PERLOMBONGAN PENDAPAT BAHASA ROJAK MENGGUNAKAN PENDEKATAN PEMBELAJARAN MESIN

NORLELA SAMSUDIN

DISERTASI YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT MEMPEROLEH IJAZAH
DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2013

PENGAKUAN

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

14 Oktober 2013

NORLELA SAMSUDIN
P 53570

PENGHARGAAN

Syukur Alhamdulillah kepada Allah S.W.T kerana memberikan saya kesihatan yang cukup, masa dan kematangan fikiran untuk menyiapkan kajian ini dalam bentuk sebegini rupa. Jutaan terima kasih yang rasanya tidak saya mampu untuk balas kembali hingga ke akhir hayat saya kepada penyelia utama Prof. Dr. Abdul Razak Hamdan atas bantuan yang begitu besar, bimbingan, teguran dan nasihat yang begitu berguna sepanjang kajian ini. Tidak lupa juga kepada penyelia bersama saya iaitu Prof. Madya Dr. Mazidah Puteh dan Dr Mohd Zakree Ahmad Nazri dengan kepakaran masing-masing yang banyak membantu menguatkan lagi semangat saya untuk menyiapkan kajian ini.

Ucapan terimakasih yang tidak terhingga saya ucapkan kepada pihak pengurusan Universiti Teknologi MARA, pihak pengurusan Universiti Teknologi MARA Caw. Terengganu dan Kementerian Pendidikan Tinggi yang memberi ruang, peluang dan membantu dari segi kewangan dari awal pengajian sehinggalah saya menamatkan kajian ini.

Ucapan terimakasih yang tidak terhingga juga saya ucapkan kepada suami tercinta En. Roslan Rani, ibu tersayang Pn Rahmah Ali dan anak Mohamed Haqim Roslan yang banyak menyokong dan membantu saya dari setiap segi disepanjang pengajian ini. Doa, pengorbanan, kasih sayang dan bantuan mereka menguatkan semangat saya untuk menamatkan pengajian.

Ucapan terimakasih juga saya hulurkan kepada pembantu penyelidik yang membantu saya mengumpul data, rakan-rakan, saudara mara dan individu-individu yang terlibat secara langsung atau tidak langsung dalam kajian ini. Hanya Allah yang dapat membalas jasa dan budi yang telah diberikan.

ABSTRAK

Penyampaian pendapat dalam talian telah menjadi perkara lumrah pada masa kini. Di Malaysia, pendapat ini ditulis dalam bahasa rojak yang menggunakan perkataan atau singkatan dari bahasa Melayu, bahasa Inggeris dan bahasa daerah. Oleh yang demikian, perlombongan pendapat dengan menggunakan kaedah pemprosesan bahasa (NLP) sukar dilaksanakan. Manakala proses perlombongan teks menggunakan kaedah pembelajaran mesin pula, menganggap semua perkataan dalam sesuatu mesej sebagai fitur. Tanpa proses pemilihan fitur yang berkesan, proses perlombongan pendapat memerlukan sumber yang banyak dan masa yang lama. Pada masa ini, pemilihan fitur dalam kategori penapisan seperti kekerapan dalam mesej, CHI Square dan kebolehcapaian maklumat hanya menyusun fitur berdasarkan nilai pemberat. Terpulanglah kepada pengguna untuk memilih fitur yang bersesuaian. Justeru itu, objektif kajian ini adalah untuk mencadangkan satu kerangka perlombongan pendapat berorientasikan pembelajaran mesin yang berkesan untuk melombong pendapat mesej dalam talian di Malaysia. Kerangka perlombongan pendapat ini akan mengambil kira dua isu utama iaitu i) memperkenalkan kaedah yang bersesuaian bagi proses penormalan teks hingar yang dikenali sebagai MyTNA dan ii) memperkenalkan kaedah pemilihan fitur berinspirasi dari teori rangkaian dalam Sistem Imun Buatan yang dikenali sebagai FS-INS. Metodologi kajian ini dibahagikan kepada empat fasa iaitu i) analisis kerangka perlombongan pendapat bahasa rojak ii) pengumpulan dan penyediaan data, iii) pembangunan MyTNA serta FS-INS, dan iv) pengujian keberkesanan kaedah yang dicadangkan. Fasa (i) dilaksanakan menerusi kajian literasi. Dalam Fasa (ii), beberapa senarai rujukan bagi proses penormalan teks hingar diperkenalkan. Selain itu, teks eksperimen yang terdiri dari 1000 pendapat positif dan 1000 pendapat negatif mengenai tayangan wayang di Malaysia juga diekstrak dalam fasa ini. Algoritma yang berkesan dirangka dan dibangunkan dalam Fasa (iii). Akhir sekali, eksperimen yang membandingkan kaedah pemilihan fitur cadangan dengan beberapa kaedah pemilihan fitur sedia ada dilaksanakan dalam Fasa (iv). Eksperimen terhadap kerangka perlombongan pendapat ini menunjukkan pengurangan sehingga 90% bilangan fitur dan berjaya meningkatkan ketepatan perlombongan pendapat sehingga 14.9% apabila pengelasan k Jiran Terdekat digunakan sebagai model pengelasan. Keputusan perlombongan pendapat dengan menggunakan pengelasan *Naïve Bayes* dan *Support Vector Machine* juga meningkat antara 6% hingga 9%. Kajian ini juga menghasilkan sebuah korpus yang terdiri dari 20,100 mesej-mesej dalam talian yang diekstrak dari forum dalam talian, aplikasi Twitter dan aplikasi Facebook.

