https://doi.org/10.17576/apjitm-2017-0601-10

http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 6 No. 1, June 2017: 127 - 141
e-ISSN: 2289-2192

GENERATING A MALAY SENTIMENT LEXICON BASED ON WORDNET

NUR SHARMINI ALEXANDER
NAZLIA OMAR

ABSTRACT

Sentiment lexicon is a list of vocabularies that consists of positive and negative words. In opinion mining, sentiment lexicon is one of the important source in text polarity classification task in sentiment analysis model. Studies in Malay sentiment analysis is increasing since the volume of sentiment data is growing on social media. Therefore, requirement in Malay sentiment lexicon is high. However, Malay sentiment lexicon development is a difficult task due to the scarcity of Malay language resource. Thus, various approaches and techniques are used to generate sentiment lexicon. The objective of this paper is to develop Malay sentiment lexicon generation algorithm based on WordNet. In this study, the method is to map the WordNet Bahasa with English WordNet to get the offset value of a seed set of sentiment words. The seed set is used to generate the synonym and antonym semantic relation in English WordNet. The highest result achives 86.58% agreement with human annotators and 91.31% F1-measure in word polarity classification. The result shows the effectiveness of the proposed algorithm to generate Malay sentiment lexicon based on WordNet.

Keywords: Sentiment Lexicon, Sentiment Dictionary, Sentiment Analysis, Opinion Mining

PENJANAAN LEKSIKON SENTIMEN DALAM BAHASA MELAYU BERASASKAN WORDNET

ABSTRAK

Leksikon sentimen merupakan perbendaharaan kata yang mempunyai polariti sama ada positif atau negatif. Dalam pelombongan pendapat leksikon sentimen adalah sumber penting dalam tugasan analisis sentimen bagi menentukan polariti pendapat. Kajian analisis sentimen dalam Bahasa Melayu (BM) semakin giat dijalankan ekoran data sentimen yang semakin banyak di media sosial. Oleh yang demikian, pembangunan leksikon sentimen dalam BM adalah perlu. Namun begitu, adalah sukar untuk membangunkan leksikon sentimen dalam BM kerana sumber digital yang diperlukan adalah terbatas. Justeru pelbagai pendekatan dan kaedah yang digunakan untuk menjana leksikon sentimen. Matlamat kajian ini ialah membangunkan algoritma bagi menjana leksikon sentimen dalam BM berasaskan WordNet. Dalam kajian ini, pendekatan yang digunakan adalah dengan memadankan WordNet Bahasa dengan WordNet Bahasa Inggeris bagi mendapatkan nilai ofset set perkataan awal. Penjanaan ke atas set perkataan awal dilakukan melalui perhubungan semantik sinonim dan antonim yang terdapat dalam WordNet Bahasa Inggeris. Peratusan persetujuan yang diberikan oleh penutur BM yang tertinggi ialah sebanyak 86.58%. Manakala peratusan ukuran-F1 bagi pengujian pengelasan polariti perkataan yang tertinggi adalah 91.31%. Keputusan pengujian telah menunjukkan keberkesanan algoritma yang dicadangkan dalam penjanaan leksikon sentimen dalam Bahasa Melayu berasaskan WordNet.

Kata Kunci: Leksikon Sentimen, Kamus Sentimen, Analisis Sentimen, Pelombongan Pendapat

PENGENALAN

Pengguna media sosial menggunakan platform seperti *Facebook*, *Twitter* dan *blog* untuk mengutarakan pendapat dan mendapatkan maklumat berkenaan produk, perkhidmatan atau topik semasa. Perkongsian pendapat adalah maklumat yang penting bagi organisasi untuk membuat analisis sentimen berkenaan produk atau perkhidmatan yang disediakan (Kouloumpis et al. 2011). Pendapat tersebut dianalisa bagi menentukan penilaian pengguna terhadap topik yang dibincangkan sama ada sentimen adalah positif atau negatif (Padmaja & Sameen Fatima, 2013). Dalam proses analisis sentimen, leksikon sentimen adalah sumber utama bagi menentukan polariti perkataan sama ada positif atau negatif dalam dokumen (Pang & Lee, 2008). Dengan itu, pendapat pengguna dan leksikon sentimen merupakan sumber utama dalam analisis sentimen bagi merancang strategi untuk meningkatkan mutu produk dan perkhidmatan (Liu, 2012).

Pada awalnya kebanyakan kajian pembangunan leksikon sentimen dijalankan dalam Bahasa Inggeris. Manakala bahasa lain seperti Bahasa Perancis (Rao & Ravichandran, 2009), Bahasa Sepanyol (Perez-Rosas et al., 2012), Bahasa Jepun (Kaji & Kitsuregawa, 2007), Bahasa Hindi (Bakliwal et al. 2012; Rao & Ravichandran 2009) dan Bahasa Indonesia (Franky et al. 2015; Vania et al. 2014) adalah sedikit. Bilangan kajian dalam bahasa lain selain Bahasa Inggeris adalah kecil kerana kekurangan sumber untuk membangunkan leksikon dalam bahasa tersebut.

Analisis sentimen semakin giat dijalankan (Ahmed & Nazlia 2015; Chen et al. 2014; Vania et al. 2014) ekoran data sentimen yang semakin banyak di media sosial. Namun kajian dijalankan terhadap ulasan Bahasa Melayu (Al-Moslmi et al. 2015; Norulhidayah et al. 2013; Nurul Fathiyah et al. 2015) adalah sedikit. Sehingga kini kajian analisis sentimen terhadap ulasan dalam Bahasa Melayu yang dijalankan menggunakan pendekatan pembelajaran mesin kaedah penyelia (Al-Moslmi et al. 2015; Al-Saffar 2015; Alshalabi et al. 2013; Norlela, Mazidah & Abdul Razak, 2011), tanpa penyelia (Nurul Fathiyah et al. 2015, 2016; Yee Liau & Pei Tan, 2014) dan gabungan kedua-dua kaedah tersebut (Ahmed & Nazlia 2015).

Bilangan yang fokus kepada penjanaan leksikon sentimen dalam Bahasa Melayu adalah kecil manakala kajian yang lain adalah fokus kepada model analisis sentimen (Nurul Fathiyah et al. 2015; Yee Liau & Pei Tan 2014). Oleh yang demikian, pembangunan leksikon sentimen dalam Bahasa Melayu adalah perlu memandangkan kajian model analisis sentimen dalam Bahasa Melayu semakin giat dijalankan (Nurul Fathiyah et al. 2015; Sonai et al. 2017). Namun begitu, sumber digital yang diperlukan untuk membangun leksikon sentimen dalam Bahasa Melayu adalah terbatas (Nurril Hirfana et al., 2011) kerana Bahasa Melayu adalah tergolong dalam kategori sumber bahasa yang terhad. Oleh yang demikian, teknik yang sesuai diperlukan dalam pembangunan leksikon sentimen dalam Bahasa Melayu. Justeru, kajian ini bertujuan membangunkan algoritma bagi penjanaan leksikon sentimen dalam Bahasa Melayu berasaskan WordNet.

