NORMALISASI LEKSIKAL BAGI PENGGUNAAN LOGHAT PERAK DALAM MEDIA SOSIAL

Maziyyah binti Ahmad Lutfi Amir

Prof. Dr Nazlia binti Omar

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Populasi penggunaan media sosial memberi impak yang tinggi kepada evolusi penggunaan bahasa. Evolusi yang terlibat adalah penggunaan perkataan yang ditransformasikan daripada perkataan bahasa piawai kepada bahasa tidak piawai. Transformasi ini bertujuan meningkatkan ekspresi emosi dan perasaan penulis. Selain itu, rakyat Malaysia juga gemar menggunakan bahasa loghat mengikut negeri mereka dalam penggunaan media sosial. Oleh yang demikian, pra pemprosesan bagi data dari media sosial seperti Twitter agak mencabar. Kajian ini akan menormalisasikan loghat Perak dalam data Twitter kepada bahasa piawai bahasa Melayu. Kajian ini bertujuan untuk pemprosesan bahasa tabii seperti pengelasan kelas kata dan pengekstrakan maklumat Selain itu, kajian ini juga akan menyingkirkan data hingar yang terdapat dalam data Tweet. Teknik yang digunakan dalam kajian ini adalah berasaskan peraturan dengan membina peraturan petua-petua loghat Perak.

1 PENGENALAN

Penggunaan media sosial dalam kalangan masyarakat sudah tidak asing lagi tidak kira peringkat umur. Hampir semua perkara dalam kehidupan seharian menggunakan media sosial. Media sosial yang popular dalam penggunaan masyarakat adalah Facebook, Twitter, Instagram dan Tumblr. Media sosial seperti Twitter atau Facebook menyediakan akses kepada jumlah data yang besar dalam masa nyata, tetapi dipenuhi dengan data hingar yang merumitkan pemprosesan teks. Normalisasi leksikal diperlukan bagi meneutralsasikan data hingar ini kepada data yang boleh diproses menggunakan alat Pemprosesan Bahasa Tabii (PBT).

Loghat adalah kelainan bahasa yang digunakan secara lazim dan ia merupakan yang paling dekat dengan individu. Menurut Siti Noraini Hamzah, variasi-variasi bahasa yang dilihat dari aspek pengguna ditinjau dari segi wilayah atau geografi penutur dan ini melahirkan loghat. (Hamzah & Jalaluddin 2018).

2 PENYATAAN MASALAH

Pengguna media sosial ini bebas untuk memuat naik dan berkongsi data seperti teks, gambar dan video di akaun media sosial mereka. Permasalahannya berbangkit apabila pengguna media sosial yang berkongsi data berbentuk teks tidak menggunakan bahasa, ejaan, tatabahasa, tanda baca dan loghat yang betul. Pengguna lebih gemar mencampur adukkan dua bahasa yang berbeza seperti bahasa Melayu dan bahasa Inggeris dalam penggunaan bahasa mereka di media sosial. Masyarakat sering menggunakan bahasa rojak dalam komunikasi seharian mereka disamping menyampaikan maklumat di media sosial. Perkara ini sudah menjadi kebiasaan dalam kalangan masyarakat apabila mereka mempraktikkan cara mereka berkomunikasi dialam maya sama seperti cara mereka berkomunikasi di dunia realiti. Permasalahan ini akan menyukarkan pemprosesan teks dan pengekstrakan maklumat dari Twitter.

Kajian ini juga memfokuskan analisis perkataan loghat Perak. Buat masa ini, masih belum ada kajian yang menormalisasikan loghat Perak kepada bahasa Melayu piawai. Penggunaan loghat Perak sangat popular dalam kalangan penduduk negeri Perak mahupun luar dari negeri ini. Kesan daripada percampuran penduduk negeri yang bermastautin di negeri yang bukan tempat kelahirannya, masyarakat sekeliling mereka juga dapat menguasai loghat asal negeri mereka. Bahkan masih ada pengguna media sosial yang menggunakan loghat Perak walaupun bukan berasal dari negeri Perak.

