PENJANAAN LEKSIKON SENTIMEN DALAM BAHASA MELAYU BERASASKAN WORDNET

NUR SHARMINI ALEXANDER

DISERTASI YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN DARIPADA SYARAT MEMPEROLEHI IJAZAH SARJANA TEKNOLOGI MAKLUMAT (SAINS MAKLUMAT)

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2017

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang setiap satunya telah saya jelaskan sumbernya.

7 Ogos 2017

NUR SHARMINI ALEXANDER
P83886

PENGHARGAAN

Alhamdulillah, bersyukur kehadrat Ilahi kerana dengan izin kurniaNya saya dianugerahkan kesihatan yang baik, kematangan fikiran dan kemudahan serta nikmat masa yang cukup untuk saya menyiapkan kajian ini. Setinggi-tinggi penghargaan saya ucapkan kepada penyelia Prof. Madya Dr. Nazlia Omar atas bimbingan, kepakaran, bantuan, teguran dan nasihat yang amat berguna serta semangat yang diberikan di sepanjang pengajian terutamanya dalam menyiapkan kajian ini.

Saya ingin merakamkan ucapan terima kasih yang tidak terhingga kepada suami tercinta, Yusri bin Yusoff di atas segala pengorbanan, kesabaran dan dorongan yang membawa saya mengharungi segala cabaran di sepanjang tempoh pengajian. Begitu juga dengan sumber inspirasi saya di sepanjang pengajian ini iaitu kelima-lima cahaya mata kesayangan Nur Aina, Danish, Nur Alia, Danial Hakim dan Nur Alisha yang amat sabar dan memahami situasi saya sebagai ibu dan juga pelajar di sepanjang pengajian. Semoga tesis ini menjadi pendorong kepada mereka dalam usaha mencari ilmu untuk mencapai kejayaan.

Saya juga ingin mengucapkan terima kasih kepada ahli keluarga saya yang telah banyak memberi semangat dan dorongan dalam tempoh pengajian saya ini. Selain itu, saya turut mengucapkan terima kasih kepada rakan sekuliah dan rakan sekerja atas kerjasama dan tunjuk ajar dalam menyempurnakan tugasan ini dengan jayanya. Terima kasih juga kepada rakan seperjuangan di bawah penyelia yang sama atas kerjasama yang diberikan dalam merealisasikan kajian ini.

Ucapan ini juga ditujukan kepada semua pihak yang terlibat sama ada secara langsung atau tidak langsung dalam menyiapkan tugasan ini. Segala sokongan dan dorongan yang diberikan amatlah dihargai.

ABSTRAK

Leksikon sentimen merupakan perbendaharaan kata yang mengandungi perkataan sama ada berunsur positif atau negatif. Dalam pelombongan pendapat, leksikon sentimen merupakan sumber utama yang digunakan dalam pengelasan polariti unit teks bagi menentukan sentimen sesebuah dokumen pendapat. Kajian model analisis sentimen dalam Bahasa Melayu (BM) semakin giat dijalankan. Dengan itu keperluan bagi sumber leksikon sentimen BM adalah tinggi. Namun pembangunan leksikon sentimen dalam BM merupakan proses yang sukar dan rumit. Ini adalah kerana sumber bahasa bagi keperluan pembangunan leksikon adalah terhad. Justeru pelbagai pendekatan dan kaedah yang digunakan untuk menjana leksikon sentimen. Selain daripada itu, kajian sedia ada tidak menjurus kepada pembangunan sentimen leksikon itu sendiri terutamanya dalam BM. Matlamat kajian ini ialah membangunkan algoritma bagi menjana leksikon sentimen dalam BM berasaskan WordNet. Seterusnya dengan menggunakan algoritma ini, leksikon sentimen dapat dihasilkan dan digunakan untuk menganalisa sentimen dalam BM. Secara umum, proses penjanaan sentimen leksikon dimulakan dengan pemilihan set perkataan awal positif dan negatif. Set perkataan awal yang dipilih seterusnya dikenal pasti dalam WordNet Bahasa yang mana dipadankan dengan WordNet Bahasa Inggeris melalui nilai ofset yang sama. Setelah itu penjanaan dilakukan dengan melalui perhubungan semantik sinonim dan antonim yang terdapat dalam WordNet Bahasa Inggeris. Akhir sekali, nilai ofset yang terdapat dalam hasil penjanaan sinonim dan antonim akan dipadankan semula dengan WordNet Bahasa bagi mendapatkan perkataan dalam BM. Penjanaan leksikon sentimen menghasilkan sebanyak 14337 lema iaitu sebanyak 6915 lema adalah positif dan 7422 lema adalah negatif. Terdapat tiga jenis eksperimen bagi menilai ketepatan leksikon sentimen iaitu penilaian piawai emas oleh penutur BM, penilaian polariti perkataan dengan persilangan kata leksikon *General Inquirer* (GI) dan penilaian pengelasan polariti perkataan. Peratusan persetujuan yang diberikan oleh penutur BM yang tertinggi ialah sebanyak 86.58%. Manakala ukuran ketepatan polariti perkataan dengan GI yang tertinggi adalah 0.906 dan peratusan ukuran-F1 bagi pengujian pengelasan polariti perkataan ke atas tiga data ulasan pelbagai domain yang tertinggi adalah 91.31%. Keputusan pengujian telah menunjukkan keberkesanan algoritma yang dicadangkan dalam penjanaan leksikon sentimen dalam Bahasa Melayu berasaskan WordNet. Namun begitu, hasil penjanaan perlu disemak bagi menghasilkan leksikon sentimen yang lebih mantap. Walau bagaimanapun, hasil kajian dapat membantu para penyelidik menggunakan leksikon sentimen dalam menjalankan analisis sentimen khususnya dalam BM.

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 to assign the sentiment of a document. Studies in Malay sentiment analysis model is growing. Therefore, requirement in Malay sentiment lexicon is high. However, Malay sentiment lexicon development is a difficult and complex task, due to the scarcity of Malay language resource. Thus, various approaches and techniques are used to generate sentiment lexicon. Besides, previous studies are not focusing on Malay sentiment lexicon sentiment development. The objective of this study is to develop Malay sentiment lexicon algorithm based on WordNet. By using this algorithm, sentiment lexicon can be generated and used to analyse the sentiment in Malay reviews. Generally, the sentiment lexicon generation process is initiated with positive and negative seed sets selection. The selected seed set are then identified in WordNet Bahasa which is mapped with English WordNet by the same offset value. Afterwards, lexicon is generated by synonym and antonym semantic relation in English WordNet. Finally, the offset values produced by synonym and antonym generation are mapped with WordNet Bahasa to produce the Malay words. The sentiment lexicon generation has produced 14337 lemmas in which 6915 are positive and 7422 are negative lemmas. Three experiments were conducted to evaluate the accuracy of the sentiment lexicon which are human judgement by Malay native speakers as a gold standard, text polarity evaluation against General Inquirer (GI) lexicon and text polarity classification. Percentage of agreement achieved is 86.58% meanwhile the best text polarity evaluation against GI is 0.906 and F1-measure result in text classification of three reviews in multiple domain is 91.31%. The result shows the effectiveness of the proposed algorithm to generate Malay sentiment lexicon based on WordNet. However, the generated sentiment lexicon need to be manually checked to produce a reliable lexicon. Nevertheless, the outcome of this study can assist researchers in developing sentiment analysis model by using sentiment lexicon particularly in Malay.

KANDUNGAN

Halaman

SENARAI JADUAL

SENARAI ILUSTRASI

BAB I

PENDAHULUAN

1.1 PENGENALAN

Teknologi sesawang dipacu pengguna mencetus revolusi media sosial seperti *Facebook*, *Twitter* dan *blog* sebagai platform utama bagi mengutarakan pendapat (Smith 2009). Berdasarkan statistik kajian media sosial global 2017, seramai 2.789 bilion bilangan pengguna media sosial (Chaffey 2017) dan menurut Heinonen (2011) perkongsian pendapat adalah aktiviti utama pengguna media sosial. 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 penulis terhadap keseluruhan konteks dokumen sama ada sentimen adalah positif atau negatif (Padmaja & Sameen Fatima 2013). Dalam proses analisis sentimen melalui pendekatan tanpa penyelia, leksikon sentimen dijana bagi menentukan polariti unit teks dalam dokumen pendapat sama ada positif atau negatif (Bo Pang & Lee 2008). Dengan itu, pendapat pengguna sama ada positif atau negatif merupakan analisis sentimen yang diimplemen bagi merancang strategi meningkatkan mutu produk dan perkhidmatan (Liu 2012).