OPINION MINING FOR BAHASA ROJAK USING MACHINE LEARNING APPROACH

ABSTRACT

Nowadays, it is normal to express opinions through online applications. In Malaysia, the opinions are expressed using various words and abbreviations in mix languages using the Malay words, the English words and local dialects. As a result, it is difficult to execute opinion mining process using natural language processing (NLP) approach. On the other hand, opinion mining using machine learning approach incorporates all words in a message as features. Without effective activity to reduce the number of these features, the opinion mining process requires more resources and longer time to complete. Current feature selection techniques in filter category such as Document Frequency, CHI Square and Information Gain will sort all features based on certain calculated values. It is up to the user to select the appropriate features based on the calculated values. Therefore, the objective of this study was to introduce suitable opinion mining framework based on the machine learning approach to mine opinions from online messages that are created in Malaysia. The activities of the framework will concentrate of two main issues i.e. (i) normalization of noisy texts using algorithm known as MyTNA and (ii) introduction of a feature selection technique named FS-INS that was inspired from Artificial Immune Network theory. The methodology adopted in this study was divided into 4 phases i.e. (i) development of suitable framework for opinion mining using mixed languages, ii) data collection and data preparation (iii) development of methods to normalize noisy text and FS-INS, and iv) testing and evaluation suggested methods. Phase (i) involved literature study. In Phase (ii) several references for normalization of noisy texts were introduced. Other than that, 1000 positive movie reviews and 1000 negative movie reviews were also collected in this phase. The algorithm for normalization of noisy texts named MyTNA and FS-INS algorithm were developed in Phase (iii). Lastly, several experiments were conducted to compare the performance of suggested feature selection method with several currently used feature selection methods in Phase (iv). Experiments using the new opinion mining framework indicated a reduction up to 90% of the features. Performing normalization of noisy text and FS-INS feature selection activities had also successfully improved accuracy of opinion mining by 14.9% in k-Nearest Neighbor classifier. Similarly, the accuracy of opinion mining processing using Naïve Baiyes and Support Vector Machine had been increased by 6% and 9% respectively. Other than that, an online corpus, consists of 20,100 online messages were also introduced in this study.

KANDUNGAN

SENARAI JADUAL

SENARAI ILUSTRASI

SENARAI SINGKATAN

Singkatan	Bahasa Inggeris	Bahasa Melayu
ab	Antibody	Antibodi
AIN	Artificial Immune Network	Rangkaian Imun Buatan
AIS	Artificial Immune System	Sistem Imun Buatan
ASR	Automatic Speech Recognition	Kenal pasti bunyi percakapan secara automatik
AT	Affinity Threshold	Ambang Afiniti
BIC	Biology Inspired Computing	Pengkomputeran Berasaskan Biologi
BOW	Bag of Word	Sehimpun perkataan
CHI	CHI Square	Gandaan CHI
DF	Document Frequency	Kekerapan Perkataan Dalam Mesej
DBP	-	Dewan Bahasa dan Pustaka
FS-INS	Feature Selection Based on Immune Network System	Pemilihan Fitur Berdasarkan Sistem Rangkaian Imun
GA	Genetic Algorithm	Algoritma Genetik
IG	Information Gain	Kedapatan Maklumat
KMDT	Corpus of Online Messages	Korpus Mesej Dalam talian
kNN	K Nearest Neighbor	K Jiran Terdekat
MC	Memory Cell	Sel memori
MyTNA	Malay Text Normalization Algorithm	Algoritma Penormalan Teks Bahasa Melayu
NB	Naïve Bayes	-
NLP	Natural Language Processing	Teknik Pemprosesan Bahasa
OOV	Out of Vocabulary	Tiada dalam kamus
POS	Part Of Speech	Sebahagian pertuturan
SMS	Short Message Service	Khidmat Pesanan Ringkas
SPMDT	Online Message Translation System	Sistem Penterjemahan Mesej Dalam talian
SVM	Support Vector Machine	-
TH	Noisy text	Teks hingar

SENARAI ISTILAH

Istilah (bahasa Melayu)	Istilah (bahasa Inggeris)
Akronim	Acronym
Algoritma Imun	Immune Algorithm
Ambang	Threshold
Analisis sentimen	Sentiment Analysis
Aplikasi	Application
Atribut	Attribute
Dalam talian	Online
Fitur	Feature
Fonetik	Phonetic
k-Jiran Terdekat	k-Nearest Neighbor
Kerangka	Framework
Lipatan	Fold
Mekanisme pembelajaran	Learning mechanism
Pangkalan data	Database
Pemangkasan	Stemming
Pembelajaran mesin	Machine Learning
Penormalan	Normalization
Pengelasan	Classification
Pengelompokan	Clustering
Pengukuran affinity	Affinity measurement
Penukaran ke perkataan dasar	Lemmatization
Perlombongan pendapat	Opinion Mining
Perlombongan teks	Text Mining
Perwakilan data	Data representation
Pilihan negatif	Negative Selection
Populasi	Population
Rangkaian Imun	Immune Network
Rentetan	String
Sel memori	Memory cell
Singkatan	Abbreviation
Teks Hingar	Noisy Text
Teori bahaya	Danger Theory

BAB 1

PENDAHULUAN

1.1 PENGENALAN

Teknologi internet yang dilancarkan pada awal tahun 1990-an telah berkembang dengan pesat dan mempengaruhi setiap sudut kehidupan sejagat. Salah satu bidang ilmu yang berkembang selaras dengan perkembangan Internet ialah perlombongan pendapat yang juga dikenali sebagai analisis sentimen (Pang & Lee 2008). Penyelidikan ini tertumpu kepada penerokaan ilmu baru dalam bidang tersebut. Bab ini menerangkan latar belakang dan permasalahan yang cuba diselesaikan oleh penyelidik. Seksyen 1.2 menerangkan latar belakang kajian. Seksyen 1.3 pula membincangkan permasalahan kajian. Berdasarkan permasalahan tersebut, Seksyen 1.4 menyenaraikan objektif dalam menangani permasalahan kajian. Skop kajian pula dinyatakan dalam Seksyen 1.5. Ringkasan metodologi kajian dinyatakan dalam Seksyen 1.6. Seksyen 1.7 menjelaskan kepentingan kajian. Ringkasan hasil kajian dinyatakan dalam Seksyen 1.8. Definisi perkataan utama dibincangkan dalam Seksyen 1.9. Bab ini diakhiri dengan ringkasan kandungan di setiap bab yang dijelaskan dalam Seksyen 1.10.