LEKSIKON SENTIMEN

Leksikon sentimen merupakan perbendaharaan kata yang mempunyai polariti sama ada positif atau negatif. Perkataan sentimen positif menggambarkan keadaan yang diingini iaitu polariti yang positif sebagai contoh “baik, bagus, suka”. Manakala perkataan sentimen negatif menggambarkan keadaan yang sebaliknya iaitu keadaan yang tidak diingini atau polariti yang tidak disukai sebagai contoh “jahat, buruk, teruk”(Liu, 2012).

Leksikon sentimen adalah sumber penting dalam tugasan analisis sentimen (Vania et al. 2014) bagi menentukan polariti pendapat dalam model analisis sentimen. Dengan itu, kajian dalam model analisis sentimen yang semakin giat dijalankan telah mendesak keperluan dalam pembangunan leksikon sentimen.

Pendekatan pembangunan leksikon sentimen adalah penting dalam menentukan haluan polariti dokumen (positif dan negatif) dan kekuatan sentimen terhadap topik-topik yang dibincangkan (Neviarouskaya et al., 2009). Pendekatan berasaskan leksikon adalah teknik berorientasikan semantik iaitu kaedah tanpa penyelia dalam pendekatan analisis sentimen. Teknik ini menggunakan peraturan leksikal dalam pengelasan sentimen iaitu pengukuran jarak perkaitan antara istilah kata sifat perkataan dalam menentukan sentimen (Kamps et al., 2004).

PENDEKATAN PENJANAAN LEKSIKON

Terdapat tiga pendekatan utama dalam pembangunan leksikon sentimen iaitu pendekatan manual, pendekatan berasaskan korpus dan pendekatan berasaskan kamus. Kebanyakan kajian dilakukan menggunakan pendekatan automatik iaitu berasaskan kamus (Andreevskaia & Bergler, 2006; Hu & Liu 2004; Kim & Hovy, 2004) dan pendekatan berasaskan korpus (Turney, 2002) memandangkan pendekatan manual memerlukan masa dan tenaga yang banyak (Liu, 2012). Selain itu, terdapat juga kajian yang menggabungkan kedua-dua pendekatan berasaskan korpus dan kamus (Franky et al. 2015).

Pendekatan penjanaan leksikon sentimen berasaskan kamus adalah set perkataan yang berupa pendapat dipadankan dengan kamus elektronik atau pangkalan data leksikal (Liu 2010). Kamus elektronik yang digunakan adalah seperti kamus dwi-bahasa manakala pangkalan data leksikal adalah seperti WordNet, SentiWordNet, Leksikon Pendapat Bing Liu, Leksikon *Havard General Inquirer*, Leksikon MPQA (*Multi-Perspective Question Answering)* dan *OpinionFinder*. Terdapat beberapa kaedah digunakan untuk membangun leksikon sentimen melalui pendekatan berasaskan kamus iaitu seperti penjanaan sinonim dan antonim meggunakan set perkataan awal berasaskan WordNet, penterjemahan ke bahasa lain melalui penggunaan SentiWordNet dan analisis semantik berasaskan kamus dalam talian.

Pangkalan data leksikal WordNet mempunyai ciri-ciri hubungan semantik antara perkataan seperti hipernim, hiponim, sinonim dan antonim. Kelebihan WordNet telah dimanfaatkan dalam kajian sebelum ini. Kamps et al. (2004) mengenal pasti sentimen bagi kata adjektif menggunakan WordNet melalui kaedah ukuran persamaan semantik berdasarkan jarak antara calon perkataan dan polariti perkataan sebagai contoh '*baik'* dan '*jahat'*. Manakala Hu dan Liu (2004) pula menjana perkataan yang berupa sentimen melalui algoritma bustrap (*bootstrap*) dengan memanfaatkan kelebihan WordNet yang mempunyai ciri-ciri hirarki dan hubungan sinonim dan antonim. Kim & Hovy (2004) membangunkan dua senarai perkataan awal yang mengandungi kata kerja dan kata adjektif yang berpolariti positif dan negatif. Proses penyarian bagi perkataan sinonim dan antonim dalam WordNet dilakukan bagi mengembangkan senarai perkataan yang telah dibangunkan dan diikuti dengan pengiraan kekuatan sentimen positif dan negatif bagi setiap perkataan tersebut. Pendekatan algoritma *Markov random walk* digunakan oleh Hassan & Radev (2010) untuk mengenal pasti polariti perkataan. Dalam kajian ini, WordNet digunakan sebagai sumber utama dalam penjanaan leksikon sentimen dalam Bahasa Arab dan Bahasa Inggeris.

Kaedah aruhan (*induction*) bagi mengembangkan polariti leksikal iaitu menggunakan graf WordNet telah diperkenalkan oleh Rao dan Ravichandran (2009). Hubungan graf yang digunakan adalah untuk meluaskan pengelasan polariti ke perkataan lain menggunakan graf berasaskan

algoritma pembelajaran semi penyelia seperti *mincuts*, *mincuts* terawak dan label rambatan. Kaedah algoritm label rambatan dalam Hindi WordNet digunakan bagi menjana leksikon sentimen dalam Bahasa Hindi. Manakala bagi penjanaan leksikon sentimen dalam Bahasa Perancis pula, kaedah kamus *French OpenOffice* digunakan.

Banea, Mihalcea & Wiebe (2008) membangunkan leksikon sentimen bagi Bahasa Romania berdasarkan kamus dalam talian dan koleksi dokumen-dokumen. Set kata subjektif dijadikan sebagai set perkataan awal digunakan sebagai pertanyaan melalui kamus dalam talian. Senarai perkataan yang dihasilkan daripada pertanyaan tersebut disaring melalui pengiraan persamaan dengan set kata subjektif sebelum ini menggunakan Analisis Semantik *Latent* (*Latent Semantic Analysis*).

Pendekatan berasaskan kamus adalah lebih berkesan dan mengandungi hampir kesemua perkataan kerana bergantung kepada koleksi perbendaharaan sedia ada yang telah piawai. Di samping itu pendekatan ini sesuai untuk menghasilkan leksikon sentimen pelbagai domain iaitu tidak bergantung kepada domain tertentu (Sonai et al. 2017). Namun Taboada et al. (2011) menyatakan bahawa pendekatan ini kurang menghasilkan asas yang kukuh dalam menjana leksikon yang lebih tepat berbanding dengan pendekatan secara manual.

