Nombor Prosiding 38

Pra-pemprosesan untuk Data Hingar di dalam Analisis Sentimen untuk Dokumen Bahasa Melayu

Mazlina Mustapha[1]*, Ramlah Md. Zain[1]

[1]Department of Information and Communication Technology, Politeknik Sultan Mizan Zainal Abidin, 23000 Dungun, Terengganu.
***Corresponding author E-mail**: mazlina@psmza.edu.my

Abstrak

Perkembangan teknologi telah mengubah cara berkomunikasi berkenaan pendapat servis dan produk. Dalam kepenggunaan, cabaran sebenar adalah untuk memahami trend terkini dan merumuskan pendapat umum mengenai sesuatu produk kerana saiz data dan kepelbagaian media sosial di dalam talian. Sebanyak 2000 komen filem telah dikumpul dari laman web forum dan blog Malaysia. Kajian ini membincangkan kerja analisis pra-pemprosesan untuk dokumen Bahasa Melayu dari dua perspektif. Pertama, beberapa alternatif perwakilan teks disiasat. Kedua, strategi pra-pemprosesan iaitu proses penormalan, pemecahan teks kepada token, menyingkir kata henti dan proses *stemming*. Apabila pra-pemprosesan dilaksanakan, sentimen akan menunjukkan aksi pengguna ke atas produk samada positif atau negatif. Kebanyakan organisasi mendapati bahawa keputusan penentuan pendapat adalah satu bahagian penting sebagai kayu ukur untuk menyokong bisnes secara pintar. Pendapat sosial ini akan diaplikasi secara efektif untuk kajian analisis sentimen. Kajian ini menentukan teknik pra-pemprosesan yang berkesan untuk menentukan sentimen positif atau negatif.

Kata Kunci: pra-pemprosesan; data hingar; analisis sentimen; dokumen BM

1.0 PENGENALAN

Dengan kemunculan Web 2.0, internet boleh diakses pada bila-bila masa dan di mana sahaja mengikut topik yang menarik perhatian pengguna. Kemudahan untuk mengakses internet terutamanya dengan menggunakan telefon pintar telah membuka ruang kepada pengguna menggunakan blog-mikro secara meluas. Kemudahan ini telah membuka ruang untuk pengguna menggunakan bahasa secara tidak formal apabila pengguna berkomunikasi sesama mereka merentasi sempadan negara. Lantas, singkatan perkataan dan penggunaan simbol mulai muncul satu persatu dan menjadi popular. Budaya ini akan menyebabkan penggunaan BM tercemar terutamanya di kalangan generasi muda dan ia akan menjadi satu cabaran untuk mengekalkan ketulenan BM kepada generasi akan datang.

Twitter merupakan salah satu sistem blog-mikro yang membenarkan pengguna menulis menggunakan mesej pendek yang dikenali sebagai *tweet*. Melalui data *tweet* pengguna boleh mendapatkan maklumat yang berguna dengan mengkelaskan kepada beberapa kategori. Contohnya, kandungan blog-mikro boleh dikelaskan berdasarkan sentimen, kandungan, pendapat, berita, topik dan sebagainya. Walaubagaimanapun, kandungan teks blog-mikro adalah hingar, ambiguiti, menggunakan singkatan dan bahasa dialek dan pelbagai gaya penulisan. Tambahan pula, teks blog-mikro juga mempunyai

simbol, pautan, emoji dan sebagainya yang menyebabkan pengkaji perlu menukar data hingar ini kepada bahasa piawai. Lantaran itu, banyak kajian perlu dilakukan untuk mengekalkan ketulenan BM ini. Kerja pra-pemprosesan adalah penting untuk mengatasi masalah data hingar dimana ia adalah salah satu kerja di dalam kajian analisis sentimen (Roy, Dhar, Bhattacharjee, & Das, 2013; Saloot, Idris, & Mahmud, 2014).