Penjanaan sentimen leksikon mempunyai tiga pendekatan utama iaitu pendekatan manual, pendekatan berasaskan korpus dan pendekatan berasaskan kamus. Kebanyakan kajian dilakukan menggunakan pendekatan automatik iaitu berasaskan kamus(Hassan et al. 2014; Mahyoub et al. 2014; Wu et al. 2016) dan pendekatan berasaskan korpus (Huang et al. 2014) 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) bagi meningkatkan pencapaian kejituan bagi keputusan pengujian leksikon sentimen.

Kajian yang dijalankan pada awalnya kebanyakan dalam Bahasa Inggeris manakala bahasa lain adalah sedikit kerana kekurangan sumber leksikon dalam bahasa tersebut. Antara bahasa yang digunakan dalam kajian penjanaan leksikon sentimen ialah 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). Bagi Bahasa Melayu pula, kajian yang fokus kepada penjanaan leksikon sentimen adalah sedikit (Sonai et al. 2017) manakala yang lain adalah fokus kepada model analisis sentimen yang menggunakan pendekatan tanpa penyelia (Nurul Fathiyah et al. 2015; Yee Liau & Pei Tan 2014).

Analisis sentimen semakin giat dilaksanakan (Ahmed & Nazlia 2015; Chen et al. 2014; Vania et al. 2014) ekoran data sentimen yang semakin banyak di media sosial. Kajian dijalankan terhadap ulasan Bahasa Inggeris secara meluas berbanding Bahasa Melayu (Al-Moslmi et al. 2015; Norulhidayah et al. 2013; Nurul Fathiyah et al. 2015) kerana sumber digital dalam Bahasa Melayu adalah terhad (Nurril Hirfana et al. 2011). 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). Jadual 1.1 menunjukkan ringkasan pendekatan dan teknik yang digunakan dalam kajian analisis sentimen yang telah dijalankan ke atas ulasan dalam Bahasa Melayu.

Jadual 1.1 **Ringkasan pendekatan analisis sentimen dalam Bahasa Melayu**

No	Kajian	Pendekatan	Teknik
1.	Samsudin, Puteh & Hamdan (2011)	Penyelia	Mesin vektor sokongan, Naïve Bayes dan Jiran K-terhampir
2.	Alshalabi et al. (2013)	Penyelia	Naïve Bayes, Jiran K-terhampir dan N-gram

bersambung…

…sambungan

3.	Norlela Samsudin (2013)	Penyelia	MyTNA (Malay Mixed Text Normalization Approach)
4.	Mazidah Puteh et al. (2013)	Penyelia	Sistem Imun Buatan
5.	Norulhidayah et al. (2013)	Penyelia	Sistem Imun Buatan
6.	Saloot et al. (2014)	Tanpa Penyelia	Sumber diterjemah ke Bahasa Melayu dan WordNet sebagai rujukan.
7.	Yee Liau, & Pei Tan (2014)	Tanpa Penyelia	Sumber leksikon Bahasa Melayu dibangunkan secara manual dan penggunaan SentiStrength
8.	Al-Moslmi et al. (2015)	Penyelia	Mesin vektor sokongan, Naïve Bayes dan Jiran K-terhampir
9.	Al-Saffar (2015)	Penyelia	Mesin vektor sokongan, Naïve Bayes dan Jiran K-terhampir
10.	Nurul Fathiyah et al. (2015)	Tanpa Penyelia	WordNet dan SentiWordNet sebagai rujukan untuk kamus skor.
11.	Ahmed & Nazlia (2015)	Gabungan pendekatan (Penyelia dan tanpa penyelia)	WordNet diterjemah ke Bahasa Melayu
12.	Nurul Fathiyah et al. (2016)	Tanpa Penyelia	WordNet sebagai rujukan untuk kamus skor

Jadual 1.1 memperlihatkan bahawa pendekatan tanpa penyelia kurang digunakan berbanding dengan pendekatan penyelia kerana kekurangan sumber bagi pembangunan leksikon sentimen dalam Bahasa Melayu (Nurul Fathiyah et al. 2015). Berdasarkan Jadual 1.1 terdapat 5 kajian yang menggunakan pendekatan tanpa penyelia iaitu berasaskan leksikon berbanding 7 kajian yang menggunakan pendekatan penyelia. Manakala teknik yang digunakan dalam penjanaan leksikon adalah pembangunan secara manual, terjemahan daripada kamus dan rujukan WordNet. Berdasarkan kajian-kajian tersebut, tidak pernah dinyatakan bahawa leksikon sentimen yang dibangunkan, diuji dengan piawai emas atau penilaian oleh pakar analisis sentimen dalam Bahasa Melayu.

Sehingga kini, kajian leksikon sentimen yang fokus kepada perbendaharaan kata berpolariti secara automatik dalam Bahasa Melayu hanya sedikit seperti Sonai et al. (2017), Nasharuddin et al. (2017) dan Darwich et al. (2016). Walaupun pengujian bagi kajian-kajian tersebut menunjukkan keputusan ketepatan leksikon sentimen yang

baik namun penambahbaikkan dari segi kaedah yang digunakan dalam pembangunan leksikon sentimen adalah perlu bagi menghasilkan pendekatan pembangunan yang lebih mantap.

1.2 PENYATAAN MASALAH

Leksikon sentimen dalam Bahasa Inggeris adalah banyak dan meluas manakala dalam bahasa lain seperti Perancis, Cina, Hindi dan Sepanyol adalah sedikit (Bakliwal et al. 2012; Chen et al. 2014; Perez-Rosas et al. 2012). Analisis sentimen dalam bahasa lain semakin giat dijalankan namun pembinaan leksikon sentimen dalam bahasa yang khusus (selain daripada Bahasa Inggeris) masih sedikit berbanding bahasa Inggeris memandangkan sumber bahasa adalah terhad (Chen et al. 2014). Oleh kerana kajian leksikon sentimen dalam bahasa selain Inggeris adalah terhad ekoran daripada kekangan sumber bahasa, pembinaan leksikon sentimen selain Bahasa Inggeris adalah perlu (Vania et al. 2014).

Bahasa Melayu adalah tergolong dalam kategori sumber bahasa yang terhad. Sehubungan itu, sumber yang diperlukan untuk membina leksikon sentimen adalah terbatas (Nurril Hirfana et al. 2011). Permasalahan utama Bahasa Melayu yang mempunyai kekurangan sumber ialah hampir tiada leksikon sentimen yang piawai dibangunkan (Nasharuddin et al. 2017). 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, penjanaan leksikon sentimen secara manual mengambil masa yang lama dan tenaga yang banyak (Liu 2012) serta adakalanya meletakkan polariti kata yang tidak tepat dan berat sebelah (Andreevskaia & Bergler 2006). Manakala penjanaan secara automatik melalui kaedah penterjemahan dari Bahasa Inggeris ke Bahasa Melayu, walaupun mudah, namun perkataan yang diterjemah boleh mempunyai makna yang berbeza dan ini mempengaruhi polariti perkataan tersebut (Sonai et al. 2017). Dengan itu, penjanaan secara automatik berasaskan kamus seperti WordNet Bahasa Inggeris adalah pendekatan yang telah berjaya digunakan dalam bahasa lain (Bakliwal et al. 2012; Darwich et al. 2015, 2016; Kim & Hovy 2004).

WordNet Bahasa iaitu pangkalan data leksikal Bahasa Melayu yang dibangunkan berasaskan WordNet Bahasa Inggeris tidak mempunyai ciri-ciri hubungan semantik antara perkataan seperti sinonim dan antonim (Bond et al. 2014). Oleh yang demikian, WordNet Bahasa Inggeris dijadikan sebagai asas utama dalam penjanaan leksikon sentimen dan WordNet Bahasa adalah sebagai rujukan bagi penterjemahan perkataan melalui nilai ofset. Namun begitu, polariti perkataan yang dihasilkan berasaskan penjanaan perkataan sinonim dan antonim adakalanya kurang tepat (Darwich et al. 2016) kerana tiada perbandingan dengan piawai emas. Sehubungan itu penggunaan kaedah dan algoritma dalam pembangunan leksikon sentimen dan sumber yang mantap adalah perlu bagi memelihara ketepatan polariti perkataan dalam Bahasa Melayu (Nurul Fathiyah et al. 2015). Oleh itu, berpandukan kepada kajian sedia ada, ia boleh membantu pembangunan leksikon sentimen dalam Bahasa Melayu berasaskan WordNet.