LEKSIKON SENTIMEN DALAM BAHASA MELAYU

Kajian analisis sentimen dalam Bahasa Melayu semakin berkembang namun leksikon sentimen yang digunakan sebagai sumber utama masih terhad. Saloot, Norisma & Rohana (2014) membangunkan leksikon sentimen sebagai sumber kepada analisis sentimen dengan menggunakan kamus elektronik dan terjemah ke Bahasa Melayu. WordNet Bahasa hanya digunakan sebagai rujukan bagi perkataan yang tidak kelihatan dan polariti perkataan ditentukan secara manual. Bagi Yee Liau & Pei Tan (2014) pula, penjanaan leksikon sentimen dilakukan secara manual dan telah menghasilkan sebanyak 54 kata positif, 109 kata negatif dan 181 kata henti. Manakala polariti perkataan dalam leksikon tersebut ditentukan berdasarkan sumber leksikon sentimen SentiStrength. Nurul Fathiyah et al. (2015) pula membangunkan lesksikon sentimen yang dikenali sebagai kamus skor menggunakan WordNet 3.0 dan SentiWordNet sebagai rujukan polariti dan skor. Sehingga kini, kajian leksikon sentimen yang fokus kepada perbendaharaan kata berpolariti secara automatik dalam Bahasa Melayu hanya sedikit iaitu Sonai et al. (2017), Nasharuddin et al. (2017) dan Darwich et al. (2016).

Darwich et al. (2016) membangunkan leksikon sentimen dalam Bahasa Melayu dengan menggunakan kaedah penjanaan perkataan sinonim dan antonim berasaskan WordNet 3.0. Proses dimulakan dengan pemilihan dua set perkataan awal iaitu set positif dan set negatif. Setiap set mengandungi lima perkataan. Setelah itu proses memadankan perkataan yang terdapat dalam WordNet Bahasa ke dalam WordNet 3.0 dilakukan melalui nilai ofset. Pemadanan ini dilakukan untuk mendapatkan maksud perkataan yang sama dalam Bahasa Inggeris. Dalam kajian ini WordNet 3.0 digunakan untuk menjana perkataan sinonim dan antonim melalui kaedah rambatan. Jumlah set positif yang dihasilkan adalah 2139 lema melalui hubungan sinonim iaitu sebanyak 1858 lema dan hubungan antonim sebanyak 281 lema. Manakala set negatif yang dihasilkan adalah 2281 kata melalui hubungan sinonim iaitu sebanyak 1974 lema dan hubungan antonim sebanyak 307 lema. Leksikon sentimen yang dibangunkan dinilai dengan membuat perbandingan dengan General Inquirer. Keputusan ukuran kejituan bagi keseluruhan pada lelaran 4 adalah 0.643 dengan ukuran ketepatan polariti perkataan positif ialah 0.546 dan perkataan negatif ialah 0.727. Manakala ukuran kejituan bagi keseluruhan pada lelaran 5 adalah 0.567 dengan ukuran ketepatan

polariti perkataan positif ialah 0.482 dan perkataan negatif ialah 0.644. Keputusan bagi pengelasan polariti pula mencapai kejituan sebanyak 0.720. Namun begitu, oleh kerana leksikon sentimen yang dihasilkan adalah dalam Bahasa Melayu, perbandingan yang dibuat menggunakan sumber dalam Bahasa Inggeris adalah kurang sesuai. Ini adalah kerana perkataan yang terdapat dalam Bahasa Melayu kadang kala tidak sama maksudnya dalam Bahasa Inggeris. Justeru, maksud yang berbeza akan memberi kesan kepada polariti perkataan dalam leksikon sentimen.

Sonai et al.(2017) turut menjalankan kajian pembangunan leksikon sentimen dalam Bahasa Melayu iaitu Malay Lexicon (Mlex) menggunakan pendekatan berasaskan kamus. Teknik yang digunakan adalah penjanaan perkataan sinonim berasaskan pangkalan data Malay synset. Pangkalan data tersebut dibangunkan sebelum ini dengan mengumpulkan set perkataan sinonim daripada pelbagai sumber yang mana sumber utama ialah WordNet. Set perkataan awal dipilih adalah kata kerja dan kata adjektif yang berupa pendapat. Perkataan diekstrak daripada ulasan Bahasa Melayu yang terdapat di Twitter. Perkataan yang terdapat dalam set tersebut disemak terlebih dahulu sama ada telah ada dalam Mlex. Sekiranya tiada, penjanaan sinonim dijalankan. Hasil penjanaan perkataan sinonim akan menentukan polariti perkataan tersebut. Sekiranya terdapat perkataan sinonim yang mempunyai polariti yang sama, perkataan tersebut akan diberikan polariti yang sama. Namun sekiranya terdapat banyak perkataan sinonim yang dijumpai dan mengandungi polariti yang berlainan, pengiraan berdasarkan pemberat usia dilakukan bagi menentukan skor polariti perkataan tersebut. Penilaian eksperimen dilakukan terhadap penjanaan leksikon secara automatik menunjukkan pendekatan dan kaedah yang digunakan adalah baik dengan keputusan ketepatan sebanyak 86%. Keputusan yang baik ini turut disokong kaedah pengiraan berdasarkan pemberat usia dilakukan bagi menentukan skor polariti perkataan. Namun bagi perkataan yang tiada dalam pangkalan data Malay *synset* dan Mlex, perkataan tersebut dikategorikan sebagai perkataan tidak diketahui. Perancangan berkenaan perkara ini tidak dinyatakan bagi menyelesaikan masalah perkataan yang tidak diketahui.

Nasharuddin et al. (2017) membangunkan leksikon sentimen dalam Bahasa Melayu dengan fokus kepada penterjemahan silang bahasa secara automatik bagi Bahasa Melayu dan Bahasa Inggeris. Sumber yang digunakan ialah WordNet Bahasa bagi Bahasa Melayu dan SentiWordNet bagi Bahasa Inggeris, memandangkan penterjemahan silang bahasa dilakukan secara automatik iaitu melalui nilai ofset yang sama. Data yang digunakan adalah terdiri daripada 883 artikel berita dalam Bahasa Melayu dan Bahasa Inggeris daripada Bernama. Artikel ini mengandungi pelbagai domain iaitu politik, ekonomi, laporan eksekutif dan sukan. Langkah dimulakan dengan pra-pemprosesan dokumen iaitu penyingkiran simbol dan kata henti serta proses kata dasar dan pelabelan golongan kata. Perkataan yang berupa sentimen dijana bagi mendapatkan senarai perkataan sinonim dalam WordNet Bahasa. Setelah itu dengan menggunakan nilai ofset dan golongan kata yang sama dalam SentiWordNet, polariti dan skor perkataan diberikan. Penilaian eksperimen dilakukan terhadap penjanaan leksikon sentimen secara automatik menunjukkan pendekatan dan kaedah yang digunakan adalah kurang memuaskan. Secara keseluruhannya keputusan ketepatan adalah sebanyak 34%. Keputusan ini dipengaruhi oleh proses di peringkat pra-pemprosesan yang kurang mantap dan pemberian skor polariti perkataan yang sama bagi konteks yang berlainan. Berdasarkan keputusan tersebut, skor polariti yang diambil daripada SentiWordNet adalah kaedah yang kurang tepat. Ini telah pun dibuktikan oleh Das & Bandyopadhyay (2010) di mana skor polariti Bengali SentiWordNet adalah 47.6%. Sehubungan itu, penterjemahan dilakukan akan memberikan kesan terhadap polariti perkataan. Ini adalah kerana makna perkataan yang diterjemah kadangkala berbeza. Ini bererti polariti perkataan perlu disemak bagi menentukan skor dan polariti perkataan adalah boleh dipercayai.