Analisis Sentimen merupakan kajian yang menganalisa pendapat, sentimen, penilaian, kelakuan dan emosi daripada tulisan seseorang (Liu, 2010b). Ia melibatkan pengkelasan polariti teks di dalam dokumen atau ayat yang menyatakan pendapat dalam kelas positif, negatif, atau neutral (Pang & Lee, 2008). Secara umumnya, Analisis Sentimen adalah proses pengkelasan, tetapi hakikatnya ia tidak semudah proses pengkelasan biasa kerana ia bengantung kepada penggunaan bahasa penulis yang mana terdapat rujukan taksa dalam penggunaan kata, tiada anotasi dalam sesebuah teks dan perkembangan bahasa itu sendiri (Liu, 2010a). Motivasi utama kajian ini adalah memfokus kepada Fasa Pra-pemprosesan di dalam pengkelasan sentimen untuk dokumen BM.

2.0 PERNYATAAN MASALAH

Terdapat beberapa isu berkaitan bahasa singkatan dan elemen visual semasa penggunaan bahasa dalam komunikasi digital tidak formal samada dalam emel, twitter, forum, chat dan SMS. Supyan (2010) mendapati terdapat dua pola bahasa yang digunakan: (a) bahasa alih mengalih-kod dan (b) bahasa rojak. Penggunaan pola bahasa ini telah menyebabkan pencemaran ejaan dalam BM dan menyebabkan data hingar. Masalah ini menjadi satu cabaran kepada pengkaji semasa proses pra-pemprosesan data ulasan dan komen filem BM. Pengkaji akan memfokus kepada dua proses di dalam Fasa Pra-pemprosesan isitu proses penormalan dan *Stemming*.

Kebanyakan kajian lepas menganalisa pendapat menggunakan dokumen Bahasa Inggeris. Tetapi malangnya hanya sedikit kajian tentang Analisis Sentimen bagi dokumen BM dijalankan. Menurut (Norlela et al. 2011), sumber yang terhad bagi kajian Analisis Sentimen bagi dokumen BM mendorong pengkaji untuk menerima cabaran di dalam bidang ini. Pengkaji akan membangunkan satu Korpus lengkap bagi dokumen BM kerana buat masa sekarang korpus ini masih belum ada.

3.0 KAJIAN BERKAIT

Fasa Pra-pemprosesan ini mempunyai proses penormalan, membuang kata henti, pemecahan teks kepada token dan *Stemming*. Pengkaji memilih dua jenis algoritma *stemmer* iaitu *Stemmer* Othman dan *Stemmer* Fatimah. Proses penormalan adalah proses mengurangkan ambiguiti dan mengurangkan data hingar dengan meneliti setiap teks di dalam sentimen. Pengkaji juga menterjemah komen yang menggunakan bahasa dialek kepada BM formal dan menterjemah perkataan Bahasa Inggeris yang terdapat di dalam komen campuran.

Proses *Stemming* adalah proses mencantas imbuhan yang terdapat di dalam sesuatu perkataan kepada kata akar mengunakan *Stemmer*. Proses *Stemming* bagi dokumen BM adalah lebih kompleks jika dibandingkan dengan proses *Stemming* bagi dokumen Bahasa Inggeris. Kebanyakan algoritma *Stemming* untuk dokumen BM menggunakan strategi penghapusan imbuhan berasaskan peraturan berbanding dengan strategi lain kerana BM adalah bahasa yang rumit (Rayner et al. 2014). Pengkaji memilih *Stemmer* Othman (Othman, 1993) kerana *Stemmer* ini sesuai untuk teks umum dan aplikasi di dalam BM dan *Stemmer* Fatimah (Fatimah et al. 1996) adalah *stemmer* yang telah melalui proses penambahbaikan daripada *Stemmer* Othman.

3.1 Kajian Lepas Pra-pemprosesan yang menggunakan Proses Penormalan Data Hingar

Terdapat banyak kajian penormalan teks BI (Roy et al. 2013; Saloot et al. 2014) dan bukan BI seperti Bahasa German (Anjaria et al. 2014) dan Bahasa Cina (Hu & Zhao 2010). Malangnya, kajian Analisis Sentimen yang memfokus kepada proses penormalan di dalam Fasa Pra-pemprosesan untuk dokumen BM adalah amat terhad (Norlela et al. 2011). Terdapat beberapa kajian yang menggunakan proses penormalan bagi dokumen BM telah dilaksanakan oleh (Basri et al. 2012; Saloot et al. 2014; Norlela et al. 2013). Walaubagaimanapun, kajian ini memfokus kepada penormalan teks BM.