1.3 OBJEKTIF

Matlamat kajian adalah untuk mencadang pembangunan algoritma bagi menjana leksikon sentimen yang dapat digunakan untuk menganalisa sentimen dalam Bahasa Melayu. Bagi menghasilkan algoritma penjanaan leksikon sentimen ini, beberapa objektif perlu dicapai iaitu:

i. Membangun algoritma penjanaan leksikon sentimen berasaskan WordNet yang dapat digunakan untuk menganalisa sentimen dalam Bahasa Melayu.

ii. Menilai ketepatan leksikon sentimen yang dicadangkan berbanding dengan piawai emas.

1.4 SKOP KAJIAN

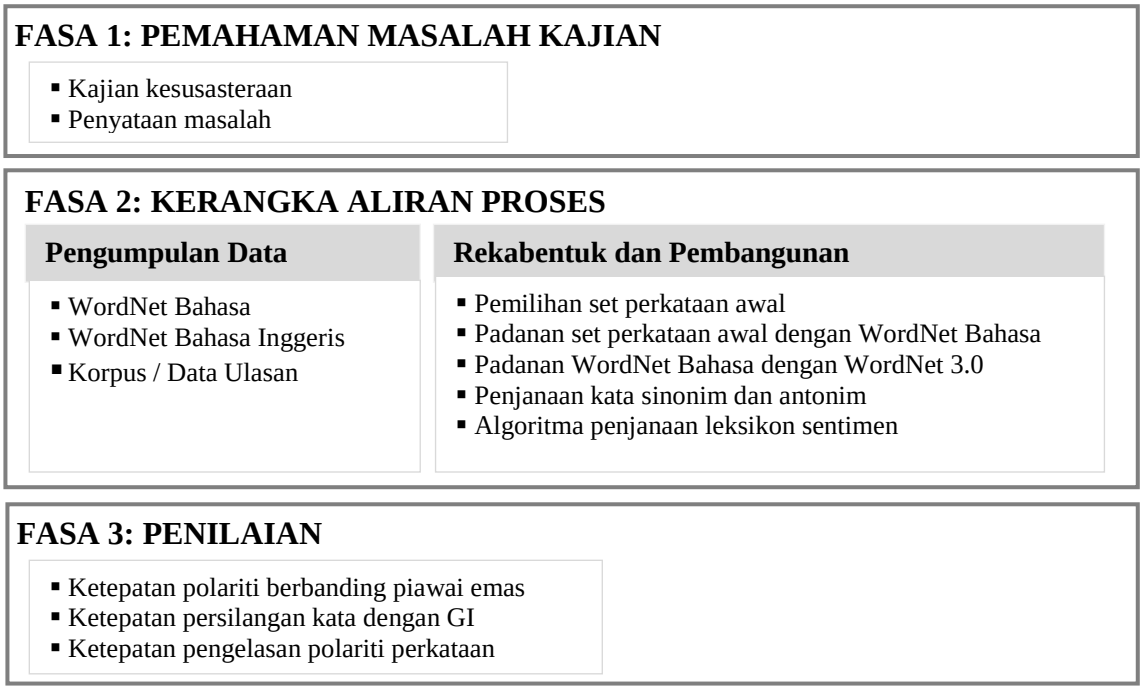
Skop kajian adalah terhad kepada pembangunan algoritma penjanaan leksikon sentimen dalam Bahasa Melayu. Kaedah pembangunan adalah menggunakan pendekatan kamus setelah mempertimbangkan bahawa kaedah ini berjaya digunakan dalam bahasa selain daripada Bahasa Melayu. Data yang digunakan untuk

pembangunan adalah WordNet Bahasa Melayu dan WordNet Bahasa Inggeris dengan memilih kata adjektif sahaja. Pemilihan kata adjektif dalam kajian ini adalah kerana kata adjektif menerangkan sifat atau perkara tentang sesuatu atau seseorang. Contoh yang menerangkan sifat perasaan seperti suka dan benci. Contoh lain iaitu sifat pandang seperti cantik, jelita, hodoh dan buruk. Perkataan-perkataan ini sering digunakan untuk menggambarkan sentimen terhadap ulasan pengguna.

WordNet Bahasa Melayu merupakan asas dalam pembangunan leksikon sentimen. Manakala WordNet Bahasa Inggeris yang mempunyai ciri-ciri hubungan antara perkataan secara semantik membolehkan teknik butstrap dilakukan bagi mendapatkan senarai sinonim dan antonim perkataan. Dengan itu, hasil kajian adalah algoritma penjanaan leksikon sentimen dalam Bahasa Melayu, hasil penilaian ketepatan keputusan leksikon sentimen dan data hasil dari penjanaan yang mengandungi golongan kata, polariti kata, dan kata synset.

1.5 REKA BENTUK KAJIAN

Reka bentuk kajian adalah penting dalam menggambarkan aktiviti penyelidikan yang menyeluruh untuk dilakukan disepanjang kajian. Reka bentuk kajian terbahagi kepada tiga fasa utama iaitu Fasa 1 Pemahaman masalah kajian, Fasa 2 Kerangka aliran proses dan Fasa 3 Penilaian. Semua modul yang terdapat dalam setiap fasa adalah seperti yang ditunjukkan dalam Rajah 1.1.



Rajah 1.1 Reka bentuk kajian

1.5.1 Fasa 1: Pemahamanan Masalah Kajian

Fasa 1 iaitu Fasa Input Kajian merupakan fasa yang paling penting dalam peringkat awal kajian. Aktiviti pada peringkat awal kajian adalah menganalisa kandungan terhadap kajian kesusasteraan bagi meninjau kajian terdahulu yang berkaitan dengan penjanaan leskikon sentimen. Fasa ini menyumbang ke arah pemahaman tentang sesuatu permasalahan kajian yang dikenalpasti. Selain itu, pendekatan yang digunakan dalam kajian lepas turut dikenalpasti. Seterusnya, permasalahan kajian akan dihurai dengan jelas bagi memastikan objektif kajian dapat dibina dengan tepat. Kupasan lanjut tentang permasalahan kajian dan objektif telah dinyatakan dalam Bab I.

1.5.2 Fasa 2: Kerangka Aliran Proses

Fasa 2 iaitu Kerangka Aliran Proses adalah fasa penting untuk menjelaskan pendekatan dan kaedah yang digunakan bagi menyelesaikan permasalahan serta mencapai objektif yang telah dikenalpasti. Fasa ini penting bagi memastikan aktiviti-aktiviti yang dilakukan adalah jelas dan teratur. Terdapat empat modul utama yang terlibat dalam fasa ini iaitu sumber data yang digunakan dalam kajian, proses

penetapan set perkataan awal, pembangunan algoritma leksikon sentimen dan rekabentuk eksperimen bagi menguji ketepatan leksikon sentimen yang telah dihasilkan. Semua empat modul tersebut dibincangkan dalam bab seterusnya iaitu dalam Bab III.

1.5.3 Fasa 3: Penilaian

Fasa 3 iaitu Penilaian menerangkan teknik yang digunakan untuk menilai hasil kajian iaitu untuk menguji ketepatan polariti perkataan dalam leksikon sentimen yang dibangunkan. Terdapat tiga jenis penilaian dilakukan iaitu menguji ketepatan polariti perkataan berbanding piawai emas, menguji ketepatan polariti ketepatan persilangan kata berbanding leksikon sentimen *General* Inquirer (GI) dan menguji pengelasan polariti perkataan. Perincian kaedah dan alasan penilaian ini digunakan diterangkan dalam Bab III.

1.6 STRUKTUR ORGANISASI TESIS

Struktur kandungan penulisan tesis ini telah dibahagikan kepada lima bab dijelaskan seperti yang dibawah.

Bab II Sorotan Kesusasteraan memfokuskan perbincangan mengenai kajian lampau iaitu pendekatan yang digunakan dalam pembinaan leksikon sentimen dan sumber pembangunan leksikon sentimen. Bab ini juga turut mengupas pembangunan leksikon sentimen dalam Bahasa Melayu.