Selain itu, data teks di media sosial mempunyai perkataan yang tidak piawai, kesalahan tatabahasa dan kesalahan sintaks. Teks sosial media seperti tweets, messenger, dan komen di dalam Facebook merupakan cabaran dalam bidang NLP (Heureux 2005). Teks daripada media sosial lebih hingar berbanding teks berita. Data hingar ini didefinisikan sebarang perbezaan antara kehendak teks yang ditunjuk dengan teks yang sebenar. Kesannya, data hingar akan menyukarkan alat Pemprosesan Bahasa Tabii (PBT) yang sedia ada untuk memproses data teks kerana tidak menepati piawai dalam bahasa Melayu.

3 OBJEKTIF KAJIAN

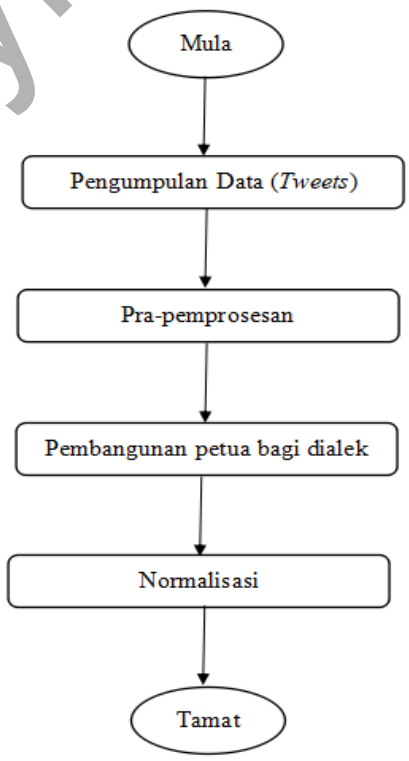
Sumber data teks yang didapati di Twitter amat hingar akan menyebabkan kesukaran alat Pemprosesan Bahasa Tabii (PBT) atau aplikasi lain untuk menterjemah dan mentafsir data ini kepada bahasa Melayu. Objektif kajian ini adalah seperti berikut:

- Membangunkan algoritma loghat Perak bagi proses normalisasi teks
- Menilai prestasi algoritma yang dibangunkan

4 METODOLOGI KAJIAN

Fasa pertama kajian ini akan dimulakan dengan mengumpul data teks yang terdapat di Twitter. Data teks tersebut mestilah mengandungi sekurang-kurangnya satu perkataan yang berunsurkan loghat Perak. Fasa kedua merupakan fasa pra-pemprosesan iaitu fasa yang akan mengeluarkan beberapa data hingar dalam data Tweet sekali gus melakukan proses pembersihan *'cleaning'*. Terdapat tujuh proses yang dilakukan dalam fasa kedua ini. Seterusnya, fasa ketiga merupakan fasa dimana pembangunan petua telah dibangunkan. Kesemua perkataan yang telah melalui fasa kedua akan diuji dengan petua loghat Perak yang telah dibangunkan.

Selepas itu,fasa terkakhir iaitu fasa normalisasi akan diteruskan dengan proses mengenalpasti perkataan yang telah melalui kesemua fasa sebelumnya. Fasa ini juga akan mengeluarkan 'output' bagi data Tweet yang hingar pada permulaannya. Rajah 3.1 menerangkan proses untuk menormalisasikan data teks kepada bahasa Melayu.



Rajah 3.1 Carta aliran normalisasi leksikal

Pada permulaan proses ini, data akan dikumpulkan. Korpus yang akan digunakan adalah korpus Twitter. Mikroblog kini sangat popular dalam kalangan masyarakat dalam aspek berkomunikasi melalui internet. Jutaan mesej akan dihantar dan diterima dalam aplikasi yang popular seperti Twitter. Penulis mesej ini akan menceritakan tentang kehidupan, berkongsi pendapat dalam pelbagai topik dalam mesej yang disampaikan (Záruba 1971).