3.2 Kajian Lepas Pra-pemprosesan yang menggunakan *Stemmer*

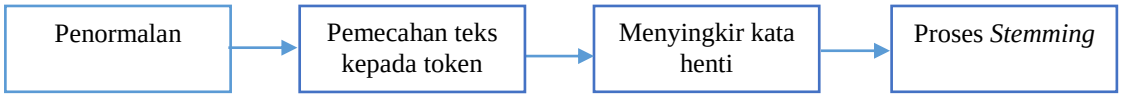
Proses *Stemming* adalah proses untuk mengurangkan saiz perkataan yang mempunyai kata imbuhan kepada kata akar dengan menyingkir imbuhan awalan, imbuhan tengah dan imbuhan akhir. Saiz dokumen juga dapat dikurangkan kerana perkataan yang mempunyai kata akar yang sama akan diproses tanpa mengira imbuhan asal pada perkataan tersebut. Sebagai contoh, untuk perkataan *makan*, *pemakanan* dan *makanan* mempunyai kata akar yang sama iaitu perkataan *makan*. Proses *Stemming* ini dapat mempercepatkan proses mencapai maklumat dan digunakan secara meluas di dalam pelbagai bidang seperti pengkomputeran linguistik dan capaian maklumat, pengurusan maklumat dan Analisis Sentimen (Vallbé et al. 2007; Mazidah et al. 2013; Norulhidayah et al. 2013) sebagai salah satu kaedah pra-pemprosesan untuk membantu meningkatkan keputusan keupayaan yang efisyen.

Terdapat sedikit kajian yang menggunakan *Stemmer* di dalam fasa pra-pemprosesan untuk Analisis Sentimen bagi teks BM (Mazidah et al. 2013; Norulhidayah et al. 2013). Kebanyakan pengkaji di dalam Analisis Sentimen tidak menggunakan proses *Stemming* dan menggunakan proses lain di dalam fasa pra-pemprosesan. Beberapa pengkaji yang menggunakan teks Bahasa Arab (Duwairi & El-orfali 2013), Bahasa Portugis (Santos & Ladeira 2014) dan Bahasa Rusia (Yussupova et al. 2012) telah menggunakan proses *Stemming* di dalam fasa pra-pemprosesan untuk Analisi Sentimen, tetapi sumber amat terhad

bagi kajian Analisis Sentimen bagi dokumen BM. Oleh sebab itu, pengkaji akan menyiasat kepentingan proses *Stemming* di dalam Fasa Pra-pemprosesan bagi dokumen BM.

4.0 EKSPERIMEN PRA-PEMPROSESAN

Fasa Pra-Pemprosesan adalah proses untuk menyediakan data untuk proses pengkelasan. Rajah 1 menunjukkan empat jenis proses pra-pemprosesan iaitu penormalan, pemecahan teks kepada token, menyingkir kata henti dan proses *Stemming*. Pengkaji lepas hanya memfokus kepada salah satu proses samada proses penormalan sahaja atau proses *Stemming* sahaja di dalam fasa Pra-Pemprosesan. Sumber yang terhad bagi kajian Analisis Sentimen untuk dokumen BM mendorong pengkaji untuk mencari kesan Fasa Pra-Pemprosesan yang mengandungi proses penormalan dan proses *Stemming*.



Rajah 1: Langkah-langkah pra-pemprosesan

4.1 Proses Penormalan

Penormalan ialah proses untuk mengurangkan ambiguiti dan data hingar di dalam sentimen. Terdapat banyak cara menulis sentimen untuk menyatakan pendapat di dalam media sosial seperti *Twitter*, *Facebook* dan forum dalam talian. Kajian ini memfokus kepada dokumen BM sahaja. Sebelum proses penormalan, aktiviti awal berikut telah dilaksanakkan:

i) Komen BI akan dibuang daripada komen campuran BM dan BI di dalam satu paragraf.
ii) Komen campuran BM dan BI di dalam satu ayat yang ringkas akan diterjemah. kepada BM formal. Contoh: *teaser dia menarik...harap2 jalan cite pun best la.*
iii) Komen BI sepenuhnya dalam satu ayat akan dibuang.
iv) Komen yang mempunyai alamat pautan yang bermula dengan “http://” atau “www” telah dibuang.
v) Komen yang menggunakan simbol seperti ikon senyum, ikon animasi, alias (@), hash (#) dan tanda tanya (?) di buang semasa proses ini.
vi) Komen yang menggabungkan lebih dari satu perkatan dan mengandungi tanda baca akan diasingkan. Contoh : jom#tonton@filemcinderella

Kebanyakan pengguna Melayu menggunakan bahasa dialek semasa menulis pendapat yang menyebabkan proses kajian menjadi semakin kompleks. Lagipun, maksud frasa akan berubah semasa proses transformasi ke BM formal dan ia akan membawa maksud yang berbeza daripada teks asal. Jadual 1 menunjukkan contoh bahasa dialek yang digunakan oleh bangsa Melayu. Masalah tersebut akan menjadi satu cabaran kepada kepada pengkaji semasa menjalankan proses ini.

Jadual 1: Contoh Bahasa Dialek

Perkataan dialek	BM formal	BI
Magghii	Mari	come
Takboh	Tidak mahu	Don't want
Hang	Kamu	you

Setiap perkataan yang menjalani proses penormalan bagi kajian ini diterjemahkan berdasarkan kepada ("Dewan Bahasa dan Pustaka Malaysia," 2008; Pustaka, 2008); dan langkahnya adalah seperti berikut:

i) **Komen yang menggunakan Bahasa dialek akan diterjemah kepada BM formal**: diterjemahkan kepada BM formal yang mempunyai maksud yang sama.
cite → cerita *antoo → hantu*
watpe → buat apa *semo → semua*

ii) **Perkataan yang mengandungi nombor**: Perkataan yang berakhir dengan nombor yang membawa maksud perkataan berulang dua kali di eja dengan betul. Nombor akan diasingkan dengan betul.
Mati2 → mati-mati *3jam → 3 jam*

iii) **Perkataan yang menggunakan satu aksara sahaja**: diterjemahkan kepada BM formal.
g → *pergi* *d → di*

iv) **Perkataan yang menggunakan beberapa aksara di depan**: diterjemahkan kepada BM formal.
tel→ telefon *no → nombor*

v) **Perkataan yang menggunakan konsonan sahaja**: diterjemahkan kepada BM formal.
pd → pada *kpd → kepada*

vi) **Perkataan yang menggunakan aksara di depan dan di belakang sahaja**: diterjemahkan kepada BM formal.
yg → yang

vii) **Perkataan yang menggunakan suku kata akhir sahaja**: diterjemahkan kepada BM formal.
dah → sudah *mak → emak*

viii) **Huruf yang berulang**: hadkan pengulangan perkataan yang sama kepada 3 suku kata.
Hehehehehehehe → hehehe

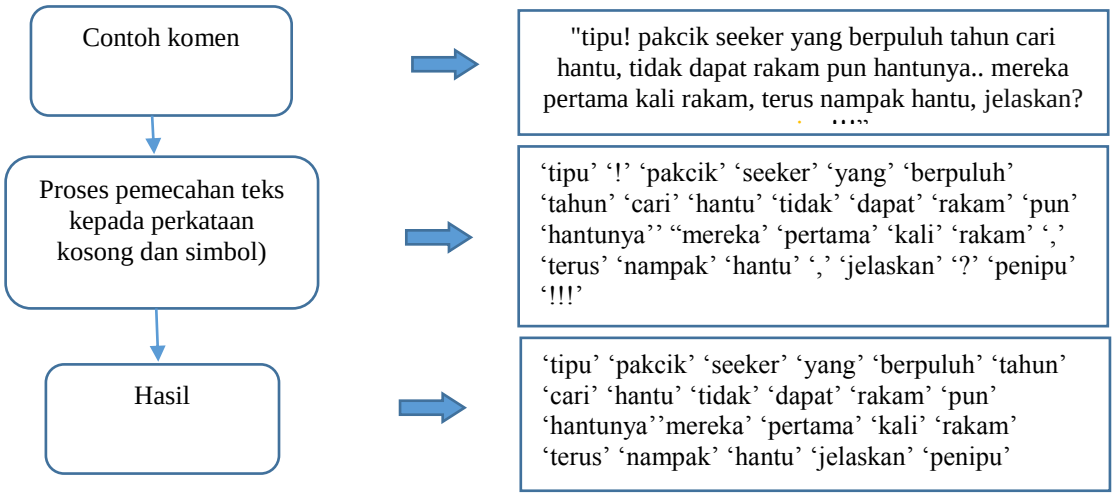
hahahahahaha → hahaha
uwaaaaaaaaaa → uwaaa