Bab III Metodologi Kajian menerangkan kaedah dan rekabentuk kajian yang dijalankan serta algoritma yang dibangunkan bagi menjana leksikon sentimen berasaskan WordNet.

Bab IV Keputusan dan Penilaian membincangkan hasil ukuran ketepatan leksikon sentimen yang telah dijana berdasarkan penilaian manual. Selain itu penilaian algoritma berasaskan garis asas juga turut dibincangkan. Penilaian leksikon sentimen ke atas beberapa data ulasan juga turut dilakukan.

Bab V Rumusan dan Penutup menyenaraikan sumbangan kajian bersama dengan cadangan kajian lanjutan setelah mengenal pasti kekurangan yang telah dijumpai sepanjang kajian dilakukan.

1.7 KESIMPULAN

Bab ini memberi penerangan yang menyeluruh tentang kajian yang akan dijalankan. Penerangan diberi mengenai penyataan masalah, objektif kajian dan skop kajian. Selain daripada itu, bab ini memberikan pemahaman konteks kajian yang menyeluruh sebelum memahami dengan lebih mendalam dalam bab yang seterusnya.

BAB II

SOROTAN KESUSASTERAAN

2.1 PENGENALAN

Bab ini menerangkan secara terperinci kajian menyeluruh status terkini dalam bidang leksikon sentimen yang berkait dengan kajian ini. Oleh itu kajian ini tertumpu kepada kaedah atau pendekatan penjanaan leksikon sentimen atau leksikon pendapat. Isu utama dibawah tumpuan penyelidikan semasa, dinyatakan serta turut menggambarkan kedudukan Bahasa Melayu dalam bidang penjanaan leksikon sentimen.

Sorotan kesusasteraan dimulai dengan kajian lampau penjanaan leksikon sentimen dan diikuti dengan definisi leksikon sentimen. Bab ini juga turut menjelaskan pendekatan penjanaan leksikon sentimen. Setelah itu, kaedah pembangunan menggunakan pendekatan kamus turut dikupas. Kaedah kajian penjanaan leksikon sentimen menggunakan sumber bahasa yang terhad seperti Bahasa Sepanyol, Bahasa Turki, Bahasa Hindi dan Bahasa Indonesia akan dinyatakan. Seterusnya pembangunan leksikon sentimen berasaskan WordNet akan dijelaskan. Setelah itu, sumber yang digunakan untuk menjana leksikon sentimen akan dinyatakan. Kebanyakkan sumber yang digunakan adalah dalam Bahasa Inggeris namun sumber dalam Bahasa Melayu dijadikan sebagai asas. Akhir sekali leksikon sentimen dalam Bahasa Melayu akan dijelaskan dengan terperinci.

Penyelidikan dalam bidang penjanaan leksikon sentimen sedang berkembang ekoran keperluan dalam analisis sentimen di era teknologi sesawang dipacu pengguna. Pada mulanya kajian tertumpu kepada Bahasa Inggeris dan pembangunan secara manual (Bo Pang et al. 2002a) untuk menentukan polariti dokumen melalui pengiraan kata positif dan negatif. Pendekatan yang sama digunakan dengan menggabungkan

sumber-sumber lain seperti kamus dan korpus dari laman web (Kennedy & Inkpen 2006) untuk pengelasan sentimen ulasan filem.

Pembinaan leksikon secara manual memerlukan tenaga dan masa yang banyak (Das & Chen 2007). Bagi menangani masalah ini, pendekatan automatik digunakan seperti penggunaan kata hubung set sinonim kata adjektif dalam WordNet bagi mengembangkan saiz leksikon pendapat (Mahyoub et al. 2014). Pendekatan ini pernah dilakukan untuk mengesan kesubjektifan (Wilson et al. 2005) dan mengelaskan sentimen (Pang & Lee 2004; Salvetti et al. 2004). Kaedah lain seperti Markov random walk digunakan bagi mengenalpasti polariti perkataan berasaskan rangkaian perkataan yang dibina berasaskan sumber leksikal (Hassan et al. 2014). Selain itu pembangunan leksikon sentimen berasaskan sentimen pengetahuan digunakan iaitu pemilihan perkataan sentimen yang diwakili emotikon, persamaan sentimen yang disari daripada perkataan yang digabung dalam ulasan dan mengenalpasti perkataan baru yang sering digunakan dalam ulasan (Wu et al. 2016).

Kajian analisis sentimen semakin berkembang dalam bahasa lain selain Bahasa Inggeris dan keadaan ini mendesak keperluan sumber leksikon sentimen dalam bahasa selain Bahasa Inggeris. Terdapat 4 pendekatan yang sering digunakan untuk membangunkan leksikon pendapat atau leksikon sentimen bagi bahasa lain selain Bahasa Inggeris (Banea et al. 2011) iaitu i) anotasi korpus secara manual, ii) unjuran silang bahasa berasaskan korpus, iii) pembangunan berasaskan kamus dan iv) terjemahan menggunakan kamus. Ringkasan pendekatan yang digunakan bagi pembangunan leksikon sentimen selain Bahasa Inggeris adalah seperti yang ditunjukkan dalam Jadual 2.1.

Jadual 2.1 Ringkasan bagi pendekatan yang digunakan dalam pembangunan leksikon bukan Bahasa Inggeris

Pendekatan	Kajian	Bahasa
Anotasi korpus secara manual	Muhammad, Mona & Mohammed (2011)	Arab
Pembangunan berasaskan korpus	Mihalcea, Banea & Wiebe (2007)	Romania
	Wan (2009)	Cina

bersambung…

…sambungan

	Banea, Mihalcea & Wiebe (2008)	Romania
Pembangunan berasaskan kamus	Rao, & Ravichandran (2009)	Hindi dan Perancis
	Vania et al. (2014)	Indonesia
	Perez-Rosas et al. (2012)	Sepanyol
Penterjemahan terus menggunakan kamus	Franky et al.(2015)	Indonesia

Pembangunan leksikon sentimen berasaskan anotasi korpus secara manual amat jarang dilakukan kerana memerlukan masa yang lama dan tenaga yang banyak. Muhammad, Mona & Mohammed (2011) telah membangunkan korpus yang dianotasikan dengan Bahasa Arab moden bersama-sama dengan leksikon polariti yang baru. Kajian tersebut menggunakan ciri-ciri bahasa tak bersandar dengan melabelkan setiap frasa dengan domain dokumen secara manual. Ciri-ciri lain yang turut digunakan iaitu morfologi yang dimodelkan melalui pembentukan kata seperti orang, aspek dan tempoh. Kajian ini telah memberi impak yang besar dari segi prestasi dalam mengeksploit leksikon sentimen yang dibangunkan.

Manakala bagi unjuran silang bahasa berasaskan korpus, kaedah ini merupakan silang bahasa antara bahasa utama dengan bahasa sasaran iaitu penterjemahan dilakukan dalam dua arah. Kaedah ini diperkenalkan oleh Mihalcea, Banea & Wiebe (2007) bagi unjuran label kesubjektifan silang selari bagi teks Bahasa Romania-Bahasa Inggeris. Kemudian kaedah ini digunakan dalam gabungan bersama penterjemahan mesin bagi unjuran label kesubjektifan dalam Bahasa Romania (Banea, Mihalcea, Wiebe et al. 2008) dan anotasi sentimen dalam Bahasa Cina (Wan 2009).

Penjanaan berasaskan kamus adalah antara pendekatan yang banyak digunakan dalam pembangunan leksikon sentimen. Pendekatan ini menggunakan kaedah butstrap (*bootstrap*) iaitu penggunaan set perkataan awal (*seed set*) bagi rambatan perkataan sinonim dan antonim dalam kamus elektronik (Banea, Mihalcea & Wiebe 2008), WordNet (Rao & Ravichandran 2009), OpinionFinder (Vania et al. 2014) atau MPQA (Perez-Rosas et al. 2012).

Pendekatan penterjemahan terus menggunakan kamus merupakan salah satu kaedah yang digunakan dalam pembangunan leksikon sentimen. Sumber leksikon dalam Bahasa Inggeris digunakan sebagai rujukan dalam penjanaan leksikon sentimen dan diterjemahkan ke bahasa sasaran seperti yang dilakukan oleh Franky et al.(2015). Penterjemahan yang digunakan adalah secara atas talian seperti *Terjemah Google*, *MOSES* dan *Kamus Online Dwi-Bahasa*.