4.1 FASA 1 : PENGUMPULAN DATA

Fasa ini akan mengumpulkan data daripada media sosial seperti Twitter dan Facebook. Data – data ini akan digunakan dalam setiap fasa. Data yang diambil mestilah mempunyai unsur perkataan loghat Perak bagi memudahkan kajian ini dilakukan. Selepas itu, data ini akan diumpukkan dalam satu fail dan dibahagikan kepada dua bahagian iaitu ‘data set’ dan ‘training set’.

Berikut merupakan ayat contoh data. Ayat di bawah merupakan ayat berunsur loghat Perak.

Sebelum pra-pemprosesan:

"Teman nak balik ghomah dekat Bote, nak masak gulei tempoyok."

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

4.2 FASA 2 : PRA – PEMPROSESAN

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

4.2.1 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 Tweet juga akan dibuang. Twitter merupakan antara media sosial yang menggunakan tanda pagar bagi mengaitkan tweet 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 Tweet dan juga mennyebut nama pengguna tersebut dalam Twitter. Jadual 4.1 contoh data Twitter. Perkataan bergaris merupakan contoh data hingar, URL, tanda pagar atau sebutan pengguna.

Kategori	Sebelum	Selepas
URL	*Tak semestinya orang Perak kena bunyi Kuale yeop. http://wp.me/paA6t9-1cC*	*Tak semestinya orang Perak kena bunyi Kuale yeop.*
Tanda pagar '#'	*10 berita pilihan sepanjang hari ini **#AWANInews** #AWANI745*	*10 berita pilihan sepanjang hari ini*
Sebutan pengguna '@'	*Selamat pagi @ahmad*	*Selamat pagi*

Jadual 4.1 Contoh kategori data

Jadual 4.1 merupakan contoh data Twitter yang digunakan di lama sosial Twitter. Data yang mempunyai unsur kategori di atas akan dikeluarkan daripada teks data asal tersebut.

4.3 FASA 3 : PEMBANGUNAN PETUA BAGI LOGHAT

Pendekatan yang akan digariskan dalam menghasilkan petua loghat Perak ini adalah kaedah berasaskan peraturan (*rule-based approach*). Peraturan ini menggunakan peraturan jika (if) dan maka (then) untuk menguji data. Peraturan akan memproses data menggunakan dua teknik iaitu rantaian kedepan dan rantaian belakang. Melalui kajian (Poria et al. 2015), pendekatan ini dapat memanfaatkan pengetahuan umum dan kebergantunan ayat bagi aspek yang tersirat dan jelas. Melalui fasa ini, peraturan akan diwujudkan mengikut kesesuaian perkataan loghat Perak. Peraturan akan digunakan bagi mengesan perkataan loghat Perak yang mempunyai persamaan ejaan diawal, ditengah atau diakhir perkataan kepada bentuk piawai. Jika petua yang dibangunkan tidak dapat menukarkan loghat Perak, kamus makna loghat Perak akan digunakan. Beberapa peraturan telah di bina seperti berikut dalam jaduaL

4.2.

Jadual 4.2 Petua Perak digunakan dalam loghat Perak

No. Petua	Petua
Petua 1	Jika huruf akhiran **'e'**, maka tukarkan kepada **'a'**
Petua 2	Jika huruf akhiran **'**ei**'** , maka tukarkan kepada **'ai'**
Petua 3	Jika huruf akhiran '**en',** maka tukarkan kepada **'ing'**
Petua 4	Jika huruf awalan '**gho',** maka tukarkan kepada **'ru'**
Petua 5	Jika huruf akhiran '**mor,** maka tukarkan kepada **'mar'**
Petua 6	Jika huruf akhiran **'o'**, maka tukarkan kepada **'ar'**
Petua 7	Jika huruf akhiran **'oh'**, maka tukarkan kepada **'ah'**
Petua 8	Jika huruf akhiran **'ok'**, maka tukarkan kepada **'ak'**
Petua 9	Jika huruf akhiran **'or'**, maka tukarkan kepada **'ur'**
Petua 10	Jika huruf akhiran **'ui'**, maka tukarkan kepada **'ul'**

Melalui peraturan yang telah dibina dalam jadual 4.2, perkataan yang menepati syarat peraturan tersebut akan mengalami normalisasi untuk mengubah perkataan tersebut dalam bahasa Melayu piawai