PEMBANGUNAN LEKSIKON SENTIMEN DALAM BAHASA MELAYU

Terdapat empat aktiviti utama dalam proses pembangunan leksikon sentimen berasaskan WordNet iaitu:

1. Pemilihan set perkataan awal,
2. Pemadanan set perkataan dengan WordNet Bahasa,
3. Penjanaan kata sinonim dan antonim di WordNet Bahasa Inggeris dan
4. Penterjemahan ke Bahasa Melayu dalam WordNet Bahasa.

PEMILIHAN SET PERKATAAN AWAL

Pemilihan set perkataan awal dilakukan dengan mengenal pasti lima perkataan adjektif berupa sentimen untuk dua set iaitu set positif dan set negatif. Set perkataan awal dipilih berdasarkan kekerapan perkataan tertinggi dalam korpus dengan mengambil kira bahawa tiada persamaan maksud antara perkataan tersebut dalam polariti atau set yang sama. Dalam kajian ini, korpus yang dipilih adalah ulasan daripada gabungan domain perkhidmatan hotel dan ulasan filem. Berdasarkan pemilihan set perkataan awal tersebut, terdapat lima perkataan yang dikenal pasti dalam setiap set. Senarai bagi set perkataan awal positif ialah S^+ = {selesa, terbaik, bersih, bagus, suka} manakala senarai bagi set perkataan awal negatif ialah S^- = {teruk, mahal, dasyat, buruk, karut}.

PEMADANAN SET PERKATAAN AWAL DENGAN WORDNET BAHASA MELAYU

Pangkalan data leksikal WordNet Bahasa Melayu (Bond et al., 2014) dibangun berdasarkan pangkalan data leksikal WordNet Bahasa Inggeris (Miller et al. 1990) dengan menggunakan nilai ofset yang sama. Pemadanan set perkataan awal dengan WordNet Bahasa Melayu dilakukan bagi mendapatkan nilai ofset yang terdapat dalam jadual di pangkalan data WordNet Bahasa Inggeris. Proses ini dimulakan dengan dua set perkataan awal iaitu set positif dan set negatif yang terpilih dipadankan dengan lema yang terdapat dalam pangkalan data WordNet Bahasa Melayu. RAJAH 1 adalah algoritma bagi padanan set perkataan awal dengan WordNet Bahasa Melayu bagi mendapatkan nilai ofset. Dengan ini setiap set perkataan awal bagi set positif dan set negatif mempunyai nilai ofset untuk digunakan dalam aktiviti seterusnya. Algoritma yang ditunjukkan pada RAJAH 1 menunjukkan bahawa, setiap perkataan sama ada kata_pos atau kata_neg dipadankan dengan jadual di dalam pangkalan data WordNet Bahasa Melayu bagi mendapatkan nilai ofset perkataan. Nilai ofset perkataan yang padan dengan kata_neg atau kata_pos akan dimasukkan ke dalam senarai ofset set positif atau set negatif iaitu sama ada ‘list_ofset_pos^i’ atau ‘list_ofset_neg^i’. Nilai i adalah ‘0’ memberikan maksud bahawa senarai adalah set perkataan awal yang mengandungi nilai ofset, contoh adalah seperti ‘list_ofset_pos^0’ atau ‘list_ofset_neg^0’.

0. Mula
1. kata_pos = S^+ {selesa, terbaik, bersih, bagus, suka} // set positif
2. kata_neg = S^- {teruk, mahal, dasyat, buruk, karut} // set negatif
3. wnb = data WordNet Bahasa
4. tlemma = tlemma dalam jadual wnb
5. nilai_ofset_pos = nilai ofset bagi kata pos
6. nilai_ofset_neg = nilai ofset bagi kata neg
7. list_ofset_pos = senarai nilai_ofset bagi kata pos
8. list_ofset_neg = senarai nilai_ofset bagi kata neg
9. nilai i adalah 0
10. Untuk setiap kata_pos
    padankan kata_pos dengan wnb.tlemma
    dapatkan nilai_ofset_pos bagi padanan wnb.tlemma
    Tamat untuk setiap kata_pos
    masukkan nilai_ofset_pos dalam senarai list_nilai_ofset_pos^i
11. Untuk setiap kat_neg
    padankan kata_neg dengan wnb.tlemma
    dapatkan nilai_ofset_neg bagi padanan wnb.tlemma
    Tamat untuk setiap kata_neg
    masukkan nilai_ofset_neg dalam senarai list_ofset_neg^i
12. Tamat

RAJAH 1 Algoritma padanan set perkataan awal dengan WNB bagi mendapatkan nilai ofset

PENJANAAN KATA SINONIM DAN ANTONIM

Proses penjanaan kata sinonim dan antonim dalam WordNet Bahasa Inggeris adalah bagi mengembangkan senarai perkataan dalam set perkataan positif dan set perkataan negatif. Oleh yang demikian, penjanaan kata sinonim dan antonim adalah proses untuk mengumpul senarai perkataan yang mempunyai makna yang sama (sinonim) dan mengumpul senarai perkataan yang mempunyai makna yang berlawanan (antonim) menggunakan kaedah rambatan.

Dalam pangkalan data leksikal WordNet Bahasa Inggeris, setiap lema yang mempunyai makna yang sama direkodkan dalam set sinonim yang sama melalui nilai ofset. Manakala bagi lema yang mempunyai kata berlawan direkodkan melalui kod hubungan antonim. Proses penjanaan kata sinonim dan antonim dilakukan menggunakan senarai nilai ofset yang mewakili set perkataan awal bagi set positif dan set negatif dengan menggunakan kaedah rambatan perkataan yang mempunyai hubungan semantik sinonim dan antonim dalam WordNet Bahasa Inggeris.

Kaedah rambatan yang dilakukan melalui hubungan semantik sinonim dan antonim memelihara sentimen polariti perkataan yang merupakan asas penting dalam penjanaan leksikon sentimen. Senarai perkataan dalam set positif dan set negatif semakin berkembang melalui kaedah rambatan yang digunakan bagi menjana kata sinonim dan antonim. Proses rambatan berlaku apabila penjanaan sinonim dilakukan ke atas satu perkataan, iaitu beberapa perkataan yang mempunyai makna yang sama disenaraikan sebagai sinonim. Setelah itu, setiap perkataan daripada senarai perkataan tersebut pula akan melalui proses yang sama sehingga ke lelaran yang kelima.