2.2 LEKSIKON SENTIMEN

Leksikon adalah perbendaharaan kata bagi sesuatu bahasa atau lebih dikenali sebagai kamus. Leksikon sentimen merupakan perbendaharaan kata yang mempunyai polariti sama ada positif atau negatif. Perkataan sentimen positif menggambarkan keadaan yang diingini iaitu polariti yang positif. Manakala perkataan sentimen negatif menggambarkan yang sebaliknya iaitu keadaan yang tidak diingini atau polariti yang tidak disukai (Liu 2012). Sebagai contoh, “baik, bagus, suka” adalah dalam kumpulan perkataan polariti positif manakala “jahat, buruk, teruk” adalah dalam kumpulan perkataan polariti negatif.

Leksikon sentimen merupakan sumber linguistik yang digunakan dalam analisis sentimen (Darwich et al. 2015). Ia adalah sumber penting dalam tugasan analisis sentimen (Vania et al. 2014) bagi menentukan polariti pendapat. Pendekatan leksikon sentimen penting dalam menentukan haluan polariti (positif dan negatif) dan kekuatan sentimen (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).

Sehingga kini, kajian berkaitan analisis sentimen semakin berkembang khususnya pembangunan pendekatan leksikon sebagai sumber utama dalam analisis sentimen. Namun kajian leksikon sentimen banyak tertumpu kepada Bahasa Inggeris (Franky et al. 2015; Perez-Rosas et al. 2012; Vania et al. 2014) berbanding dengan bahasa-bahasa utama yang lain seperti Bahasa Peracis, Sepanyol dan Cina. Kajian leksikon sentimen yang fokus kepada perbendaharaan kata berpolariti secara

automatik dalam Bahasa Melayu adalah sangat sedikit (Ahmed & Nazlia 2015; Nurul Fathiyah et al. 2015).

2.3 PENJANAAN LEKSIKON

Leksikon sentimen dibangunkan sebagai sumber pengelasan polariti perkataan dalam model analisis sentimen. Pembangunan leksikon dilakukan dengan menambah perkataan atau lema dalam senarai mengikut polariti sama ada positif atau negatif bergantung kepada maksud sesuatu perkataan tersebut. Penambahan perkataan ke dalam senarai dilakukan sama ada secara manual iaitu dimasukkan sendiri oleh pembangun atau secara automatik melalui penjanaan perkataan. Penjanaan perkataan dilakukan adalah berasaskan sama ada kamus atau korpus.

2.3.1 Set Perkataan Awal

Penjanaan perkataan secara automatik memerlukan, set perkataan awal sebagai rujukan kepada sumber untuk penjanaan. Set perkataan awal merupakan perkataan yang digunakan sebagai rujukan untuk mendapatkan perkataan lain yang mempunyai hubungan dengannya. Perkataan lain yang diperoleh akan dimasukkan sebagai senarai dalam leksikon.

Menurut Liu (2010), set perkataan awal yang mengandungi perkataan yang berupa pendapat perlu disediakan. Penyediaan set perkataan awal ini adalah secara manual. Set perkataan ini terdiri daripada set perkataan positif dan set perkataan negatif. Set perkataan tersebut boleh mengandungi kata kerja dan kata adjektif (Kim & Hovy 2004), kata nama, kata kerja dan kata adjektif (Andreevskaia & Bergler 2006) atau perkataan yang mempunyai nilai kekuatan polariti yang tinggi (Perez-Rosas et al. 2012).

2.3.2 Penjanaan Perkataan

Penambahan perkataan ke dalam senarai leksikon secara automatik dilakukan melalui kaedah penjanaan perkataan. Selain daripada set perkataan awal, penjanaan perkataan secara automatik juga perlukan sumber seperti kamus atau korpus. Sumber digunakan

sebagai rujukan dan input kepada senarai perkataan yang dimasukkan ke dalam leksikon.

Penjanaan dilakukan adalah untuk menambahkan senarai perkataan dalam leksikon mengikut polariti. Dengan menggunakan dua set perkataan awal, iaitu set positif dan set negatif, penjanaan dilakukan dengan mencari hubungan perkataan yang dikehendaki dalam kamus seperti WordNet atau korpus. Setelah itu, perkataan yang dijumpai, dimasukkan ke dalam senarai set perkataan. Kemudian, set perkataan ini digunakan dalam penjanaan bagi lelaran yang seterusnya seperti mana proses penjanaan dilakukan ke atas set perkataan awal tadi.

Penjanaan perkataan dilakukan berdasarkan hubungan perkataan dengan perkataan lain. Kaedah yang telah dilakukan dalam kajian sebelum ini adalah seperti penjanaan perkataan sinonim dan antonim (Bakliwal et al. 2012; Hu & Liu 2004; Kim & Hovy 2004), persamaan semantik *Latent* (Banea, Mihalcea & Wiebe 2008), kata hubung (Hatzivassiloglou & McKeown 1997) dan berasaskan corak sentimen (Vania et al. 2014).

2.3.3 Terjemahan Perkataan

Pembangunan leksikon sentimen selain daripada Bahasa Inggeris berhadapan dengan sumber leksikon yang terhad. Dengan itu, adalah perlu menggunakan sumber Bahasa Inggeris dalam proses penjanaan leksikon dan menterjemahkan hasil penjanaan ke bahasa yang dikehendaki.

Penjanaan leksikon yang berasaskan WordNet Bahasa Inggeris perlu menterjemahkan perkataan yang dihasilkan ke bahasa yang dikehendaki. Proses penterjemahan adalah mudah sekiranya mempunyai WordNet dalam bahasa lain seperti WordNet Sepanyol (Perez-Rosas et al. 2012) dan Hindi WordNet (Bakliwal et al. 2012). Penterjemahan dilakukan dengan menggunakan nilai ofset bagi *synset* yang sama memandangkan WordNet dalam bahasa lain dibangunkan berasaskan WordNet Bahasa Inggeris.

Penterjemahan menggunakan kamus dwi-bahasa juga boleh digunakan dalam pembangunan leksikon sentimen. Mihalcea et al. (2007) membangunkan leksikon sentimen dalam Bahasa Romania berasaskan pangkalan data leksikal OpinionFinder. Sumber kamus yang digunakan untuk menterjemahkan hasil penjanaan adalah kamus dwi-bahasa (*Bahasa Inggeris-Bahasa Romania).* Selain daripada itu, hasil penjanaan leksikon sentimen yang berasaskan kamus dalam Bahasa Inggeris juga diterjemahkan secara terus menggunakan terjemah *Google* (Franky et al. 2015).

Terdapat beberapa kaedah yang digunakan bagi penterjemahan Bahasa Inggeris – Bahasa Melayu bagi sumber korpus dan leksikon. Antara kaedah yang digunakan adalah penterjemahan silang bahasa melalui padanan nilai ofset WordNet Bahasa Inggeris ke WordNet Bahasa bagi yang menggunakan sumber leksikon WordNet (Darwich et al. 2016; Nurul Amelina et al. 2017). Selain daripada itu penterjemah *Google* juga digunakan untuk penterjemahan Bahasa Inggeris – Bahasa Melayu (Hijazi et al. 2017).

Hasil penjanaan perkataan yang berasaskan sumber dalam Bahasa Inggeris diterjemahkan ke bahasa yang dikehendaki. Walau pun sumber leksikon dalam bahasa selain Bahasa Inggeris adalah terhad, pembangunan leksikon sentimen dalam bahasa lain menjadi mudah dengan adanya sumber dan teknik sedia ada. Namun demikian, penterjemahan yang dilakukan kadang kala tidak memberikan maksud yang sebenar dan adakalanya memberikan maksud yang berlainan (Dehkharghani et al. 2016). Hasil terjemahan yang tidak tepat akan memberi kesan terhadap pengelasan polariti perkataan itu sendiri (Sonai et al. 2017).

2.4 PENDEKATAN PENJANAAN LEKSIKON

Model sentimen analisis yang menggunakan pendekatan berasaskan leksikon memerlukan sumber linguistik bagi menentukan polariti perkataan atau ayat yang mengandungi sentimen. Terdapat tiga pendekatan yang digunakan untuk pembangunan leksikon sentimen iaitu pendekatan manual, berasaskan kamus dan berasaskan korpus. Jadual 2.2 menunjukkan ringkasan pendekatan yang digunakan dalam kajian pembangunan leksikon sentimen.