ix) **Perkataan yang menggunakan singkatan**: diterjemahkan kepada BM formal berdasarkan rujukan di dalam (Pustaka, 2008).
bl → bila *krn → kerana*

x) **Perkataan ambiguiti:** diterjemahkan kepada BM formal.
dh → sudah *sdh → sudah* *sudaa → sudah* *sdah → sudah*

4.2 Pemecahan Teks kepada token

Setiap ayat komen akan dipecahkan kepada siri perkataan berdasarkan pengesanan ruang kosong *(space)* di antara dua perkataan dan membuang karakter simbol seperti tanda koma, kata seru, tanda titik dan lain-lain.



Rajah 2: Proses pemecahan teks kepada token

Rajah 2 menggambarkan contoh bagaimana teknik pemecahan teks kepada token dilakukan. Bagi contoh di atas, proses pemecahan teks kepada token ini berjaya mengurangkan komen daripada 27 perkataan kepada 22 perkataan untuk diproses di peringkat yang seterusnya.

4.3 Tapisan Kata Henti

Proses tapisan kata henti membolehkan sistem menyingkirkan senarai kata henti yang kerap wujud di dalam komen tetapi tidak memberi makna dan mempengaruhi maksud sesuatu ayat. Jenis perkataan yang dianggap kata henti adalah seperti kata ganti nama, kata hubung, dan kata tunjuk (Samarakoon et al. 2011).

Jadual 2: Contoh senarai Kata henti

Senarai Kata Henti yang diabaikan semasa Pra-pemprosesan				
ada	dahulu	hendaklah	kecuali	merekapun
adakah	dalam	hingga	kelak	meskipun
adakan	dalamnya	ia	kembali	mu

Jadual 2 menunjukkan contoh senarai kata henti adalah seperti perkataan dan, akan, atau, yang, tetapi dan sebagainya. Penyingkiran kata henti ini membolehkan sistem mengambil hanya data penting sahaja untuk di proses ke peringkat seterusnya iaitu proses *Stemming*. Penyingkiran kata henti ini juga boleh mengurangkan saiz Korpus BM dan seterusnya boleh mempercepatkan proses pengkelasan dan capaian maklumat.

4.4 Proses *Stemming*

Proses *Stemming* adalah proses mencantas perkataan yang mengandungi kata imbuhan awal, tengah dan akhir bagi setiap perkataan bagi mendapatkan kata akar yang sebenar dengan menggunakan *Stemmer*. Sebagai contoh, kata akar ‘baca’ mempunyai pelbagai variasi perkataan imbuhan seperti “bacaan”, “dibaca”, “bacalah”, “terbaca” dan banyak lagi tetapi semua perkataan tersebut menjurus kepada satu kata akar yang sama.

Proses *Stemming* digunakan secara meluas di dalam bidang perkomputeran linguistik, capaian maklumat, pengurusan maklumat dan Analisis Sentimen (Mazidah et al. 2013; Norulhidayah et al. 2013; Vallbé et al. 2007). Proses *Stemming* ini boleh mengurangkan saiz indeks fail dan meningkatkan keupayaan capaian maklumat dengan mengurangkan saiz data. Apabila semua perkataan digunakan sebagai atribut input, maka pangkalan data leksikal akan menjadi besar (Nor Azman & Nazlia 2008). Sehubungan itu, proses *Stemming* merupakan cara terbaik bagi menangani masalah tersebut kerana proses ini akan menghasilkan data yang penting sahaja dan dapat membantu meningkatkan keputusan keupayaan Analisis Sentimen yang lebih baik. Proses ini juga dapat meningkatkan Kejituan dan Dapatan Semula bagi Analisis Sentimen untuk Korpus BM.