1.2 LATAR BELAKANG KAJIAN

Pendapat atau sentimen adalah pernyataan yang menyatakan penilaian seseorang terhadap sesuatu perkara. Perlombongan pendapat adalah satu proses untuk mendapatkan maklumat berkaitan pendapat-pendapat tertentu yang berada dalam sekumpulan mesej. Perkembangan teknologi Internet telah meningkatkan kaedah perkongsian pendapat secara dalam talian melalui elektronik forum dan aplikasi sosial seperti Facebook dan Twitter. Selain itu, ruang maklum balas pelanggan adalah salah satu elemen yang perlu ada dalam setiap laman web sesuatu organisasi. Dengan melombong pendapat daripada

maklum balas dalam talian, peniaga dapat mengenal pasti sejauh mana produk atau servis yang ditawarkan boleh diterima oleh pengguna (Archak et al. 2007; Lee et al. 2008; Wang & Ren 2007;). Pada masa ini, pendapat positif atau pendapat negatif mengenai sesuatu perkara boleh tersebar dengan meluas dalam masa yang singkat. Justeru itu, pihak pengurusan syarikat perlu menganalisis sentimen terhadap produk mereka dengan segera. Teknologi perlombongan pendapat juga boleh digunakan bagi menyalurkan maklum balas pelanggan kepada bahagian atau unit yang terlibat secara automatik agar ia dapat diproses dengan segera. Maklumat yang diekstrak dari pengguna ini dapat membantu pihak pengurusan sesuatu organisasi untuk membuat keputusan dengan tepat dan cepat. Keupayaan tersebut akan membantu meningkatkan profit seterusnya mengekalkan organisasi dalam perniagaan.

Di pihak pelanggan pula, mereka boleh menilai pendapat pengguna-pengguna yang terdahulu sebelum membuat keputusan untuk membeli sesuatu produk atau melanggan sesuatu servis. Ramai penyelidik terdahulu telah memberikan tumpuan untuk mendapatkan pendapat dari pengguna terdahulu. Penyelidik seperti Pang et al. (2002) dan Zheng & Ye (2009) menjalankan kajian mengenai melombong maklum balas dari pelanggan. Dey & Haque (2008), Gamon (2004), Hu & Liu (2004) , dan Stoyanov & Cardie (2006) pula menghasilkan ringkasan pendapat pengguna terdahulu bagi memudahkan penilaian terhadap sesuatu produk atau servis. Pendapat mengenai sesuatu isu yang diutarakan oleh kumpulan berhaluan kiri pula menjadi bahan kajian kepada Abbasi (2008), Chen & Salem (2008), Tchalakova (2007). Manakala Hongfei & Zhihao (2007) dan Stavrianou & Chauchat (2008) pula melihat penggunaan perlombongan pendapat bagi mengemas kini dan meningkatkan penggunaan pembelajaran secara dalam talian.

Dilaporkan, pada 30 Jun 2011, 60.7% atau 17.7 juta rakyat Malaysia menggunakan Internet (internetworldstats 2012). Facebook adalah aplikasi dalam talian yang paling kerap digunakan oleh rakyat Malaysia. Selain itu, laman dan aplikasi yang membolehkan pengguna berinteraksi di antara satu sama lain seperti *blogger.com*, *cari.com*, *mudah.com* dan Twitter adalah di antara sepuluh perkhidmatan elektronik tertinggi yang digunakan oleh rakyat Malaysia (Alexa.com 2013) . Situasi ini mewujudkan berbagai maklumat yang boleh dilombong dari mesej dalam talian yang diwujudkan oleh pengguna di Malaysia. Walau bagaimanapun, perlombongan pendapat kurang mendapat perhatian dari penyelidik di Malaysia. Sehingga tesis ditulis, penulis masih belum berjumpa kajian yang

melibatkan perlombongan pendapat menggunakan bahasa rojak yang sering digunakan di mesej dalam talian. Kajian ini menyumbang ke arah peningkatan aktiviti perlombongan pendapat di mesej dalam talian yang dihasilkan oleh pelanggan di Malaysia dengan memperkenalkan kerangka perlombongan pendapat mesej yang mengandungi bahasa rojak. Kerangka ini akan menggunakan algoritma baru dikenali sebagai MyTNA untuk menormalkan teks hingar dan algoritma FS-INS berdasarkan sistem imun buatan untuk memilih satu set fitur yang bersesuaian untuk aktiviti perlombongan pendapat menggunakan pendekatan pembelajaran mesin.

1.3 PERMASALAHAN KAJIAN

Berikut adalah contoh maklum balas yang diambil dari mesej dalam talian mengenai tayangan sesuatu filem di Malaysia:

a) *"oh bestnya, best giler serius. Nak kasik 5 bintang plus2"*
b) *Aku bg 4.9 out of 5stars.Yg 0.1 xcukup to sbb aku xfaham.. masa aku tgk aritu pon xfull"*
c) *Ksian ngan kawan aku. Coz abihkan duit utk film nih"*

Contoh-contoh tersebut menunjukkan dengan jelas ciri-ciri mesej dalam talian seperti berikut:

1) **Peratus penggunaan teks hingar yang tinggi** (Wong et al. 2006; Dev & Haque 2008). Penggunaan perkataan yang tiada dalam kamus (OOV) adalah tinggi. Sebagai contoh, hanya beberapa perkataan sahaja yang betul ejaannya di maklum balas (a) iaitu *oh, best* dan *bintang*.
2) Perkataan yang digunakan adalah **campuran perkataan dalam bahasa Inggeris dan bahasa Melayu,** atau lebih dikenali dengan **bahasa rojak**. Selain itu, gabungan suku kata dari kedua-dua bahasa juga digunakan. Perkataan-perkataan seperti *bestnya*, *plus2* dan *xfull* tidak wujud sama ada dalam koleksi perkataan bahasa Inggeris atau bahasa Melayu. Walau bagaimanapun, perkataan-perkataan ini digunakan dengan meluas dalam mesej-mesej elektronik ini.
3) Ayat yang dibentuk **tidak mengikut struktur pengendalian bahasa** yang ditetapkan seperti penggunaan simbol berhenti yang salah, penggunaan huruf besar dan huruf kecil yang tidak kemas dan tidak menggunakan nahu bahasa yang betul.

4) Penggunaan **ejaan perkataan yang tidak mengikut ejaan** suku kata tetapi mengikut ejaan fonetik seperti perkataan *giler* dan *abiskan* dalam maklum balas (c).
5) **Pendapat dinyatakan secara tidak langsung**. Sebagai contoh, dalam maklum balas (c) tiada perkataan seperti *'tidak bagus'* atau *'booring'* yang menggambarkan pendapat negatif digunakan.

