan Inggeris standard yang telah dieja secara tidak rasmi. Kamus ini juga mengandungi perkataan yang telah dieja dengan angka atau hanya angka sahaja yang mempunyai makna. Jadual 4.2 menunjukkan senarai kata bagi Kamus dialek dan Jadual 4.3 menunjukkan senarai kata bagi Kamus hingar.

Jadual 4.2 Kamus dialek

Dialek	**Maksud**
habaq	beritahu
duk	duduk
mai	mari
geqam	geram
pelaq	kurang ajar
awat	kenapa
bantei	pukul
sat	sekejap
bengkin	garang
guano	bagaimana
berdangoh	berdarah
bibior	bibir
kitak	awak
teman	saya

Jadual 4.3 Kamus hingar

Perkataan	Maksud
dgn	dengan
ktorg	kita orang
x	tak
tau	tahu
tue	tu
nie	ni
yg	yang
astu	selepas itu
tgh2	tengah-tengah
pulak	pula
mkn	makan
2dung	tudung
slmt	selamat
se7	setuju

Berdasarkan jadual 4.2 dan jadual 4.3, perkataan tersebut akan ditukarkan kepada maksud sebenar perkataan tersebut dalam bahasa Melayu piawai.

FASA 4 : PENTERJEMAHAN BAHASA

Pengguna media sosial sering menggunakan dialek yang pelabagai dalam menyampaikan mesej di laman sosial seperti Twitter dan Facebook. Penggunaan bahasa asing yang popular dalam kalangan masyarakat di Malaysia adalah percampuran antara dua bahasa iaitu bahasa Melayu dan bahasa Inggeris. Penterjemahan bahasa Inggeris kepada bahasa Melayu perlulah dilakukan supaya output akhir bagi data merupakan bahasa Melayu. Penterjemahan secara langsung bagi menterjemahkan bahasa asing dalam data Twitter dengan menggunakan *googleTrans* akan dilakukan dalam fasa ini.

Sebelum penterjemahan : *Aku rasa amazing dengan hang.*

Selepas penterjemahan: *Aku rasa kagum dengan hang.*

FASA 5 : PENILAIAN

Pada fasa ini, penilaian output akan dilakukan melalui dua cara. Pertama, percubaan untuk menilai prestasi normalisasi teks bahasa Melayu pemerhatian hasil output. Kemudian, untuk menilai prestasi normalisasi teks bahasa Melayu menggunakan kaedah ketepatan dan dapatan semula (precision and recall).

5 HASIL KAJIAN

Hasil pengujian mendapati fasa pertama hingga fasa akhir bagi kajian ini hampir menepati objektif kajian iaitu menormalisasi data yang mengandungi perkataan hingar dan loghat Kedah kepada data bahasa Melayu. Jadual 5.1 merupakan keputusan pengujian ini.

Jadual 5.1 Hasil ujian bagi setiap fasa

Pengujian	**Penerangan**	**Berjaya/Gagal**	**Penambahbaikan**
Pengujian 1 : Pra-pemprosesan	Pengujian pembuangan simbol dan lain-lain	Berjaya	Menambah lagi pra – pemprosesan untuk menyingkirkan simbol yang lain.
Pengujian 2 : Pembangunan kamus loghat Kedah	Pengujian penukaran secara terus bagi perkataan dialek Kedah	Berjaya	Pengujian dapat menukarkan perkataan loghat Kedah kepada bahasa Melayu
Pengujian 3 : Pembangunan kamus hingar	Pengujian penukaran secara terus bagi perkataan hingar	Berjaya	Pengujian dapat menukarkan perkataan hingar kepada bahasa Melayu
Pengujian 4 : Implementasi peraturan	Pengujian penukaran perkataan selepas implementasi peraturan	Separa berjaya	Beberapa perkataan mengeluarkan output yang tidak tepat selepas implementasi peraturan

Pengujian 5 : Implementasi peraturan loghat Kedah	Pengujian penukaran perkataan selepas implementasi peraturan	Separa berjaya	Beberapa perkataan mengeluarkan output yang tidak tepat selepas implementasi peraturan
Pengujian 6 : Penterjemahan bahasa	Pengujian menterjemah bahasa Inggeris	Separa berjaya	Beberapa perkataan mengeluarkan output yang tidak diterjemahkan selepas diterjemahkan.

Mengikut jadual di atas, setiap fasa masih dapat mencapai objektif masing -masing sekaligus mampu menormalisasikan data hingar Twitter. Penambah baikan boleh dilakukan dari masa ke masa supaya lebih banyak data boleh diuji dan mendapat output yang tepat.

6 KESIMPULAN

Kesimpulannya, kajian normalisasi leksikal teks di media sosial telah dibangunkan mengikut objektif kajian dan metodologi kajian yang telah ditetapkan. Kajian ini diharapkan dapat memberi kaedah-kaedah baik kepada pengguna-pengguna.