Jadual 3: Set kata imbuhan *(affix)* bagi perkataan BM

Imbuhan Awal *(Prefix)*	*‘ber’, ‘per’, ‘ter’, ‘mem’, ‘pem’, ‘menge’, ‘penge’, ‘meng’, ‘peng’,* *‘men’, ‘pen’, ‘me’, ‘pe’, ‘be’, ‘ke’, ‘se’, ‘te’, ‘di’*
Imbuhan Akhiran *(Suffix)*	*‘nya’, ‘kan’, ‘an’, ‘i’, ‘kah’, ‘lah’, ‘pun’, ‘ita’, ‘man’, ‘wan’, ‘wati’,* *‘ku’, ‘mu’*
Imbuhan Awalan dan Akhiran	*‘ber…an’, ‘per…an’, ‘ter…kan’, ‘mem…kan’, ‘pem…an’, ‘pen…an’,* *'pe…an’, ‘ke…an’, ‘se…an’, ‘te…kan’, ‘di…kan’, ‘ber…kan’, ‘me…i’,* *‘men…i’, ‘meng…i’, ‘menge…kan’, ‘penge…an’,*

	'peng…an'
Penggunaan dua atau lebih imbuhan	*'diper…', '…kannya', 'memper…i', 'berke…an', 'men…inya', 'di…kannya'*

Stemmer pertama adalah *Stemmer* BI yang dihasilkan oleh Lovin pada tahun 1968, diikuti oleh Dowson pada tahun 1974 dan Porter pada tahun 1980. Terdapat juga *Stemmer* bukan BI yang telah dibangunkan seperti *Stemmer* BM, Bahasa Arab, Spanish dan Cina. Kajian ini memfokus kepada *Stemmer* untuk dokumen BM.

Penyelidikan awal dalam *Stemmer* untuk dokumen BM telah dilaksanakan oleh Othman (1993). Berikutan daripada itu beberapa *Stemmer* BM telah dihasilkan seperti *Stemmer* menggunakan Algoritma Fatimah (Fatimah et al. 1996), Idris & Syed Mustapha (2001) dan Yasukawa et al. (2009) (Salhana et al. 2012). Bagi tujuan kajian ini, penyelidik menggunakan 2 jenis algoritma cantasan iaitu SAO (Othman 1993) dan SAF (Fatimah et al. 1996) sebagai rujukan dalam membangunkan algoritma cantasan kata akar. Kata akar bagi perkataan BM untuk *Stemmer* SAO (Othman, 1993) dan SAF (Ahmad et al, 1996) berpandukan ("Dewan Bahasa dan Pustaka Malaysia," 2008). Pengkaji akan membandingkan hasil cantasan dua jenis *Stemmer* ini ke atas data kajian.

5.0 KEPUTUSAN DAN PERBINCANGAN

Sebanyak 2000 komen filem telah dikumpul dari laman web forum dan blog Malaysia. Tujuan utama eksperimen ini adalah menghasilkan satu korpus BM dengan membuang data hingar dari laman web forum dan blog BM untuk analisa lanjutan. Objektif kerja pra-pemprosesan adalah untuk membuang semua aksara yang tidak bermakna dan hanya mengambil perkataan yang bermakna sahaja.

Secara umumnya, pra-pemprosesan boleh dikelaskan kepada dua kategori iaitu pra-pemprosesan biasa dan pra-pemprosesan spesifik (Hidayatullah & Ma'arif, M, 2016). Pra-pemprosesan yang biasa dilaksanakan adalah pemecahan teks kepada token, menyingkir kata henti, membuang tanda bacaan, simbol dan proses *stemming*. Manakala pra-pemprosesan spesifik dilaksanakan ke atas sentimen dari media sosial seperti blog, twitter, facebook dan web forum. Sentimen media sosial adalah hingar dimana pengguna gemar menggunakan simbol, hashtag, emoji, bahasa dialek, bahasa rojak dan bahasa singakatan. Proses pra-pemprosesan spesifik ini amat mencabar kepada pengkaji dimana pengkaji perlu melaksanakan proses penormalan untuk mengurangkan ambiguiti dan data hingar di dalam sentimen.