Terdapat dua kaedah bagaimana perlombongan pendapat dilaksanakan iaitu kaedah pemprosesan bahasa dan kaedah pembelajaran mesin. Dalam kaedah yang pertama, struktur pembinaan ayat bagi setiap perkataan akan dikenal pasti. Setiap perkataan akan di label dengan struktur nahu bahasa tertentu seperti *frasa kata nama, frasa kata adjektif* atau *frasa kata kerja.* Kaedah ini menggunakan teknik pemprosesan bahasa (NLP) yang tertentu seperti pembahagian kepada unsur-unsur kecil (*tokenization),* pelabelan nahu bahasa (*part of speech (POS)),* pemangkasan perkataan *(stemming)* dan penukaran kepada kata dasar (*lemmatization)*. Teknik ini digunakan oleh penyelidik seperti Nasukawa & Yi (2003), Turney (2002) dan Wang & Ren (2007) .

Walau bagaimanapun, teknik ini tidak sesuai untuk digunakan dengan mesej dalam talian yang mengandungi bahasa rojak. Ini kerana mesej dalam talian mengandungi peratus teks hingar yang tinggi dan menggunakan istilah bahasa Melayu yang bercampur aduk dengan bahasa Inggeris dan bahasa dialek. Selain itu, pembentukan ayat dalam mesej ini tidak mengikut kaedah pembentukan ayat yang betul sama ada dalam bahasa Inggeris atau dalam bahasa Melayu. O'Neill (2009) menyuarakan kesukaran ini dalam kenyataan berikut:

> *One drawback of an NLP based approach is that it would likely perform very poorly when used on grammatically incorrect text… methods to detect and possibly correct bad English would be necessary before use on a large scale*.

Kaedah kedua pula melibatkan penggunaan aktiviti-aktiviti perlombongan teks yang digunakan oleh penyelidik seperti Pang et al. (2002), Pang & Lee (2005), Tsitsumia et al. (2007) dan Zheng & Ye (2009). Kaedah ini menggunakan proses-proses yang sama seperti proses pengelasan teks untuk mengelas mesej-mesej kepada kelas positif atau kelas negatif. Kaedah perlombongan pendapat yang menggunakan kaedah pembelajaran

mesin melibatkan bilangan fitur yang tinggi. Ini berlaku kerana proses pengelasan teks akan menganggap setiap perkataan dalam mesej sebagai atribut atau fitur. Keadaan ini memerlukan masa pemprosesan yang lama dan memerlukan ruang pemprosesan yang banyak. Pada masa ini Kebanyakan kajian-kajian perlombongan pendapat yang menggunakan kaedah pembelajaran mesin menggunakan teknik pemilihan fitur yang dikategorikan sebagai penapisan (*filter*) (Boiy & Moens 2009; Dave et al. 2003; Pang et al. 2002). Melalui kaedah ini, sub set fitur yang dianggap berkait dengan pendapat dipilih berdasarkan kriteria tertentu tanpa mengambil kira kaitan perkataan dengan model pengelasan. Kekerapan perkataan dalam mesej (DF), kedapatan maklumat (IG) dan CHI Square (CHI) adalah contoh teknik pemilihan fitur dalam kategori penapisan. Terdapat dua kelemahan pada teknik pemilihan fitur semasa iaitu

a) Isu sentimen tidak diambil kira sewaktu pemilihan fitur.
b) Satu nilai diberikan kepada setiap fitur berdasarkan formula tertentu. Fitur-fitur kemudian disusun berdasarkan nilai-nilai tersebut. Sekiranya pengatur cara tidak menetapkan bilangan fitur yang dipilih, semua fitur akan digunakan dalam proses perlombongan pendapat. Pada masa ini pemilihan fitur tidak dilaksanakan secara automatik.

Bagi memilih perkataan yang berkait dengan sentimen, penyelidik-penyelidik terdahulu menggunakan kaedah-kaedah berikut sebelum aktiviti pemilihan fitur dilaksanakan

a) Menggunakan proses POS untuk melabel struktur nahu setiap ayat dan memilih perkataan-perkataan yang tergolong dalam struktur nahu tertentu seperti <Kata Adjektif> atau <Kata Kerja>.
b) Menggunakan pangkalan-pangkalan data tertentu seperti WordNet atau SentiWordNet untuk memilih perkataan-perkataan yang menyatakan pendapat seperti bagus, baik dan lain-lain

Kajian literatur menunjukkan tiada kajian perlombongan pendapat yang melibatkan mesej dalam talian khususnya mesej yang diwujudkan oleh rakyat Malaysia. Kajian lepas berkaitan penggunaan bahasa rojak tertumpu pada kesan negatif hasil penggunaan bahasa rojak terhadap penggunaan Bahasa Melayu (Husni 2009, Latisha & Surina 2013) dan jati diri bangsa Melayu (Zaharani et al. 2011). Penyelidik berpendapat, ciri-ciri mesej dalam talian menyebabkan tiada kajian perlombongan pendapat terhadap mesej dalam talian. Kandungan teks hingar dalam mesej-mesej ini perlu dibetulkan sebelum sebarang proses perlombongan pendapat dilaksanakan. Aktiviti ini akan

mengurangkan bilangan perkataan yang membawa maksud yang sama. Sebagai contoh istilah *tak, dak, tk, x, tidaakk* digunakan di mesej dalam talian bagi mewakili istilan *tidak*. Bagi melombong pendapat menggunakan teknik pemprosesan bahasa, setelah teks hingar dinormalkan, perkataan-perkataan yang berada dalam mesej tersebut perlu diubah atau disusun mengikut struktur pembentukan ayat yang betul bagi membolehkan setiap perkataan di label dengan struktur nahu bahasa yang betul. Sehingga kajian ini dilaksanakan, penyelidik tidak berjumpa kajian yang boleh mengubah mesej dalam talian yang mengandungi bahasa rojak kepada ayat yang menggunakan istilah dan struktur pembentukan ayat yang betul. Oleh yang demikian, kaedah pembelajaran mesin adalah lebih sesuai untuk melombong pendapat mesej dalam talian yang mengandungi bahasa rojak.