Jadual 2.2 Ringkasan pendekatan yang digunakan dalam kajian penjanaan leksikon sentimen

Pendekatan	Kajian	Fitur	Bahasa
Manual	Das & Chen (2007)	berdasarkan perkataan yang mempunyai nilai pembeza yang tinggi dengan memilih perkataan berdasarkan pembacaan mengikut domain.	Inggeris
	Muhammad et al. (2011)	leksikon sentimen bagi domain agensi berita yang dianotasikan dengan Bahasa Arab moden.	Arab
Kamus	Kamps et al. (2004)	kaedah ukuran persamaan semantik berdasarkan jarak antara calon perkataan dan polariti perkataan menggunakan WordNet	Inggeris
	Kim & Hovy (2004)	membangunkan dua senarai perkataan awal yang mengandungi kata kerja dan kata adjektif yang positif dan negatif bagi menyari kata sinonim dan antonim dalam WordNet.	Inggeris
	Banea, Mihalcea & Wiebe (2008)	Set kata subjektif dijadikan sebagai set perkataan awal digunakan sebagai pertanyaan melalui kamus dalam talian dan disaring menggunakan ukuran pengiraan persamaan.	Romania
	Rao & Ravichandran (2009)	pengelasan polariti ke perkataan lain menggunakan graf berasaskan algoritma pembelajaran semi-penyelia.	Hindi Perancis
	Hassan & Radev (2010)	algoritma *Markov random walk* untuk mengenal pasti polariti perkataan.	Arab Inggeris
Korpus	Hatzivassiloglou & McKeown (1997)	menggunakan kata hubung '*dan'* dan '*tetapi'* bagi menentukan polariti korpus.	Inggeris
	Turney (2002)	pendekatan berorientasikan semantik purata dalam mengenal pasti ulasan sama ada positif atau negatif.	Inggeris
	Kaji & Kitsuregawa (2007)	Penggunaan struktur susun atur seperti pembutiran atau jadual dan saranan perkataan seperti *pro* dan *kontra* untuk menyari positif dan negatif ayat penilaian.	Jepun

bersambung…

…sambungan

Bautin, Vijayarenu & Skiena (2008)	dokumen diterjemah ke Bahasa Inggeris diikuti dengan pengiraan polariti terhadap entiti menggunakan ukuran sekutuan.	Jerman
Vania et al. (2014)	kaedah pengembangan leksikon berasaskan corak sentimen dan terjemahan secara automatik.	Indonesia

Berikut ialah penerangan tentang kajian pembangunan leksikon sentimen yang telah dijalankan menggunakan pendekatan manual, berasaskan kamus dan berasaskan korpus.

2.4.1 Pendekatan Manual

Pendekatan penjanaan leksikon sentimen secara manual adalah aktiviti menganotasikan perkataan mengikut polariti sama ada dalam kelas positif, negatif atau neutral. Perkataan dipilih berdasarkan korpus sedia ada dan ditentukan sama ada perkataan tersebut menggambarkan sentimen iaitu positif, negatif atau neutral. Pemilihan perkataan dilakukan melalui kaedah penyelia skim pemberat istilah dan kemudiannya anotasi secara manual sama ada positif, negatif atau neutral (Kotelnikov et al. 2016).

Muhammad et al. (2011) telah membangunkan leksikon sentimen bagi domain agensi berita yang dianotasikan dalam Bahasa Arab moden iaitu sebanyak 3982 kata adjektif yang dilabelkan sama ada positif, negatif atau neutral. Manakala Das & Chen (2007) menjana leksikon dalam Bahasa Inggeris berdasarkan perkataan yang mempunyai nilai pembeza yang tinggi dengan memilih perkataan berdasarkan pembacaan mengikut domain secara manual. Setakat ini saiz leksikon adalah sebanyak 300 perkataan.

Dalam kajian Dehkharghani et al. (2016), leksikon sentimen dalam Bahasa Turki dibangunkan menggunakan sumber WordNet Turki dengan menganotasikan polariti mengikut kesesuaian secara manual. Polariti yang diberikan adalah sama ada positif, negatif atau objektif ke atas 14,795 lema.

Pendekatan secara manual membolehkan leksikon sentimen yang dibangunkan adalah fleksible berdasarkan pengelasan pengguna (Das & Chen 2007). Selain itu, penjanaan secara manual menghasilkan leksikon sentimen yang mempunyai kategori mengikut domain (Muhammad et al. 2011). Namun begitu, tugasan ini memerlukan masa yang lama dan tenaga yang banyak untuk menganotasi korpus yang bersaiz besar. Di samping itu, kebanyakkan keputusan adalah berat sebelah dari segi kesubjektifan semasa membuat anotasi memandangkan pertimbangan bagi setiap tenaga yang melakukan tugasan adalah berbeza (Andreevskaia & Bergler 2006).

2.4.2 Pendekatan Berasaskan Kamus

Pendekatan berasaskan kamus adalah menggunakan set perkataan yang mengandungi pendapat dan dipadankan dengan kamus elektronik atau pangkalan data leksikal dalam talian (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 membangunkan 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.

WordNet yang mempunyai hubungan semantik antara perkataan seperti hipernim, hiponim, sinonim dan antonim 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 mewakili 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 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. Rangkaian perkataan dibina berasaskan pelbagai sumber dan sumber utama adalah WordNet. Perkataan yang berada dalam set sinonim yang sama dihubungkan antara satu sama lain bagi membentuk graf yang mengandungi set perkataan dan golongan kata. Selain itu, makluamt kata hubung dan nod penghubung juga terdapat dalam rangkian graf perkataan tersebut. Dalam kajian lain, Hassan et al. (2014) melakukan kajian ke atas pelbagai bahasa seperti Bahasa Inggeris, Bahasa Arab dan Bahasa Hindi dengan membina rangkaian menggunakan kaedah yang sama dengan Hassan & Radev (2010). Bagi perkataan yang tiada dalam rangkaian, Hassan et al. (2014) menggunakan pendekatan pengagihan berasaskan web iaitu hasil pencarian perkataan di internet perlu melalui pra-pemprosesan terlebih dahulu. Setelah itu pemilihan perkataan dilakukan berdasarkan jumlah kekerapan perkataan dalam dokumen sebagai set perkataan yang berkaitan.

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.

2.4.3 Pendekatan Berasaskan Korpus

Pendekatan berasaskan korpus adalah penjanaan leksikon sentimen yang bergantung kepada kewujudan pola berulang. Perkataan yang disusun mengikut turutan berstruktur bagi membentuk ayat boleh memberi makna sentimen dalam korpus. Ulasan biasanya mempunyai pola perkataan yang sama terutamanya dalam domain yang sama (Vania et al. 2014). Pola perkataan yang sama merupakan peluang yang digunakan sebagai teknik dalam penjanaan leksikon sentimen. Perkataan berbentuk pendapat yang terdapat di dalam senarai set perkataan awal akan digunakan untuk mencari perkataan pendapat yang lain di dalam korpus (Liu 2010).

Teknik penggunaan kata hubung telah dicetuskan oleh Hatzivassiloglou & McKeown (1997) bagi menentukan polariti perkataan. Penjanaan leksikon bagi dua perkataan yang mempunyai kata hubung ‘*dan’* diberi nilai polariti yang sama manakala dua perkataan yang mempunyai kata hubung ‘*tetapi’* diberi polariti yang berlawanan. Penjanaan leksikon sentimen ini menggunakan korpus jurnal *Wall Street* dalam Bahasa Inggeris.

Selain itu, kaedah semantik turut digunakan dalam pendekatan korpus bagi menjana leksikon. Kajian Turney (2002) menggunakan pendekatan berorientasikan semantik purata dalam mengenal pasti ulasan sama ada positif atau negatif. Terdapat tiga langkah yang digunakan dalam kajian ini iaitu 1) proses penyarian frasa yang mengandungi kata adjektif atau kata keterangan, 2) membuat anggaran

berorientasikan semantik bagi setiap frasa dan 3) mengelaskan ulasan berdasarkan frasa berorientasikan semantik purata.