Penjanaan kata sinonim dan antonim dalam WordNet Bahasa Inggeris dilakukan menggunakan set perkataan awal positif iaitu $list\_nilai\_ofset\_pos^0$ = {selesa, terbaik, bersih, bagus, suka} manakala senarai bagi set perkataan awal negatif ialah $list\_nilai\_ofset\_neg^0$ = {teruk, mahal, dasyat, buruk, karut}. RAJAH 2 menunjukkan algoritma bagi penjanaan sinonim dan antonim menggunakan set perkataan awal positif diwakilkan $list\_nilai\_ofset\_pos^0$ dan set perkataan awal negatif adalah $list\_nilai\_ofset\_neg^0$ menggunakan kaedah rambatan untuk

mengembangkan senarai set perkataan positif dan negatif pada lelaran pertama. Hasil penjanaan kata sinonim bagi set perkataan positif disimpan dalam senarai set perkataan positif dan set perkataan negatif disimpan dalam senarai set perkataan negatif. Manakala hasil penjanaan kata antonim bagi perkataan positif disimpan dalam set senarai perkataan negatif dan hasil perkataan negatif disimpan dalam senarai perkatan positif. Set perkataan baru yang dihasilkan pada lelaran pertama iaitu *list_nilai_ofset_pos*^{1} bagi set positif dan *list_nilai_ofset_neg*^{1} bagi set negatif merupakan set perkataan yang dijana untuk lelaran kedua. Dengan itu set perkataan yang dihasilkan pada lelaran semasa akan digunakan sebagai set perkataan baru yang dijana pada lelaran yang seterusnya. Oleh yang demikian, setiap lelaran menggunakan set perkataan yang baru untuk dijana dengan menggunakan kaedah rambatan bagi menghasilkan senarai perkataan sinonim dan antonim.

0. Mula
1. list_nilai_ofset_pos^{i} = senarai nilai ofset bagi perkataan positif
2. nilai_ofset_pos = nilai ofset perkataan positif
3. list_nilai_ofset_neg^{i} = senarai nilai ofset bagi perkataan negatif
4. nilai_ofset_neg = nilai ofset perkataan negatif
5. wn = data WordNet
6. nilai i adalah 0
7. Jika i kurang daripada 5
    a. DapatkanSinonim(list_nilai_ofset_pos^{i})
        dapatkan wn.nilai_ofset_pos yang mempunyai *senses* sama
        masukkan nilai_ofset_pos ke dalam list_nilai_pos^{i+1}
    b. DapatkanAntonim(list_nilai_ofset_pos^{i})
        dapatkan wn.nilai_ofset_pos yang berlawan
            masukkan nilai_ofset_pos ke dalam list_nilai_neg^{i+1}
    c. DapatkanSinonim(list_nilai_ofset_neg^{i})
        dapatkan wn.nilai_ofset_neg yang mempunyai *senses* sama
        masukkan nilai_ofset_neg ke dalam list_nilai_neg^{i+1}
    d. DapatkanAntonim(list_nilai_ofset_neg^{i})
        dapatkan wn.nilai_ofset_neg yang berlawan
        masukkan nilai_ofset_neg ke dalam list_nilai_pos^{i+1}
    nilai i+1
7. Tamat Jika
8. Tamat

RAJAH 2 Algoritma bagi penjanaan sinonim dan antonim menggunakan set perkataan awal

PENTERJEMAHAN KE BAHASA MELAYU

Penjanaan sinonim dan antonim yang dilakukan dalam WordNet Bahasa Inggeris menghasilkan set perkataan bagi polariti positif dan negatif dalam Bahasa Inggeris. Oleh yang demikian, penterjemahan ke Bahasa Melayu perlu dilakukan melalui padanan nilai ofset WordNet Bahasa Inggeris dengan nilai ofset WordNet Bahasa Melayu. Berdasarkan RAJAH 3 nilai ofset yang terdapat dalam *list_nilai_ofset_pos*^{i} dan *list_nilai_ofset_neg*^{i} dipadankan dengan nilai ofset yang terdapat dalam pangkalan data WordNet Bahasa Melayu.

```
list_nilai_ofset_pos^i = senarai nilai ofset bagi perkataan positif
nilai_ofset_pos = nilai ofset perkataan positif
list_nilai_ofset_neg^i = senarai nilai ofset bagi perkataan negatif
nilai_ofset_neg = nilai ofset perkataan negatif

DapatkanKataBM(list_nilai_ofset_pos^i+1)
        Dapatkan wnb.tlemma bagi padanan wn.nilai_ofset_pos
        Masukkan wnb.nilai_ofsef_pos, wnb.tlema ke dalam list_pos^i+1

DapatkanKataBM(list_nilai_ofset_neg^i+1)
        Dapatkan wnb.tlemma bagi padanan wn.nilai_ofset_neg
        Masukkan wnb.nilai_ofsef_pos, wnb.tlema ke dalam list_ neg^i+1
```

RAJAH 3 Algoritma bagi penterjemahan perkataan ke Bahasa Melayu

PENILAIAN LEKSIKON SENTIMEN

Leksikon sentimen yang dibangunkan perlu dinilai bagi memastikan polariti perkataan yang dijana adalah tepat. Penilaian leksikon sentimen yang dibangunkan diterangkan seperti dibawah.

PENILAIAN KETEPATAN POLARITI PERKATAAN LEKSIKON SENTIMEN

Penilaian ini adalah untuk menguji ketepatan polariti perkataan leksikon sentimen yang telah dibangunkan berbanding piawai emas. Piawai emas merupakan kaedah penilaian yang dilakukan oleh penutur yang fasih dalam sesuatu bahasa ke atas leksikon sentimen yang dibangunkan dalam bahasa tersebut. Dalam kajian ini, penilaian diberikan oleh orang yang fasih bertutur dalam Bahasa Melayu iaitu pelajar Fakulti Teknologi dan Sains Maklumat. Penilai juga mempunyai pengetahuan tentang perkara yang berkaitan dengan analisis sentimen. Tujuan eksperimen ini dijalankan ialah untuk mendapatkan nilai peratusan persetujuan polariti perkataan leksikon sentimen daripada penilai yang fasih bertutur dalam Bahasa Melayu. Dalam penilaian ini, penilai akan menentukan polariti perkataan yang terdapat dalam leksikon sentimen yang dihasilkan sama ada betul atau salah bagi setiap lelaran penjanaan. Setiap perkataan yang dinilai betul adalah polariti perkataan yang dipersetujui oleh penilai. Nilai peratus persetujuan ini dikira dengan menjumlahkan semua bilangan polariti perkataan yang betul berdasarkan jumlah kesemua perkataan yang dijana.