Permasalahan apabila kaedah pembelajaran mesin digunakan adalah teknik pemilihan fitur yang tidak automatik dan tidak mengambil kira isu sentimen sewaktu memilih set fitur. Selain itu, rujukan yang serupa seperti *WordNet* dan *SentiWordNet* yang digunakan oleh penyelidik-penyelidik yang melombong pendapat dalam bahasa Inggeris tidak wujud dalam bahasa Melayu. Oleh yang demikian terdapat keperluan untuk:

a) Mewujudkan satu kerangka perlombongan pendapat menggunakan teknik pembelajaran mesin bagi melombong pendapat di mesej dalam talian yang mengandungi bahasa rojak.
b) Mewujudkan satu kaedah penormalan teks hingar yang sesuai bagi mesej dalam talian yang menggunakan bahasa rojak; dan
c) Mewujudkan satu kaedah pemilihan fitur untuk melombong pendapat dari mesej dalam talian. Kaedah ini perlu mengambil kira isu sentimen dan tidak merujuk kepada kamus-kamus tertentu.

Berikut adalah persoalan dalam kajian ini:

1) Apakah kerangka perlombongan pendapat yang sesuai bagi mesej dalam talian yang menggunakan bahasa rojak?
2) Apakah kaedah penormalan teks hingar yang sesuai bagi mesej dalam talian yang menggunakan bahasa rojak?
3) Apakah kaedah pemilihan fitur yang bersesuaian dengan perlombongan pendapat dari mesej dalam talian?

1.4 OBJEKTIF KAJIAN

Objektif kajian ini adalah untuk

1) Mencadangkan satu kerangka perlombongan pendapat, kaedah pembelajaran mesin bagi melombong mesej dalam talian yang mengandungi bahasa rojak.
2) Mencadangkan satu kaedah penormalan teks hingar yang wujud di mesej dalam talian; dan
3) Mencadangkan satu kaedah pemilihan fitur bagi perlombongan pendapat berdasarkan Sistem Imun Buatan dengan mengambil kira faktor sentimen yang wujud di sesuatu mesej.

1.5 SKOP KAJIAN

Skop kajian ini adalah seperti berikut:

1) Kajian ini tertumpu pada mewujudkan kerangka perlombongan pendapat menggunakan kaedah pembelajaran mesin dan kaedah pemilihan fitur yang relevan.
2) Kajian ini menggunakan mesej dalam talian yang diekstrak dari elektronik forum, aplikasi Facebook dan aplikasi Twitter. Mesej-mesej ini dipercayai diwujudkan oleh rakyat Malaysia dan mengandungi perkataan-perkataan yang dikategorikan sebagai bahasa rojak.
3) Penilaian keberkesanan proses dilaksanakan dengan menjalankan eksperimen perlombongan pendapat menggunakan teknik yang sering digunakan iaitu *Naïve Bayes (NB), k Nearest Neighbour (kNN)* dan *Support Vector Machine (SVM).*

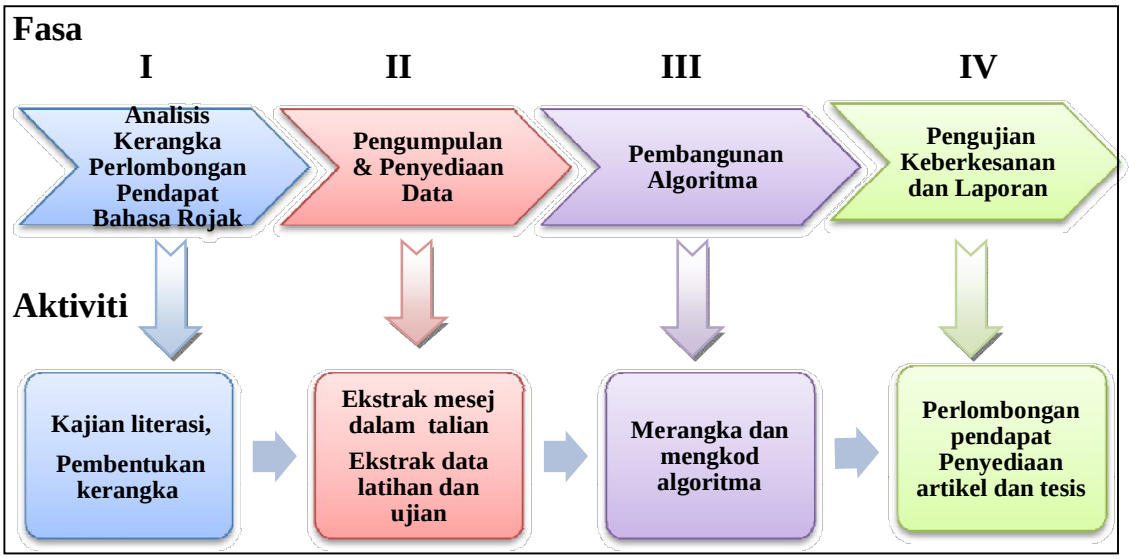
1.6 METODOLOGI KAJIAN

Seksyen ini menjelaskan secara ringkas, aktiviti-aktiviti yang dilaksanakan bagi menyelesaikan permasalahan dan mencapai objektif-objektif kajian yang dibahagikan kepada empat fasa iaitu

I. Analisis Kerangka Perlombongan Pendapat Bahasa Rojak;
II. Pengumpulan & Penyediaan Data;
III. Pembangunan Algoritma; dan

IV. Pengujian Keberkesanan dan Laporan.

Rajah 1.1 memaparkan secara ringkas aktiviti dan hasil di setiap fasa kajian. Fasa pertama melibatkan analisis kerangka perlombongan pendapat bahasa rojak yang dicapai melalui analisis kajian-kajian lepas yang berkaitan dengan perlombongan pendapat, penormalan teks hingar dan pemilihan fitur.



Rajah 1.1 Ringkasan metodologi utama kajian

Fasa kedua melibatkan pengumpulan dan penyediaan data. Dua jenis data diekstrak dari forum dalam talian, aplikasi Facebook dan aplikasi Twitter. Data tersebut adalah mesej dalam talian yang membentuk Korpus Mesej Dalam Talian (KMAT) dan maklum balas positif serta maklum balas negatif yang berkait dengan tayangan filem di Malaysia sebagai data eksperimen.