Teknik pengecaman corak struktur ayat merupakan salah satu teknik yang digunakan dalam pendekatan korpus. Kaji & Kitsuregawa (2007) merangka pendekatan bagi membina leksikon sentimen dalam Bahasa Jepun menggunakan saranan penstrukturan menggunakan dokumen HTML. Kajian ini menggunakan senarai saranan perkataan untuk mengenal pasti penilaian klausa iaitu sama ada positif atau negatif bergantung kepada struktur ayat. Penggunaan struktur susun atur seperti pembutiran atau jadual dan saranan perkataan seperti '*pro'* dan *'kontra'* untuk menyari ayat penilaian sama ada positif atau negatif.

Vania et al. (2014) menggunakan kaedah pengembangan leksikon berasaskan corak sentimen. Teknik penjanaan leksikon sentimen yang digunakan adalah pemilihan set perkataan awal melalui penyarian ayat daripada korpus dalam Bahasa Indonesia. Set perkataan awal digunakan untuk mencari calon perkataan sentimen dalam korpus menggunakan teknik-teknik seperti golongan kata, polariti ayat dan peralihan maksud dalam ayat.

Dalam kajian yang telah dijalankan, keputusan bagi ketepatan pengelasan polariti leksikon sentimen yang dibangun menggunakan pendekatan korpus adalah baik. Namun begitu, leksikon sentimen yang dibangun menggunakan pendekatan ini hanya boleh digunakan dalam domain yang sama.

2.5 KAEDAH PEMBANGUNAN DALAM PENDEKATAN KAMUS

Pendekatan yang berasaskan kamus telah berjaya dibuktikan dalam pembangunan leksikon sentimen dalam bahasa lain selain Bahasa Inggeris (Bakliwal et al. 2012; Darwich et al. 2016; Dehkharghani et al. 2016; Perez-Rosas et al. 2012). Pendekatan ini dipilih kerana mengandungi hampir kesemua perkataan yang tidak berasaskan kepada domain yang tertentu (Sonai et al. 2017). Kaedah pembangunan leksikon sentimen berasaskan kamus terdiri daripada empat kategori iaitu 1) hubungan antara kata, 2) penterjemahan secara terus dan 3) penggunaan sumber polariti. Ringkasan

kaedah pembangunan leksikon sentimen yang menggunakan pendekatan berasaskan kamus bagi bahasa bukan Bahasa Inggeris diringkaskan seperti di dalam Jadual 2.3.

Jadual 2.3 Ringkasan kaedah pembangunan leksikon sentimen menggunakan pendekatan berasaskan kamus bukan Bahasa Inggeris

Kaedah	Kajian	Fitur	Bahasa
Kaedah hubungan antara kata dalam kamus	Bakliwal et al. (2012)	kaedah pengembangan melalui penerabasan dalam WordNet berasaskan graf.	Hindi
	Banea, Mihalcea & Wiebe (2008)	penggunaan butstrap dengan penggunaan set perkataan awal yang dijana menggunakan kamus dalam talian bagi menghasilkan senarai calon kesubjektifan yang berpotensi menggunakan pencarian sinonim perkataan.	Romania
	Eskander & Rambow (2015)	memadankan perkataan yang terdapat dalam keterangan ringkas AraMorph dengan lema yang terdapat dalam SentiWordNet	Arab
	Darwich et al. (2015, 2016)	memadankan WordNet Bahasa ke dalam WordNet melalui nilai ofset *synset* bagi mendapatkan senarai sinonim dan antonim dengan kaedah rambatan.	Melayu
	Sonai et al. (2017)	*Malay* Lexicon (MLEX) dibangunkan melalui kaedah butstrap berdasarkan perkataan sinonim dan antonim daripada pangkalan data *synset* Melayu yang berasaskan WordNet	Melayu
Kaedah penterjemahan kamus secara terus	Franky et al. (2015)	Sumber leksikon sentimen Bahasa Inggeris diterjemahkan menggunakan kamus dalam talian (*Terjemah Google)* dan dipadankan dengan korpus.	Indonesia
	Mihalcea et al. (2007)	Sumber leksikon kesubjektifan Bahasa Inggeris (*OpinionFinder)* diterjemahkan ke Bahasa Romania menggunakan kamus dwi-bahasa.	Romania
Penggunaan sumber polariti	Dehkharghani et al. (2016)	proses pengelasan *synset* SentiWordNet dan SenticNet digunakan sebagai fitur dan disari untuk mendapatkan skor positif dan negatif.	Turki

bersambung…

…sambungan

Perez-Rosas et al. (2012)	Menggunakan OpinionFinder sebagai asas perkataan yang sangat positif dan sangat negatif dan SentiWordNet untuk menetapkan skor positif dan negatif.	Sepanyol
Nasharuddin et al. (2017)	Kaedah silang bahasa menggunakan sentiWordNet dan WordNet Bahasa	Melayu

Berdasarkan ringkasan yang ditunjukkan di dalam Jadual 2.3, banyak kajian telah dijalankan dalam bahasa selain Bahasa Inggeris. Kaedah yang digunakan dalam kajian lepas diperincikan seperti dibawah.

2.5.1 Kaedah Hubungan Antara Kata

Kamus elektronik mempunyai ciri-ciri istimewa bagi hubungan antara perkataan dengan perkataan yang lain seperti sinonim, antonim dan hipernim. Kaedah hubungan antara kata dalam kamus telah digunakan dengan jayanya bagi pembangunan leksikon sentimen dalam Bahasa Inggeris (Hu & Liu 2004; Kamps et al. 2010; Kim & Hovy 2004). Selain itu, terdapat kajian yang menggunakan kaedah yang sama dalam pembangunan leksikon sentimen dalam bahasa lain selain daripada Bahasa Inggeris (Bakliwal et al. 2012; Banea, Mihalcea & Wiebe 2008; Eskander & Rambow 2015).

Bakliwal et al. (2012) dalam kajiannya menggunakan kaedah pengembangan WordNet berasaskan graf bagi membangunkan leksikon sentimen dalam Bahasa Hindi. Dengan menggunakan set perkataan awal positif dan set perkataan awal negatif, penerabasan dalam WordNet 3.0 berbentuk seperti graf iaitu setiap perkataan dihubungkan antara satu sama lain berdasarkan hubungan sinonim dan antonim. Perkataan yang telah dihubungkan tadi akan dijadikan sebagai set perkataan awal untuk dihubungkan dengan perkataan lain menggunakan kaedah yang sama. Sekiranya terdapat hubungan dengan perkataan yang telah dijumpai sebelum ini, perkataan tersebut akan diabaikan dan disambung dengan perkataan yang lain. Akhir sekali senarai perkataan yang mengandungi polariti dikumpulkan dan dijadikan sebagai leksikon sentimen. Pendekatan ini hanya menggunakan sumber WordNet 3.0.

Keputusan penilaian telah mencapai peratusan persetujuan sebanyak 70.4%. Manakala keputusan kejituan pengelasan polariti yang dicapai ialah sebanyak 79%.

Pembangunan leksikon sentimen dalam Bahasa Romania telah memperlihatkan kebaikan kaedah hubungan melalui kata dengan penggunaan butstrap seperti dalam kajian Banea, Mihalcea & Wiebe (2008). Kajian ini menggunakan beberapa perkataan sebagai set perkataan awal untuk menjana leksikon kesubjektifan secara automatik. Proses dimulakan dengan set perkataan awal yang dijana menggunakan kamus Bahasa Romania bagi menghasilkan senarai calon kesubjektifan yang berpotensi dengan menggunakan pencarian perkataan sinonim. Senarai calon perkataan tersebut akan dipilih berdasarkan ukuran persamaan *Latent Similarity Analysis* yang tertinggi dan kira-kira 4,000 kemasukkan digunakan untuk membangunkan pengelas berdasarkan peraturan kesubjektifan. Pengujian dilakukan dengan membandingkan antara piawai emas anotasi di peringkat ayat secara manual dengan anotasi secara automatik di peringkat ayat yang diberikan secara heuristik. Kajian ini telah mencapai keputusan ukuran-F bagi kesubjektifan sebanyak 66.20% dan keseluruhan ukuran-F adalah 61.69%.

Eskander & Rambow (2015) membangunkan leksikon sentimen dalam Bahasa Arab dengan menghubungkan pangkalan data leksikal AraMorph dengan SentiWordNet. AraMorph adalah sistem yang menganalisa morfologi dan melabel golongan kata bagi Bahasa Arab. Teknik yang digunakan adalah dengan memadankan perkataan yang terdapat dalam keterangan ringkas AraMorph dengan lema yang terdapat dalam SentiWordNet. Selain itu, golongan kata bagi kedua-dua sumber juga turut dipadankan. Padanan perkataan yang terhasil akan dimasukkan ke dalam senarai sama ada positif atau negatif bagi membentuk leksikon sentimen. Kajian ini dapat mencapai keputusan purata kejituan F1 sebanyak 68.6%.