4.4 FASA 4 : NORMALISASI

Pada fasa ini, kesemua leksikon akan dinormalisasikan kepada bentuk yang piawai. Data Twitter yang dimasukkan ke dalam input akan ditukarkan kepada bentuk piawai bahasa Melayu selepas melalui proses pra-pemprosesan dan pembangunan petua loghat bahasa Perak. Selepas satu petua dibina, petua tersebut akan diuji keberkesanan dan ketepatan dalam proses ini. Proses normalisasi akan bermula. Pendekatan ini telah digunakan oleh Bo Han (2014) untuk bahasa Inggeris. Perbezaan ialah kajian ini pra-pemprosesan telah dilakukan terdahulu supaya proses normalisasi lebih baik daripada kajian terdahulu, dan setakat ini, ini masih tiada kajian tersebut di dalam bahasa Melayu.

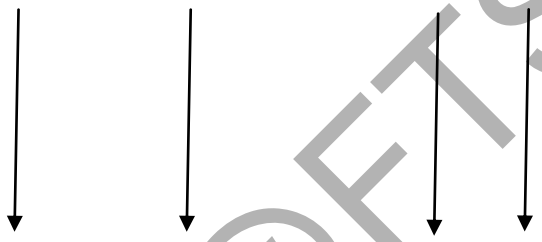
Seterusnya, perkataan yang menepati syarat atau peraturan yang pertama akan menukarkan huruf yang ia kepada huruf baru sekali gus memaparkan output perkataan baru seteleh huruf ditukar. Jika perkataan tersebut tidak menepati syarat yang pertama, ia akan diuji dengan syarat kedua dan seterusnya sehingga tamat aliran ini. Pada akhir fasa ini, kesemua perkataan dalam ayat tersebut akan mengalami normalisasi tidak kira

perkataan berunsur loghat Perak mahupun perkataan yang tidak mengandungi unsur loghat Perak. Dibawah merupakan perubahan ayat sebaik sahaja petua digunakan.

Input:

"Teman hendak balik ghomah dekat Botei, hendak masak gulei tempoyok."

Output:



"Saya hendak balik rumah dekat Bota, nak masak gulai tempoyak."

Berdasarkan contoh di atas, proses normalisasi telah berlaku bagi menormalisasikan ayat asal yang mengandungi perkataan loghat Perak.

5 HASIL KAJIAN

Bahagian ini akan membincangkan proses pembangunan kajian ini. Setiap fasa akan menunjukkan ayat sebelum dan selepas setiap fasa. Output yang dihasilkan bagi kedua-dua data yang dimasukkan akan dihuraikan dalam pengujian ini.

Ayat 1

Sebelum pra-pemprosesan :

“Bukan.. Sbb kata makcik2 zue.. Itu loghat atau dialek org perak..huhuhu RT @brianna0912 : Setahu akak meronjang tu ... http://tmi.me/6K8xv”

Selepas pra-pemprosesan :

“Bukan.. Sbb kata makcik2 zue.. Itu loghat atau dialek org perak..huhuhu : Setahu akak meronjang tu …”

Data Twitter yang dimasukkan dalam pengaturcaraan ini berubah selepas implementasi dalam fasa pertama ini. Perkataan RT, @brianna0912 dan http://tmi.me/6K8xv dalam ayat pertama dapat disingkirkan secara keseluruhannya. Rajah 5.2 merupakan ayat 2 yang digunakan untuk pengujian fasa ini.

Ayat 1

Sebelum kamus :

‘’Ate betui ke idak dapat bonus? Awok memikior dari tadi ni.’’

Selepas kamus :

‘’habis itu betui ke tidak dapat bonus? saya fikir dari tadi ini .’’