Fasa ketiga melibatkan pembangunan algoritma dan prototaip bagi kaedah penormalan teks hingar bahasa rojak dan kaedah pemilihan fitur yang berkesan untuk melombong pendapat dari mesej dalam talian.

Fasa terakhir melibatkan ujian keberkesanan kedua-dua kaedah yang dicadangkan dalam proses perlombongan pendapat. Hasil eksperimen dibincangkan dalam artikel-artikel yang dibentangkan di seminar dan diterbitkan dalam jurnal. Akhir sekali, hasil kajian direkodkan dalam tesis PhD. Lampiran A menunjukkan ringkasan kandungan kajian yang dilaksanakan.

1.7 KEPENTINGAN KAJIAN

Kajian ini memperkenalkan satu kerangka perlombongan pendapat dengan menggunakan bahasa rojak. Perlombongan pendapat mesej dalam kajian yang diwujudkan oleh rakyat Malaysia masih belum diterokai sehingga kajian ini dilaksanakan. Kewujudan teks hingar yang tinggi dan kekurangan rujukan bagi mengenal pasti perkataan-perkataan yang menyatakan sentimen menyumbang kepada situasi tersebut. Bagi menangani isu teks hingar yang tinggi, kajian ini memperkenalkan algoritma MyTNA bagi menterjemah kesilapan ejaan yang sering berlaku di mesej dalam kajian yang mengandungi bahasa rojak. Bagi mengatasi masalah kekurangan rujukan untuk mengenal pasti sentimen, kerangka perlombongan pendapat menggunakan kaedah pembelajaran mesin digunakan. Bagi memilih fitur-fitur yang relevan dengan perlombongan pendapat, algoritma FS-INS diperkenalkan yang menggunakan pendekatan berdasarkan sistem imun tabii untuk memilih fitur. Kajian ini mendapati penggunaan kerangka perlombongan pendapat yang dihasilkan menghasilkan keputusan ketepatan perlombongan pendapat yang lebih baik berbanding dengan keputusan ketepatan perlombongan pendapat tanpa menggunakan MyTNA dan FS-INS.

Selain itu, korpus mesej dalam talian yang terdiri dari 20,000 mesej dalam talian yang diekstrak secara rawak dari forum elektronik, aplikasi Facebook dan aplikasi Twitter boleh digunakan sebagai data dalam kajian yang melibatkan mesej dalam talian selain dari perlombongan pendapat seperti kajian kaedah pembentukan singkatan yang digunakan dalam mesej dalam talian dan penggunaan ikon dan simbol yang mewarnai mesej dalam talian. Data juga boleh digunakan dalam kajian yang melihat variasi kaedah berkomunikasi secara dalam talian. Korpus ini adalah yang pertama seumpamanya di Malaysia.

Kepesatan perkembangan teknologi internet telah mempengaruhi kehidupan rakyat Malaysia. Adalah diharapkan kajian ini akan menggalakkan lebih banyak kajian dalam bidang perlombongan pendapat dan perlombongan teks di kalangan pengkaji-pengkaji di Malaysia.

1.8 RINGKASAN HASIL DAN SUMBANGAN KAJIAN

Berikut adalah senarai hasil dan sumbangan kajian.

1. Kerangka perlombongan pendapat mesej dalam talian yang mengandungi bahasa rojak.
2. Algoritma penormalan teks hingar yang diwujudkan oleh pengguna dalam talian di Malaysia
3. Algoritma pemilihan fitur berasaskan sistem imun buatan bagi memilih fitur-fitur yang relevan untuk proses perlombongan pendapat.
4. Korpus dalam talian yang diekstrak dari elektronik forum, aplikasi Facebook dan aplikasi Twitter secara rawak untuk digunakan sebagai kajian-kajian ilmiah.

1.9 DEFINISI ISTILAH UTAMA

Memahami maksud perkataan utama memudahkan pemahaman topik-topik yang dibincangkan dalam tesis ini. Oleh yang demikian, seksyen ini menjelaskan definisi perkataan-perkataan utama yang digunakan dalam kajian ini seperti pendapat, bahasa rojak, teks hingar, pemilihan fitur dan sistem imun buatan.

Pendapat

Pendapat, sentimen, emosi dan spekulasi merujuk kepada suatu keadaan peribadi yang tidak dapat dilihat dengan mata kasar. Pendapat-pendapat mengenai sesuatu produk, servis atau individu dinyatakan dalam mesej dengan menggunakan perkataan-perkataan tertentu yang dikenali sebagai bahasa subjektif. Perkataan-perkataan seperti *suka*, *tidak suka*, *best* dan *seronok* adalah contoh perkataan yang menyatakan pendapat atau sentimen. Bahasa objektif pula digunakan apabila menyatakan sesuatu fakta (Wilson & Wiebe 2003) . Pendapat juga merujuk kepada pandangan seseorang individu terhadap sesuatu perkara pada sesuatu masa berdasarkan pengalaman yang dilaluinya. Kebiasaannya pendapat ini diperkukuhkan lagi dengan kehadiran fakta-fakta tertentu (Jain et. Al. 2012) .

Pendapat boleh dinyatakan secara langsung/tersurat (*explicit*) seperti '*Saya suka filem ini.*' atau tersirat *(implicit*) seperti '*Filem ini mendapat sambutan kerana lakonan Erra Fazira*.' Pendapat yang dinyatakan secara langsung menggunakan perkataan-perkataan tertentu seperti *'suka', ' baik'* atau *'tidak suka'* untuk menyatakan pandangan

penulis. Kebiasaannya perkataan-perkataan dalam kategori adjektif digunakan dalam menyatakan pendapat secara langsung. Sementara itu kebiasaannya, tiada perkataan-perkataan sentimen digunakan dalam pernyataan pendapat secara tidak langsung.