Kajian ini menunjukkan keberkesanan kerja pra-pemprosesan untuk korpus BM. Kerja pra-pemprosesan ini memberi kesan kepada keputusan pengkelasan analisa sentimen untuk dokumen BM (M. Arif & Mustapha, 2017; Norlela Samsudin, Mazidah Puteh, Ahmad

Nazmi Fadzal, & Tajuddin, 2013; Saloot, Idris, & Mahmud, 2014). Kajian ini menentukan teknik pra-pemprosesan yang berkesan seperti penormalan, pemecahan teks kepada, tapisan kata henti dan proses *stemming* adalah penting dalam untuk meningkatkan prestasi kajian analisis sentimen. Setelah pra-pemprosesan dilaksanakan, sentimen akan ditentukan samada sentimen positif atau negatif.

6.0 KESIMPULAN DAN KAJIAN MASA DEPAN

Kajian ini menunjukkan kerja pra-pemprosesan untuk dokumen BM untuk sentimen laman web forum dan blog. Kajian mendapati laman web forum dan blog BM adalah hingar disebabkan trend pengguna yang menggunakan singkatan perkataan untuk menulis komen. Penyelidik mendapati kerja pra-pemprosesan ini perlu dilaksanakan agar dapat megurangkan data hingar dan untuk meningkatkan keputusan pengkelasan Analisis Sentimen. Teknik pra-pemprosesan perlu ditentukan bergantung kepada jenis sentimen yang dipilih oleh penyelidik.

Kajian akan datang juga mengharapkan kajian Analisis Sentimen untuk dokumen BM dilaksanakan di dalam pelbagai domain dan bidang seperti produk, servis dan perkhidmatan. Pembangunan sistem penterjemahan pelbagai bahasa dialek ke bentuk BM formal di dalam talian juga amat diperlukan untuk memudahkan dan mempercepatkan proses penormalan semasa fasa pra-pemprosesan kerana proses ini memerlukan masa yang lama untuk meneliti setiap perkataan dan ayat yang dinyatakan oleh pengguna.

6.0 RUJUKAN

Anjaria, M., Mahana, R., & Guddeti, R. (2014). Influence Factor Based Opinion Mining of Twitter Data Using Supervised Learning. In *IEEE International Conference on Data Mining Workshops*.

Basri, S. B., Alfred, R., & On, C. K. (2012). Automatic spell checker for Malay blog. *2012 IEEE International Conference on Control System, Computing and Engineering*, 506–510. https://doi.org/10.1109/ICCSCE.2012.6487198

Dewan Bahasa dan Pustaka Malaysia. (2008). In *Laman Web Pusat Rujukan Persuratan Melayu (PRPM) @ DBP*. Retrieved from http://prpm.dbp.gov.my/

Duwairi, R., & El-orfali, M. (2013). A Study of the Effects of Preprocessing Strategies on Sentiment Analysis for Arabic Text. *Journal of Information Science*, 1–14. https://doi.org/10.1177/0165551510000000

Fatimah Ahmad, Mohammed Yusoff, & Tengku M. T. Sembok. (1996, December). Experiments with a stemming algorithm for Malay words. https://doi.org/10.1002/(SICI)1097-4571(199612)47:12<909::AID-ASI4>3.0.CO;2-6

Hidayatullah, A. F., & Ma'arif, M, R. (2016). International Conference on Recent Trends in Physics 2016 (ICRTP2016). *Journal of Physics: Conference Series 801*, *755*, 11001. https://doi.org/10.1088/1742-6596/755/1/011001

Hu, G., & Zhao, Q. (2010). Study to Eliminating Noisy Information in Web Pages based on Data Mining. In *Sixth International Conference on Natural Computation* (pp. 660–663).

Liu, B. (2010a). Sentiment Analysis : A Multi-Faceted Problem. In *IEEE Intelligent System*.