2.5.2 Penterjemahan Kamus

Sumber leksikon dalam Bahasa Inggeris yang digunakan sebagai rujukan dalam penjanaan leksikon sentimen akan diterjemahkan ke bahasa sasaran menggunakan kamus elektronik atau kamus dalam talian. Kaedah ini turut digunakan dalam kajian bagi membangunkan leksikon sentimen dalam bahasa selain Bahasa Inggeris.

Franky et al. (2015) membangunkan leksikon sentimen dalam Bahasa Indonesia dengan menggunakan kaedah penterjemahan leksikon Bahasa Inggeris secara automatik. Kajian ini menggunakan 446 ayat yang dipilih secara rawak daripada laman web *KitaReview* dan melakukan anotasi seperti kesubjektifan atau keobjektifan dan polariti ayat oleh penutur Bahasa Indonesia. Sumber leksikon Bahasa Inggeris yang digunakan adalah leksikon pendapat Bing Liu, *Harvard General Inquirer*, MPQA dan SentiWordNet. Sumber ini akan diterjemahkan oleh penterjemah *Google, Moses* dan kamus dwi-bahasa. Leksikon baru yang dihasilkan dinilai dalam tugas praktikal bagi meramalkan polariti ayat. Hasil daripada penilaian tersebut memberikan keputusan kejituan yang tertinggi iaitu 74.64% adalah sumber leksikon yang diterjemahkan oleh terjemah *Google*.

Mihalcea et al. (2007) membangunkan leksikon sentimen dalam Bahasa Romania dengan kaedah terjemahan menggunakan kamus dwi-bahasa (*Bahasa Inggeris-Bahasa Romania).* Sumber leksikon kesubjektifan Bahasa Inggeris yang digunakan adalah *OpinionFinder* dan diterjemahkan ke Bahasa Romania bagi menghasilkan leksikon kesubjektifan atau sentimen dalam Bahasa Romania. Sebanyak 4,983 kemasukkan yang terdapat dalam leksikon kesubjektifan Bahasa Romania. Penilaian leksikon dilakukan secara manual iaitu dinilai oleh anotasi penutur Romania ke atas 150 kemasukkan kesubjektifan secara rawak.

2.5.3 Penggunaan Sumber Polariti

Pembangunan leksikon sentimen berasaskan sumber polariti seperti SentiWordNet (Esuli & Sebastiani 2006) turut digunakan oleh Dehkharghani et al. (2016) dan Perez-Rosas et al. (2012) dalam kajian mereka. Penggunaan SentiWordNet membolehkan polariti kata diberikan skor bagi memudahkan pengiraan sentimen ke atas frasa. Berikut adalah penerangan kajian yang telah dijalankan menggunakan sumber polariti seperti SentiWordNet.

Pembangunan leksikon sentimen dalam Bahasa Turki oleh Dehkharghani et al. (2016) menggunakan sumber Bahasa Inggeris iaitu WordNet 3.0 dan sumber polariti Bahasa Inggeris seperti SentiWordNet dan SenticNet. Manakala sumber Bahasa Turki yang digunakan pula adalah WordNet Turki yang dibangunkan

berasaskan WordNet 3.0 iaitu mengandungi *synset,* keterangan ringkas, sinonim, antonim dan juga hipernim. Bagi proses pengelasan *synset* Turki, SentiWordNet dan SenticNet digunakan sebagai fitur dan disari untuk mendapatkan skor positif dan negatif bagi setiap *synset*. Keputusan yang dicapai daripada pengujian dalam kajian ini adalah kejituan sebanyak 91.99% bagi ketiga-tiga pengelasan iaitu positif, negatif dan objektif.

Kajian oleh Perez-Rosas et al. (2012) memperlihatkan bahawa sumber polariti digunakan dalam menjana leksikon sentimen dalam Bahasa Sepanyol. Proses dimulakan dengan anotasi secara manual menggunakan leksikon *OpinionFinder* bagi mendapatkan perkataan sangat positif dan sangat negatif. Setelah itu, penganotasian ini akan dipadankan dengan WordNet 3.0 secara manual dan diterjemahkan secara padanan ke WordNet Bahasa Sepanyol. Dengan itu, anotasi secara automatik dapat dilakukan menggunakan SentiWordNet dengan mendapatkan skor positif dan negatif memandangkan *OpinionFinder* hanya menentukan polar perkataan sama ada positif dan negatif sahaja. Keputusan kejituan yang dicapai melalui kaedah ini ialah 74%.

2.6 SUMBER PEMBANGUNAN LEKSIKON SENTIMEN

Perkataan atau perbendaharaan kata sesuatu bahasa atau dikenali sebagai leksikal merupakan sumber penting dalam pembangunan leksikon sentimen. Terdapat banyak leksikal di dalam Bahasa Inggeris seperti WordNet (Miller et al. 1990), SentiWordNet (Esuli & Sebastiani 2006), MPQA (Deng & Wiebe 2015) dan OpinionFinder (Wiebe & Ellen 2005). Manakala dalam bahasa lain adalah seperti WordNet Bahasa dalam Bahasa Melayu (Nurril Hirfana et al. 2011), Bahasa Hindi (Saraswati et al. 2010), Bahasa Turki (Bilgin et al. 2004) dan Bahasa Sepanyol (Fernández-Montraveta et al. 2008).

2.6.1 WordNet Bahasa Melayu

Pembangunan prototaip WordNet Bahasa Melayu dikenali sebagai WordNet Bahasa dimulai oleh Tze & Hussein (2006). WordNet Bahasa dibangun berasaskan *Princeton Wordnet 1.6* dan Kamus Inggeris Melayu Dewan. Prototaip tersebut mengandungi 12,429 synset bagi kata nama dan 5,805 synset bagi kata kerja.

WordNet Bahasa (Nurril Hirfana et al. 2011) diperkembangkan lagi sebagai sumber bagi kajian semantik leksikal dalam Bahasa Melayu dan Bahasa Indonesia. Tujuan pengembangan WordNet Bahasa adalah sebagai sumber sokongan bagi perkembangan kegiatan analisa dalam Bahasa Melayu. Leksikal ini dibangunkan daripada gabungan pelbagai sumber seperti kamus Perancis-Inggeris-Melayu, Kamus Melayu-Inggeris (KAMI) dan sumber leksikal elektronik seperti WordNet Bahasa Inggeris, WordNet Bahasa Perancis dan WordNet Bahasa Cina. WordNet Bahasa disusun dengan menggabungkan kosa kata berdasarkan kosa kata Inggeris daripada Princeton WordNet 3.0. Leksikal ini mengandungi 49,668 *synsets*, 145,696 makna perkataan dan 64,431 perkataan unik. Jadual 2.4 menunjukkan data yang terdapat dalam pangkalan data leksikal WordNet Bahasa.

Bond et al. (2014) memperkembangkan WordNet Bahasa yang dijadikan sebagai sumber untuk kajian semantik leksikal dalam Bahasa Melayu. Dalam kajian tersebut, ketiga-tiga sumber WordNet Bahasa yang dibangunkan iaitu WordNet Melayu (Tze & Hussein 2006), WordNet Indonesia (Riza et al. 2010) dan WordNet Bahasa (Nurril Hirfana et al. 2011). Gabungan WordNet Bahasa tersebut mengandungi 48,000 konsep, 58,000 perkataan dalam Bahasa Indonesia dan 38,000 konsep dan 45,000 perkataan dalam Bahasa Malaysia.

Jadual 2.4 Contoh lema yang terdapat dalam WordNet Bahasa

Ofset	Bahasa	Ketepatan padanan	Lema
01586342-a	B	Y	baik
01438304-v	B	Y	hantar
01450178-a	B	O	kurang
05840431-n	B	O	sifat
14434681-n	B	O	penting

WordNet Bahasa adalah sumber pangkalan data leksikon Bahasa Melayu yang sedia ada digunakan buat masa ini. Namun begitu pangkalan data ini tidak mempunyai hubungan perkataan semantik sinonim dan antonim. (Bond et al. 2014).