Bahasa Rojak

Dalam Kamus Dewan Bahasa dan Pustaka, bahasa rojak didefinisikan sebagai "bahasa bercampur aduk yang digunakan dalam penulisan dan lisan ". Bahasa rojak juga dikenali sebagai bahasa pasar (Karim et al. 2010). Ia adalah ragam bahasa yang digunakan oleh orang Melayu dengan bangsa lain. Dalam konteks mesej dalam talian, bahasa rojak melibatkan penggunaan perkataan bahasa Melayu dan bahasa Inggeris. Bahasa rojak juga didefinisikan sebagai penggunaan kombinasi perkataan-perkataan dari dua atau lebih bahasa dalam komunikasi di mana satu bahasa menjadi bahasa utama. Bahasa rojak juga dikenali sebagai Manglish (Malaysian English) di mana perkataan dari bahasa Hokkien, bahasa Melayu, Bahasa Mandarin dan Bahasa Telagu digabungkan dalam bahasa Inggeris (Husni 2009, Latisha & Surina 2013). Roosfa (2004) pula menganggap bahasa rojak sebagai "campuran bahasa Melayu dan Inggeris dan disulam pula dengan bahasa pasar, bahasa kasar dan bahasa yang mencarut".

Kajian ini menggunakan takrifan yang diberikan oleh Majed (2011) dalam mengenal pasti perkataan-perkataan dalam kategori bahasa rojak. Majed (2011) memperluaskan definisi bahasa rojak dengan meliputi singkatan seperti *'nak'*, *'mau'* dan *'tak'* serta bahasa remaja seperti *'awek'*, *'kantoi'* dan *'otai'*. Bahasa rojak dituturkan dalam situasi tidak formal dan tidak mengikut peraturan tertentu. Oleh kerana komunikasi dalam talian dikategorikan sebagai tidak formal, penggunaan bahasa rojak adalah lebih digemari dan digunakan dengan meluas. Isu-isu berkaitan bahasa rojak dikupas dalam laman-laman web seperti http://kodrimohd.blogspot.com/2012/03/bahasa-rojak.html dan http://www.slideshare.net/jongakeling/present-bahasa-rojak. Penggunaan bahasa rojak dianggap sebagai cabaran dalam mempertahankan penggunaan bahasa Melayu sebagai bahasa kebangsaan (Anon 2012a, Anon 2012b, Ruhaiza 2012).

Teks Hingar

Sesuatu perkataan diklasifikasikan sebagai teks hingar apabila ia berlainan dari perkataan yang sebenarnya bagi menyatakan sesuatu maksud (Subramaniam et al. 2009) . Dengan maksud yang hampir sama, Knoblock et al. (2007) mentakrifkan teks hingar sebagai *"any kind of difference between the surface form of a coded representation of the text and*

the intended, correct, or original text." Teks hingar terbentuk sewaktu sesuatu mesej dihasilkan sama ada melalui:

a) Proses penterjemahan dari sistem percakapan kepada teks;
b) Proses mengimbas dan diubah dari bentuk imej kepada teks melalui sistem OMR; atau
c) Pembentukan mesej dalam talian seperti menggunakan forum dalam talian, khidmat pesanan ringkas (SMS), blog, Facebook atau Twitter.

Sebelum tahun 2000, kebanyakan kajian yang melibatkan pemprosesan teks hingar menumpukan kajian terhadap mesej yang terhasil dari proses (a) dan (b) (Kernighan et al. 1990). Mulai tahun 2005, kajian teks hingar melibatkan data dari khidmat pesanan ringkas (Aw et al. 2006). Kini kajian mula menggunakan mesej-mesej yang dibentuk dengan aplikasi dalam talian seperti forum elektronik serta mesej-mesej di Facebook dan Twitter (Clark & Araki 2011, Pek & Paroubek 2010, Laboreiro et al. 2010).

Dalam konteks mesej dalam talian, kewujudan teks hingar berkait rapat dengan penggunaan bahasa rojak. Penggunaan peranti yang tidak bersesuaian dan mod komunikasi yang tidak formal adalah di antara sebab bahasa rojak digunakan dalam mesej dalam talian. Berikutan dari itu, kehadiran teks hingar dalam mesej dalam talian adalah tinggi dan memberi kesan kepada aktiviti pemprosesan perkataan seperti pengelasan teks dan perlombongan pendapat.

Penormalan Teks Hingar

Teks hingar di definisi sebagai perkataan yang diringkaskan atau diubah dari perkataan sebenar. Bagi mendapatkan perkataan sebenar, ia perlu melalui proses penormalan teks hingar. Yvon (2010) menganggap penormalan teks hingar sebagai menulis kembali sesuatu mesej dengan menggunakan perkataan yang betul bagi memudahkan manusia atau mesin membaca kandungan mesej tersebut. Selain itu, Vinciarelli (2005) pula merujuk proses penormalan teks hingar kepada aktiviti untuk mengganti teks, menghapus teks atau menambah teks bagi mengubah teks yang sedia ada kepada teks yang sebenarnya bagi membawa maksud tertentu. Dalam konteks kajian ini, pemprosesan teks hingar dilakukan bagi mengurangkan bilangan fitur apabila proses perlombongan

pendapat dilaksanakan. Oleh yang demikian, penormalan teks hingar dalam kajian ini membawa maksud membetulkan ejaan teks hingar kepada ejaan yang betul.

Sistem Imun Buatan

Sistem imun adalah salah satu sistem semula jadi manusia untuk menghalang serangan dari bahan asing yang dikenali sebagai antigen (*ag*) dari merosakkan sistem badan manusia. Bahan penahan yang digunakan oleh sistem imun semula jadi dikenali sebagai antibodi. Berdasarkan metafora ini, **Sistem Imun Buatan (*Artificial Immune System* (AIS)** dibentuk. De Castro dan Timmis (2002) menjelaskan maksud AIS seperti berikut:

> *" Artificial Immune System are adaptive systems, inspired by theoretical immunology and observed immune functions, principles and models, which are applied in problem solving "* (De Castro & Timmis, 2002, hlm 58)

Teori AIS banyak digunakan dalam kajian-kajian lepas bagi menyelesaikan masalah seperti pengelasan data, pengelompokan data, mengenal pasti mail elektronik yang kritikal dan mengenal pasti virus yang memasuki sistem rangkaian komputer.

1.10 ORGANISASI TESIS

Bab II membincangkan kajian-kajian lepas dalam bidang perlombongan pendapat terutamanya kajian yang menggunakan pembelajaran mesin bagi menyelesaikan masalah. Selain itu, kajian yang berkait dengan pemprosesan teks hingar dan pemilihan fitur dalam proses pengelasan teks menggunakan kaedah pembelajaran mesin juga dibincangkan. Di akhir bab ini cadangan kerangka perlombongan pendapat bahasa rojak dicadangkan.