PENILAIAN PENGELASAN POLARITI PERKATAAN

Penilaian ini adalah untuk menguji ketepatan pengelasan teks ulasan berdasarkan leksikon sentimen yang telah dibangunkan. Tujuan eksperimen ini dijalankan ialah untuk menguji kebolehpercayaan leksikon sentimen yang telah dibuat pengujian ketepatan polariti perkataan dengan mendapatkan nilai ketepatan keputusan pengelasan teks terhadap data ulasan. Kaedah pengelasan yang digunakan bagi penilaian dalam kajian ini adalah menggunakan teknik Naïve Bayes iaitu algoritma yang digunakan untuk mencari kebarangkalian tertinggi dalam pengelasan polariti perkataan. Teknik Naïve Bayes dipilih kerana hasil pengelasan bagi polariti perkataan yang dilakukan adalah lebih baik berbanding dengan teknik lain seperti Mesin Sokongan Vektor dan Entropi Maksimum (Pang et al., 2002). Seterusnya penilaian bagi ketepatan keputusan pengelasan

dilakukan bagi melihat prestasi leksikon sentimen yang digunakan sebagai sumber kepada aktiviti pengelasan polariti teks. Pengelasan polariti perkataan merupakan penilaian bagi mengelaskan dokumen ulasan sama ada mempunyai sentimen positif atau negatif dengan berasaskan leksikon sentimen yang telah dibangunkan. Contoh dokumen ulasan yang digunakan dalam eksperimen ini adalah seperti di dalam RAJAH 4 dibawah.

1. Harga yang baik dan penginapan yang terbaik. Kami berpuas hati dengan penginapan dan layanan serta lokasi hotel ini. Kami berjaya mendapatkan harga diskaun yang menguntungkan.
2. Penceritaan dari babak ke babak sangat menarik dan hebat. Filem yang sangat menarik dan terbaik.
3. Alza saya telah mengalami beberapa kerosakan, iaitu power window kiri belakang berbunyi pelik, motor wiper hadapan rosak dan juga kerosakan pada gearbox.
4. Pengalaman yang buruk berhadapan dengan pekerja yang biadap dan teruk. Pekerja yang langsung tidak membantu.

RAJAH 4 Contoh ulasan dalam penilaian pengelasan polariti perkataan

Dalam teknik Naïve Bayes, pengiraan kebarangkalian bagi dokumen bersentimen positif atau negatif dilakukan dengan mengira polariti bagi setiap teks dalam dokumen tersebut. Setiap dokumen mengandungi banyak teks dan setiap teks mempunyai nilai polariti tersendiri iaitu teks berpolariti positif bernilai 1 dan teks berpolariti negatif adalah bernilai -1. Pengiraan kebarangkalian kelas C* bagi dokumen d adalah seperti berikut.

$$C^* = \text{argmax } P(c|d) \quad (1)$$

Rumusan pengiraan kebarangkalian pengelasan Naïve Bayes adalah seperti berikut.

$$P(C_j|d_i) = \frac{P(C_j) \,.\, P(C_j|d_i)}{P(d_i)} \quad (2)$$

PERBINCANGAN

Leksikon sentimen dihasilkan melalui penjanaan perkataan sinonim dan antonim menggunakan kaedah rambatan sebanyak lima lelaran dalam WordNet Bahasa Inggeris. Penjanaan dimulakan dengan penggunaan set perkataan awal iaitu lima perkataan dalam set positif dan lima perkataan dalam set negatif. Bilangan lema yang dihasilkan pada setiap lelaran semakin meningkat dan jumlah lema yang terdapat dalam leksikon sentimen adalah sebanyak 14,337 lema. Pada lelaran yang pertama, jumlah lema yang dihasilkan adalah sebanyak 723 lema iaitu sebanyak 339 lema positif dan 384 lema negatif. Penjanaan perkataan sinonim bagi set positif ialah sebanyak 339 lema dan set negatif ialah 312 lema. Manakala penjanaan perkataan antonim pula ialah 17 lema bagi set positif dan 275 bagi set negatif. Secara keseluruhannya hasil penjanaan perkataan sinonim dan antonim adalah seperti di JADUAL 1.

JADUAL 1 Hasil penjanaan perkataan sinonim dan antonim mengikut lelaran

Bilangan Lelaran	Bilangan lema	Bilangan lema positif	Bilangan Lema negatif
1	723	339	384
2	1151	499	652
3	2836	1380	1456
4	4667	2293	2374
5	4960	2404	2556

KEPUTUSAN PENILAIAN KETEPATAN POLARITI PERKATAAN

Eksperimen ini memberi tumpuan ke atas peratus persetujuan ketepatan penilaian bagi polariti perkataan yang dibangunkan menggunakan pendekatan penjanaan perkataan sinonim dan antonim berasaskan WordNet. Hasil pengujian adalah seperti dalam JADUAL 2 menunjukkan persetujuan ketepatan penilaian bagi polariti perkataan dalam leksikon sentimen.

JADUAL 2 Persetujuan ketepatan polariti perkataan leksikon sentimen

No. Lelaran	Bil. Perkataan	Bil. Positif	Peratus Positif Betul	Bil. Negatif	Peratusan Negatif Betul	Peratusan Keseluruhan Betul
1	723	339	82.30	384	90.36	86.58
2	1151	499	60.12	652	85.58	74.54
3	2836	1380	49.71	1456	80.49	65.50
1-3	4710	2218	57.03	2492	83.35	70.96
4	4667	2293	38.51	2374	71.86	55.47
5	4960	2404	22.38	2556	68.23	46.01
Semua	14337	6915	38.84	7422	74.47	57.29

JADUAL 2 menunjukkan bahawa peratusan bagi persetujuan penilai dengan leksikon sentimen yang dijana secara automatik adalah semakin menurun. Dalam kajian Darwich et al. (2016) ketepatan keputusan juga menunjukkan penurunan secara keseluruhan dalam setiap lelaran. Penurunan ini adalah disebabkan nilai kesalinghubungan semantik antara perkataan semakin lemah dan jurang hubungan perkataan tersebut semakin besar. Jarak hubungan semantik atau kesalinghubungan semantik adalah pengukuran ke atas dua perkataan yang saling berkaitan (Siblini & Kosseim 2013). Nilai jarak kesalinghubungan semantik menentukan ketepatan hubungan antara perkataan (Budanitsky & Hirst 2006). Berdasarkan kaedah pemberat hubungan semantik berasaskan WordNet yang telah dibangunkan oleh Siblini & Kosseim (2013), nilai kesalinghubungan semantik antara lema contoh ‘berfaedah’ pada lelaran 1 dengan lema ‘mendatangkan manfaat’ pada lelaran 2 adalah 96%. Nilai ini menunjukkan bahawa nilai kesalinghubungan semantik antara dua lema ini adalah tinggi dalam hubungan sinonim. Manakala nilai kesalinghubungan semantik antara lema ‘berfaedah’ pada lelaran 1 dengan lema ‘tersusun’ pada lelaran 5 adalah 28%. Nilai ini menunjukkan bahawa hubungan semantik antara dua lema ini adalah lemah dan jurang hubungan perkataan sinonim bagi dua lema ini adalah besar. Oleh itu, nilai kesalinghubungan semantik perkataan antara lelaran yang semakin kecil menunjukkan perkaitan dengan penurunan peratusan persetujuan leksikon sentimen yang dijana secara automatik. Selain daripada itu, penjanaan perkataan sinonim yang berpolariti positif boleh juga

menghasilkan perkataan yang berpolariti negatif. Contoh hasil penjanaan sinonim bagi perkataan '*berbesar hati*' ialah '*membanggakan'*, '*bangga'*, '*angkuh*', '*bongkak*', '*megah*', '*meninggi*', '*riak*', '*sikap angkuh*' dan '*sombong*'. Perkataan polariti positif yang dihasilkan ialah '*bangga*' dan '*megah*' manakala polariti negatif ialah '*angkuh*', '*bongkak*', '*meninggi*', '*riak*', '*sikap angkuh*' dan '*sombong*'. Dengan itu, penjanaan perkataan sinonim tidak menjamin bahawa polariti perkataan yang dijana adalah sama turut mempengaruhi penurunan peratusan persetujuan penilaian.