Ayat 1 menunjukkan beberapa perkataan loghat Perak yang wujud dalam kamus Perak. Perkataan tersebut telah ditukarkan kepada perkataan bahasa Melayu. Binaan kamus ini berjaya kerana mampu menukarkan perkataan loghat Perak kepada perkataan bahasa Melayu. Penambahbaikan boleh difokuskan kepada penambahah bilangan perkataan loghat Perak supaya lebih banyak perkataan boleh ditukarkan. Pengaturcaraan bagi fasa ini juga harus diperbaiki bagi menghasilkan sistem yang efisen

Ayat 1

Sebelum petua :

“Aku rasa, negeri perak je kot, yg ada byk dialek. Kami mempunyai dialek kuale, batu gajah, tepen, lenggong. Ipoh pulak, ckp mcm org kunci menjerit. Perth”

Selepas petua :

“Aku rasa , negeri perak je kot , yg ada byk Dialek . Kami mempunyai dialek kuala , batu gajah , Teping , Lenggong . Ipah pulak , ckp mcm org kunci menjerit . Perth”

Perbandingan di atas merupakan sebelum dan selepas implementasi peraturan. Melalui perbandingan ini ayat 1, perkataan ‘kuale’ dapat ditukarkan kepada ejaan yang betul iaitu ‘kuala’. Manakala perkataan ‘Tepen’ dan ‘Ipoh’ ditukarkan kepada ejaan yang salah. Kebanyakan peraturan yang digunakan merupakan peraturan yang dibina selepas mengenalpasti ejaan dalam perkataan loghat Perak. Tetapi ia tidak dapat diaplikasikan bagi kedua-dua nama tempat berikut. Cadangan bagi penambahbaikan ini adalah dengan meletakkan ejaan nama khas ke dalam senarai kamus. Dengan cara ini, ejaan ni dapat terus ditukarkan kepada ejaan yang betul.

Keterangan di bawah merupakan contoh ayat 1 yang mengandungi perkataan unsur bahasa Inggeris.

Sebelum terjemahan :

“Aku rasa, negeri perak je kot, yg ada byk dialek. We have dialek kuale, batu gajah, tepen, lenggong. Ipoh pulak, ckp mcm org key yell. Pergh”

Selepas terjemahan :

“Aku rasa, negeri perak je kot, yg ada byk dialek. Kami mempunyai dialek kuale, batu gajah, tepen, lenggong. Ipoh pulak, ckp mcm org kunci menjerit. Perth”

Perkata diatas merupakan perbandingan data input sebelum dan selepas pengujian terjemahan. Perkataan ‘We’ dan ‘have’ dapat diterjemahkan dengan maksud yang betul manakala perkataan ‘key’ dan ‘yell’ juga diterjemahkan dengan betul namun telah mengubah maksud sebenar ayat. ‘key’ ‘yell’ yang dimaksudkan oleh penulis data adalah ‘KL’ = ‘Kuala Lumpur’. Kesalahan seperti ini merupakan kesalahan diluar jangkaan akan tetapi objektif untuk menormalisasikan ayat ini mampu dicapai dengan fasa terjemahan ini. Bagi perkataan ‘Pergh’ pula, ia merupakan yang mengekspresi perasaan dalam ayat ini. Tetapi ia telah ditukarkan kepada nama tempat ‘Perth’.

Hasil kajian mendapati fasa pertama hingga fasa akhir bagi kajian ini hampir menepati objektif kajian iaitu menormalisasi data yang mengandungi loghat Perak 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 Perak	Pengujian penukaran secara terus bagi perkataan loghat Perak	Berjaya	Pengujian dapat menukarkan perkataan loghat Perak kepada bahasa Melayu
Pengujian 3 : Implementasi peraturan loghat Perak	Pengujian penukaran perkataan selepas implementasi peraturan	Separa berjaya	Beberapa perkataan mengeluarkan output yang tidak tepat selepas implementasi peraturan
Pengujian 4 : Penterjemahan bahasa	Pengujian menterjemah bahasa Inggeris	Berjaya	Mengenalpasti bahasa yang kerap digunakan selain bahasa Inggeris

Mengikut jadual di atas, setiap fasa masih dapat mencapai objektif masing -masing sekali gus mampu menormalisasikan data loghat Perak. Penambahbaikan boleh dilakukan dari masa ke masa supaya lebih banyak data boleh diuji dan mendapat output.