Liu, B. (2010b). Sentiment Analysis and Subjectivity. In *Handbook of Natural Language Processing* (pp. 1–38).

M. Arif, S., & Mustapha, M. (2017). The Effect of Noise Elimination and Stemming in Sentiment Analysis for Malay Documents. *Proceedings of the International Conference on Computing, Mathematics and Statistics (iCMS 2015)*, (iCMS), 67–73. https://doi.org/10.1007/978-981-10-2772-7

Mazidah Puteh, Norulhidayah Isa, Sayani Puteh, & Nur Amalina Redzuan. (2013). Sentiment Mining of Malay Newspaper (SAMNews) Using Artificial Immune System. *Proceedings of the World Congress on Engineering*, *III*.

Nor Azman Mat Ariff, & Nazlia Omar. (2008). Pengelasan E-mel Menggunakan Kaedah Perambat Balik. *Jurnal Teknologi Maklumat & Multimedia*, *5*, 91–106.

Norlela Samsudin, Abdul Razak Hamdan, Mazidah Puteh, & Mohd Zakree Ahmad Nazri. (2013). Mining Opinion in Online Messages. *International Journal of Advanced Computer Science and Applications*, *4*(8), 19–24.

Norlela Samsudin, Mazidah Puteh, & Abdul Razak Hamdan. (2011). bess or xbest : Mining the Malaysian Online Reviews. In *Conference on Data Mining and Optimization (DMO)* (pp. 28–29).

Norlela Samsudin, Mazidah Puteh, Ahmad Nazmi Fadzal, & Tajuddin, M. T. H. M. (2013). Dynamic Normalization of Microtexts Using SPMAT. In *CREAM-Current Research in Malaysia* (Vol. 2, pp. 101–113).

Norulhidayah Isa, Mazidah Puteh, & Raja Mohamad Hafiz Raja Kamarudin. (2013). Sentiment Classification of Malay Newspaper Using Immune Network (SCIN). In *Proceedings of the World Congress on Engineering* (Vol. III).

Othman. (1993). *Pengakar Perkataan Melayu untuk Sistem Capaian Dokumen.* Universiti kebangsaan Malaysia.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Pustaka, D. B. dan. (2008). *Panduan Singkatan Khidmat Pesanan Ringkas (SMS) Bahasa Melayu*. (Zaidi Ismail, Ed.).

Rayner, A., Leong, C. L., On, C. K., & Patricia, A. (2014). A Literature Review and Discussion of Malay Rule - Based Affix Elimination Algorithms. In *The 8th International Conference on Knowledge Management in Organizations* (pp. 579–591). https://doi.org/10.1007/978-94-007-7287-8

Roy, S., Dhar, S., Bhattacharjee, S., & Das, A. (2013). A Lexicon based Algorithm for Noisy Text Normalization as Pre-Processing for Sentiment Analysis, 2319–2322.

Saloot, M. A., Idris, N., & Aw, A. (2014). Noisy Text Normalization Using an Enhanced Language Model, 111–122.

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

Samarakoon, U., Regier, A., Tan, A., Desany, B. a, Collins, B., Tan, J. C., … Ferdig, M. T. (2011). High-throughput 454 resequencing for allele discovery and recombination mapping in Plasmodium falciparum. *BMC Genomics*, *12*(1), 116. https://doi.org/10.1186/1471-2164-12-116

Santos, F. L. Dos, & Ladeira, M. (2014). The Role of Text Pre-processing in Opinion Mining on a Social Media Language Dataset. *2014 Brazilian Conference on Intelligent Systems*, 50–54. https://doi.org/10.1109/BRACIS.2014.20

Supyan Hussin. (2010). Pertumbuhan dan Perkembangan Bahasa dalam Komunikasi Digital : Cabaran dan Harapan. In *Kolokium Penyelididkan Siswazah, Bahasa, Komunikasi dan Penyelidikan* (pp. 1–11).

Yussupova, N., Bogdanova, D., & Boyko, M. (2012). Applying of Sentiment Analysis for Texts in Russian Based on Machine Learning Approach. In *IMMM 2012: The Second International Conference on Advances in Information Mining and Management* (pp. 8–14).