KEPUTUSAN PENILAIAN PENGELASAN POLARITI PERKATAAN

Penilaian bagi pengelasan polariti perkataan adalah menggunakan senarai lema dalam leksikon sentimen yang dipersetujui oleh penilai seperti yang telah dilakukan pada pengujian ketepatan polariti perkataan. Data ulasan yang digunakan dalam pengujian ini adalah ulasan 1 yang mewakili domain perkhidmatan hotel dan filem daripada sumber Yelp dan Rotten Tomatoes dan ulasan 2 yang mewakili domain produk seperti kereta proton, telefon pintar dan komputer riba daripada sumber forum ProtonMalaysia dan blog teknologi. Pengujian dilakukan ke atas data positif dan data negatif yang setiap satunya mengandungi 100 ayat bagi kedua-dua domain. Hasil pengujian adalah seperti yang ditunjukkan dalam JADUAL 3.

JADUAL 3 Hasil pengujian pengelasan polariti perkataan

Leksikon sentimen yang dijana mengikut lelaran	Ulasan 1 Domain perkhidmatan hotel dan filem			Ulasan 2 Domain produk (kereta proton, telefon pintar dan komputer riba)		
	Kejituan	Dapatan semula	Skor-F1	Kejituan	Dapatan semula	Skor-F1
Lelaran 1-3	0.8539	0.76	0.8042	0.7879	0.52	0.6265
Lelaran 4	0.6573	0.94	0.7736	0.6043	0.84	0.703
Lelaran 5	0.6484	0.83	0.728	0.6481	0.7	0.6731
Semua lelaran	0.8065	0.75	0.7772	0.7733	0.58	0.6628

JADUAL 3 menunjukkan nilai kejituan, dapatan semula dan ukuran skor-F1 ulasan 1 lebih tinggi berbanding ulasan 2 bagi kesemua lelaran. Nilai kejituan yang tertinggi bagi kedua-dua ulasan adalah pada gabungan lelaran 1 hingga 3 iaitu masing-masing ialah 85.39% dan 78.79%. Hasil pengujian tidak menunjukkan perbezaan nilai kejituan yang ketara dalam setiap kategori lelaran. Namun begitu, pengujian ini menggunakan kaedah unigram dalam menentukan pengelasan setiap perkataan. Satu perkataan adakalanya tidak memberikan maksud yang sebenar dalam ulasan. Contoh ulasan '*peranti ini tidak bagus*' menunjukkan polariti perkataan adalah positif bagi perkataan '*bagus*' kerana leksikon sentimen memberikan skor '1'. Namun sekiranya kaedah bigram iaitu dua perkataan digunakan, ulasan tersebut menunjukkan polariti adalah negatif iaitu bagi perkataan '*tidak bagus*' dan leksikon sentimen memberikan skor '-1'. Dengan itu petua penggunaan kaedah pengelasan perlu dititikberatkan bagi mengenal pasti maksud ulasan tersebut kerana akan mempengaruhi nilai kejituan dalam pengelasan polariti perkataan. Nilai dapatan semula yang tertinggi bagi kedua-dua ulasan adalah pada lelaran 4. Keputusan bagi nilai dapatan semula untuk ulasan 1 ialah 94% manakala ulasan 2 ialah sebanyak 84% pada lelaran 4. Nilai dapatan semula yang tinggi adalah dipengaruhi oleh jumlah padanan perkataan yang banyak dalam

ulasan dengan lema yang dihasilkan pada lelaran 4. Keputusan tertinggi bagi ukuran skor-F1 pula menunjukkan perbezaan pada kedua-dua ulasan. Ulasan 1 mendapat nilai tertinggi pada gabungan lelaran 1 hingga 3 iaitu sebanyak 80.42% manakala nilai dapatan semula tertinggi bagi ulasan 2 adalah pada lelaran 4 iaitu sebanyak 70.30%.

Secara keseluruhannya hasil pengujian pengelasan polariti perkataan secara puratanya memberikan nilai tertinggi pada gabungan lelaran 1 hingga 3 bagi ulasan 1 tetapi purata nilai tertinggi bagi ulasan 2 pula adalah pada lelaran 4. Namun begitu, keputusan bagi eksperimen ini telah memberikan keputusan yang baik secara keseluruhannya.

KESIMPULAN

Kajian ini bertujuan untuk membangunkan algoritma bagi penjanaan leksikon sentimen dalam Bahasa Melayu berasaskan WordNet. Terdapat empat proses utama yang terlibat dalam membangunkan leksikon sentimen iaitu pemilihan set perkataan awal, pemadanan WordNet Bahasa dan WordNet Bahasa Inggeris melalui nilai ofset, penjanaan perkataan sinonim dan antonim dalam WordNet Bahasa Inggeris dan penterjemahan hasil penjanaan ke Bahasa Melayu dengan memadankan nilai ofset ke WordNet Bahasa. Bagi menilai ketepatan hasil kajian, dua jenis eksperimen dijalankan iaitu menguji ketepatan leksikon sentimen berbanding piawai emas iaitu penilaian secara manual dan menguji pengelasan polariti perkataan. Hasil pengujian menunjukkan bahawa, pendekatan yang digunakan dalam kajian untuk menyelesaikan masalah yang telah dinyatakan adalah baik. Berdasarkan keputusan tersebut maka penambahbaikan boleh diteruskan iaitu memperluaskan skop jenis golongan kata seperti kata kerja dan kata tugas yang berupa pendapat dan memantapkan penghasilan pembangunan algoritma dalam penyelidikan ke atas penjanan leksikon sentimen dalam Bahasa Melayu. Dalam pada itu bidang kajian semantik boleh diterapkan bersama dalam kajian khususnya kesalinghubungan antara perkataan dalam Bahasa Melayu.

RUJUKAN

Ahmed, A. & Nazlia, O. 2015. Integrating a lexicon based approach and K Nearest Neighbour for Malay sentiment analysis. *Journal of Computer Science 2015*, hlm. 639–644.

Al-Moslmi, T., Gaber, S., Al-Shabi, A., Albared, M. & Omar, N. 2015. Feature selection methods effects on machine learning approaches in Malay sentiment analysis. *1st ICRIL-International Conference on Innovation in Science and Technology (lICIST 2015)*, hlm. 2–5.