6 KESIMPULAN

Penggunaan loghat dalam media sosial tidak asing lagi dalam kalangan masyarakat di Malaysia. Terdapat banyak loghat yang digunakan berbeza mengikut negeri, daerah dan etnik. Normalisasi bagi mengubah ayat yang mengandungi unsur loghat Perak kepada bahasa Melayu amat penting bagi menghasilkan data – data yang berkualiti untuk digunakan dalam kajian akan datang. Data yang mengandungi unsur loghat ini serba sedikit akan menganggu ketetapan bagi sesuatu pengujian data.

Kajian ini mampu menterjemahkan perkataan loghat Perak dengan menggunakan pendekatan sama ada petua Perak yang dibangunkan atau kamus Perak. Kajian ini juga mampu menyingkirkan simbol – simbol yang akan menganggu proses normalisasi. Terdapat 10 petua yang telah digunakan bagi menukar suku kata awalan atau akhiran loghat Perak. Kamus yang telah dibangunkan pula mengumpulkan 150 perkataan loghat Perak beserta makna dalam bahasa Melayu. Disebabkan rakyat Malaysia kerap menggunakan bahasa rojak iaitu percampuran dua bahasa yang berbeza dalam satu ayat, terjemahan secara terus dilakukan dalam kajian ini bagi menterjemahkan bahasa Inggeris kepada bahasa Melayu.

7 RUJUKAN

Hamzah, S. N. & Jalaluddin, N. H. 2018. Kepelbagaian Varian Leksikal Loghat di Perak : Pendekatan Geographical Information System. *Akademika* 88(April): 137–152.

Heureux, L. 2005. Yr Ig Ht Yr Ig 1–14. doi:10.1093/jac/dkq328

Nor Hashimah Jalaluddin. 2018. Loghat Melayu di Perak : Analisis Geolinguistik. *International Journal of the Malay World and Civilisation* 6(2): 69–82.

Cook, P. & S. Stevenson 2009. An unsupervised model for text message normalization. Proceedings of the workshop on computational approaches to linguistic creativity. pp. 71-78.

Ahmed, B., Cha, S.-H. & Tappert, C. 2004. Language Identification from Text Using N-gram Based Cumulative Frequency Addition Collection of Text Samples and Creation of N-gram Profiles. *Proceedings of Student/Faculty Research Day, CSIS, Pace University* 1–8.

Choochart Haruechaiyasak & Alisa Kongthon. 2013. LexToPlus: A Thai Lexeme Tokenization and Normalization Tool. *The 4th Workshop on South and Southeast Asian NLP (WSSANLP) under the 6th International Joint Conference on Natural Language Processing (IJCNLP)* (International Joint Conference on Natural Language Processing): 9–16.

Clark, E. & Araki, K. 2011. Text normalization in social media: Progress, problems and applications for a pre-processing system of casual English. *Procedia - Social and*

Behavioral Sciences 27(Pacling): 2–11. doi:10.1016/j.sbspro.2011.10.577

Dirkson, A., Verberne, S. & Kraaij, W. 2019. Lexical Normalization of User-Generated Medical Text 11–20. doi:10.18653/v1/w19-3202

Hamzah, S. N. & Jalaluddin, N. H. 2018. Kepelbagaian Varian Leksikal Dialek di Perak : Pendekatan Geographical Information System. *Akademika* 88(April): 137–152.

Heureux, L. 2005. Yr Ig Ht Yr Ig 1–14. doi:10.1093/jac/dkq328

Kane, S. N., Mishra, A. & Dutta, A. K. 2016. Preface: International Conference on Recent Trends in Physics (ICRTP 2016). *Journal of Physics: Conference Series* 755(1): 2–9. doi:10.1088/1742-6596/755/1/011001

Kotselidis, C. & Luj, M. (n.d.). Clustering JVMs with Software Transactional Memory Support.

Mansur, M. 2006. Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus. *BRAC University* (August).

Nor Hashimah Jalaluddin. 2018. Dialek Melayu di Perak : Analisis Geolinguistik. *International Journal of the Malay World and Civilisation* 6(2): 69–82.

Poria, S., Cambria, E., Ku, L.-W., Gui, C. & Gelbukh, A. 2015. A Rule-Based Approach to Aspect Extraction from Product Reviews 28–37. doi:10.3115/v1/w14-5905

Saloot, M. A., Idris, N. & Mahmud, R. 2014. An architecture for Malay Tweet normalization. *Information Processing and Management* 50(5): 621–633. doi:10.1016/j.ipm.2014.04.009

van der Goot, R. 2019. MoNoise: A Multi-lingual and Easy-to-use Lexical Normalization Tool 201–206. doi:10.18653/v1/p19-3032

Záruba, F. 1971. Problémy keratinizace. C. Predběžné závěry. *Cesko-Slovenska Dermatologie* 46(5): 223–229.

Asmah Hj. Omar. 1993. Susur galur bahasa Melayu. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Ahmed, B., Cha, S.-H. & Tappert, C. 2004. Language Identification from Text Using N-gram Based Cumulative Frequency Addition Collection of Text Samples and Creation of N-gram Profiles. *Proceedings of Student/Faculty Research Day, CSIS, Pace University* 1–8.

Choochart Haruechaiyasak & Alisa Kongthon. 2013. LexToPlus: A Thai Lexeme Tokenization and Normalization Tool. *The 4th Workshop on South and Southeast Asian NLP (WSSANLP) under the 6th International Joint Conference on Natural Language Processing (IJCNLP)* (International Joint Conference on Natural Language Processing): 9–16.

Clark, E. & Araki, K. 2011. Text normalization in social media: Progress, problems and applications for a pre-processing system of casual English. *Procedia - Social and Behavioral Sciences* 27(Pacling): 2–11. doi:10.1016/j.sbspro.2011.10.577

Dirkson, A., Verberne, S. & Kraaij, W. 2019. Lexical Normalization of User-Generated Medical Text 11–20. doi:10.18653/v1/w19-3202

Hamzah, S. N. & Jalaluddin, N. H. 2018. Kepelbagaian Varian Leksikal Dialek di Perak : Pendekatan Geographical Information System. *Akademika* 88(April): 137–152.

Heureux, L. 2005. Yr Ig Ht Yr Ig 1–14. doi:10.1093/jac/dkq328

Kane, S. N., Mishra, A. & Dutta, A. K. 2016. Preface: International Conference on Recent Trends in Physics (ICRTP 2016). *Journal of Physics: Conference Series* 755(1): 2–9. doi:10.1088/1742-6596/755/1/011001

Kotselidis, C. & Luj, M. (n.d.). Clustering JVMs with Software Transactional Memory Support.

Mansur, M. 2006. Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus. *BRAC University* (August).

Nor Hashimah Jalaluddin. 2018. Dialek Melayu di Perak : Analisis Geolinguistik. *International Journal of the Malay World and Civilisation* 6(2): 69–82.

Poria, S., Cambria, E., Ku, L.-W., Gui, C. & Gelbukh, A. 2015. A Rule-Based Approach to Aspect Extraction from Product Reviews 28–37. doi:10.3115/v1/w14-5905

Saloot, M. A., Idris, N. & Mahmud, R. 2014. An architecture for Malay Tweet normalization. *Information Processing and Management* 50(5): 621–633. doi:10.1016/j.ipm.2014.04.009

van der Goot, R. 2019. MoNoise: A Multi-lingual and Easy-to-use Lexical Normalization Tool 201–206. doi:10.18653/v1/p19-3032

Záruba, F. 1971. Problémy keratinizace. C. Predběžné závěry. *Cesko-Slovenska Dermatologie* 46(5): 223–229.

Han, B. 2014. Improving the utility of social media with Natural Language Processing.Tesis The University of Melbourne,

Gilabert, P. L., Gadringer, M. E., Montoro, G., Mayer, M. L., Silveira, D. D., Bertran, E. &

Magerl, G. 2009. An efficient combination of digital predistortion and ofdm clipping for power amplifiers. *International Journal of RF and Microwave Computer-Aided Engineering* 19(5): 583–591. doi:10.1002/mmce.20381