Al-Saffar, A.A.M. 2015. Malay sentiment classification based on machine learning and lexicon based approach. *Tesis S. T. M, Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia.*

Alshalabi, H., Tiun, S., Omar, N. & Albared, M. 2013. Experiments on the use of feature selection and machine learning methods in automatic Malay text categorization. *4th International Conference on Electrical Engineering and Informatics, ICEEI 2013*, Jil. 11, hlm. 748–754.

Andreevskaia, A. & Bergler, S. 2006. Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. *Proceedings of EACL* 6: 209–216.

Bakliwal, A., Arora, P. & Varma, V. 2012. Hindi subjective lexicon: A lexical resource for Hindi polarity classification. *The Eighth International Conference on Language Resources and Evaluation*(May): 1189–1196.

Banea, C., Mihalcea, R. & Wiebe, J. 2008. A Bootstrapping method for building subjectivity lexicons for languages with scarce resources. *Proceedings of the Language Resources Evaluation Conference (LREC)* 2764–2767.

Bond, F., Lim, L.T., Tang, E.K. & Hammam, R. 2014. The combined Wordnet Bahasa. *NUSA: Linguistic*

Studies of Languages in and around Indonesia 57 3–9.
Budanitsky, A. & Hirst, G. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics* 32(1): 13–47.
Chen, Y., Brook, S. & Skiena, S. 2014. Building sentiment lexicons for all major languages. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics Volume 2*, hlm. 383–389.
Darwich, M., Shahrul, A. & Nazlia, O. 2015. Inducing a domain-independent sentiment lexicon in Malay. *JAIST Symposium on Advanced Science and Technology*,
Darwich, M., Shahrul, A. & Nazlia, O. 2016. Automatically generating a sentiment lexicon for the Malay language. *Asia-Pacific Journal of Information Technology and Multimedia* 5(1): 49–59.
Das, A. & Bandyopadhyay, S. 2010. Towards the global SentiWordNet. *PACLIC 24 - Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, hlm. 799–808.
Franky, Bojar, O. & Veselovská, K. 2015. Resources for Indonesian sentiment analysis. *The Prague Bulletin of Mathematical Linguistics* 103(1): 21–41.
Hassan, A. & Radev, D. 2010. Identifying text polarity using random walks. *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, hlm. 395–403.
Hu, M. & Liu, B. 2004. Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD 04* 4: 168.
Kaji, N. & Kitsuregawa, M. 2007. Building lexicon for sentiment analysis from massive collection of HTML documents. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Jil. 43, hlm. 1075–1083. Association for Computational Linguistics.
Kamps, J., Marx, M., Mokken, R.J. & Rijke, M. de. 2004. Using WordNet to measure semantic orientations of adjectives. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, hlm. 1115–1118.
Kim, S.-M. & Hovy, E. 2004. Determining the sentiment of opinions. *Proceeding COLING '04 Proceedings of the 20th international conference on Computational Linguistics*, hlm. 1367.
Kouloumpis, E., Wilson, T. & Moore, J. 2011. Twitter sentiment analysis: The good the bad and the OMG! *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 11)* 538–541.
Liu, B. 2010. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, Second Edi., hlm. 1–38. Taylor and Francis Group, Boca.
Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* 5(1): 1–167.
Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K.J. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography* 3(4): 235–244.
Neviarouskaya, A., Prendinger, H. & Ishizuka, M. 2009. SentiFul: Generating a reliable lexicon for sentiment analysis. *Proceedings - 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009*,
Norlela, S., Mazidah, P. & Abdul Razak, H. 2011. Bess or xbest: Mining the Malaysian online reviews. *Conference on Data Mining and Optimization*, hlm. 38–43.
Norulhidayah, I., Mazidah, P. & Raja Mohamad Hafiz, R.K. 2013. Sentiment classification of Malay newspaper using immune network (SCIN). *Proceedings of the World Congress on Engineering 2013*, Jil. III, hlm. 1543–1548.
Nurril Hirfana, M.N., Suerya, S. & Bond, F. 2011. Creating the Open Wordnet Bahasa. *25th Pacific Asia Conference on Language, Information and Computation*, hlm. 255–264.
Nurul Amelina, N., Muhamad Taufik, A., Azreen, A. & Rabiah, A.K. 2017. English and Malay cross-lingual sentiment lexicon acquisition and analysis. *Information Science Application* 424(1): 467–475.
Nurul Fathiyah, S., Halizah, B. & Zurina, S. 2016. Lexical based sentiment analysis - Verb, adverb &

negation. *Journal of Telecommunication, Electronic and Computer Engineering* 8(2): 161–166.
Nurul Fathiyah, S., Halizah, B., Zurina, S., Ahmad Fadzli Nizam, A.R., Mohd Hafiz, Z. & Nurulhalim, H. 2015. Sentiment classification of unstructured data using lexical based techniques. *Journal Technology* 77: 1–5.
Padmaja, S. & Sameen Fatima, S. 2013. Opinion mining and sentiment analysis - An assessment of peoples' belief: A survey. *International Journal of Ad Hoc, Sensor & Ubiquitous Computing* 4(1): 21–33.
Pang, B. & Lee, L. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2(1–2): 1–135.
Pang, B., Lee, L. & Vaithyanathan, S. 2002. Thumbs up?: Sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, hlm. 79–86.
Perez-Rosas, V., Banea, C. & Mihalcea, R. 2012. Learning sentiment lexicons in Spanish. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, hlm. 3077–3081.
Rao, D. & Ravichandran, D. 2009. Semi-supervised polarity lexicon induction. *EACL '09 Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Jil. 30, hlm. 675–682.
Saloot, M.A., Norisma, I. & Rohana, M. 2014. An architecture for Malay Tweet normalization. *Information Processing and Management* 50(5): 621–633.
Siblini, R. & Kosseim, L. 2013. Using a weighted semantic network for lexical semantic relatedness. *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, hlm. 610–618.
Sonai, K., Anbananthen, M., Selvaraju, S. & Krishnan, J.K. 2017. The generation of Malay lexicon. *American Journal of Applied Sciences*
Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics* 37(2): 267–307.
Turney, P.D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*(July): 417–424.
Vania, C., Ibrahim, M. & Adriani, M. 2014. Sentiment lexicon generation for an under-resourced language. *International Journal of Computational Linguistics and Applications* 5(1): 59–72.
Yee Liau, B. & Pei Tan, P. 2014. Gaining customer knowledge in low cost airlines through text mining. *Industrial Management & Data Systems* 114(9): 1344–1359.

Nur Sharmini Alexander[1] dan Nazlia Omar[2]
Fakulti Teknologi dan Sains Maklumat (FTSM),
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor
MALAYSIA
[1]nursharmini@gmail.com; [2]nazlia@ukm.edu.my

Received: 9 January 2017
Accepted: 22 March 2017
Accepted: 25 June 2017