Bab III mengemukakan metodologi kajian yang dilaksanakan. Selain itu, kaedah penyediaan data juga dibincangkan dengan terperinci.

Bab IV membincangkan algoritma penormalan teks hingar (MyTNA) dan kaedah untuk menguji keberkesanan algoritma dalam proses penormalan teks hingar.

Bab V pula membincangkan algoritma untuk memilih fitur berdasarkan sistem imun buatan (FS-INS). Sebelum itu, beberapa aktiviti pra pemprosesan seperti menukar format

huruf kepada huruf kecil, menggabungkan perkataan '*tidak*' dengan perkataan berikutnya dan mengeluarkan perkataan tanpa maksud dijelaskan.

Bab VI membincangkan keputusan atau hasil kajian dari segi keberkesanan pembersihan teks hingar, aktiviti pra pemprosesan dan keberkesanan pemilihan fitur. Beberapa eksperimen yang melibatkan perlombongan pendapat dilaksanakan bagi menguji keberkesanan setiap aktiviti yang diperkenalkan dalam kajian ini.

Bab VII merumus kajian yang dilaksanakan secara menyeluruh dan mengemukakan kelebihan kaedah yang digunakan dalam kajian ini bagi melombong pendapat. Juga dibincangkan kelemahan pendekatan yang diambil dan ruang pengembangan kajian pada masa hadapan.

BAB II

KAJIAN LITERATUR

2.1 PENGENALAN

Bab ini membincangkan kajian-kajian lepas yang berkaitan dengan proses perlombongan pendapat. Ia dimulakan dengan kajian lepas mengenai kerangka perlombongan pendapat yang menggunakan kaedah pembelajaran mesin (Seksyen 2.2), diikuti dengan proses penormalan teks hingar (Seksyen 2.3) dan kaedah pemilihan fitur (Seksyen 2.4). Kajian ini mencadangkan pemilihan fitur berasaskan sistem imun buatan sebagai salah satu kaedah penyelesaian masalah. Justeru itu, bahagian berikutnya mengkaji kaedah pemprosesan teks dengan menggunakan sistem imun buatan (Seksyen 2.5). Bab ini diakhiri dengan perbincangan dan rumusan kajian literatur.

2.2 KERANGKA PERLOMBONGAN PENDAPAT KAEDAH PEMBELAJARAN MESIN

2.2.1 Pengenalan Perlombongan Pendapat

Perlombongan pendapat mula mendapat tempat di awal dekad ke 21 berikutan dari peningkatan teknologi perkakasan dan komunikasi serta peningkatan penggunaan teknologi Internet. Selain itu peningkatan ini dikaitkan dengan peningkatan kajian-kajian pembelajaran mesin khususnya pengelasan teks (Pang & Lee 2008). Tchalakova (2007) mentakrifkan perlombongan pendapat sebagai proses untuk mengekstrak dan mempamerkan penilaian, pendapat atau sentimen terhadap sesuatu perkara yang terkandung dalam sesuatu penulisan. Sentimen ini menyatakan perasaan apabila seseorang menggunakan sesuatu produk atau servis (Boiy & Moens 2009). Jain et al. (2012) pula menganggap perlombongan pendapat sebagai satu proses untuk mempelajari

pendapat atau emosi yang dinyatakan oleh seseorang terhadap sesuatu perkara atau aktiviti.

Di antara isu-isu berkaitan perlombongan pendapat yang diberi perhatian oleh penyelidik adalah :

- mengenal pasti polar sentimen sama ada positif atau negatif (Dey & Haque 2008; Kim & Hovy 2006; Zheng & Ye 2009;);
- pengelasan pendapat kepada kelas tertentu seperti 0/1 sebagai penanda 'sangat tidak suka' hingga 5/10 sebagai penanda 'sangat suka' (Kang et al. 2009; Kim & Hovy 2006) ;
- penyata sentimen dan sentimen yang dinyatakan (Kim & Hovy 2004; Kim & Hovy 2006); dan
- perbandingan satu jenama dengan jenama yang lain (Sun et al. 2009).

Kajian juga dilaksanakan bagi mengenal pasti sentimen di peringkat mesej (Pang et al 2002; Gamon 2004) ayat (Riloff & Wiebe 2003; Wiebe & Mihalcea 2006) atau frasa (Akkaya et al. 2009; Wilson et al. 2005). Kajian ini memberikan tumpuan kepada kaedah untuk mengenal pasti sentimen positif atau sentimen negatif dalam sesuatu mesej. Bagi tujuan tersebut, dua kaedah yang sering digunakan dalam melombong pendapat di kajian yang lepas adalah kaedah pemprosesan bahasa dan kaedah pembelajaran mesin.

Kaedah pemprosesan bahasa menggunakan aktiviti-aktiviti seperti mengenal pasti dan melabel struktur nahu bahasa dan merujuk kepada pangkalan data tertentu untuk mengenal pasti polar sentimen. Kaedah ini menghasilkan ketepatan penilaian yang tinggi. Walau bagaimanapun, kewujudan teks hingar yang tinggi akan mempengaruhi ketepatan perlombongan pendapat (Garmon 2004). Selain dari itu, perlombongan pendapat yang melibatkan bahasa selain dari Bahasa Inggeris sukar dilaksanakan kerana kekurangan bahan rujukan yang boleh mengenal pasti sama ada sesuatu perkataan adalah perkataan sentimen atau tidak seperti Senti WordNet dalam bahasa lain (Boiy & Moens 2009; Tan, S., & Zhang 2008). Kaedah pemprosesan bahasa juga memerlukan masa yang lama untuk memproses sesuatu perkataan dan kurang sesuai digunakan bersama data yang banyak. Jadual 2.1 menyatakan kelebihan dan kekurangan menggunakan kaedah pemprosesan bahasa bagi melombong pendapat. Maklumat lanjut berkaitan perlombongan pendapat menggunakan kaedah pemprosesan bahasa dinyatakan di Lampiran L.

Jadual 2.1 Kelebihan dan kekurangan melombong pendapat menggunakan kaedah pemprosesan bahasa

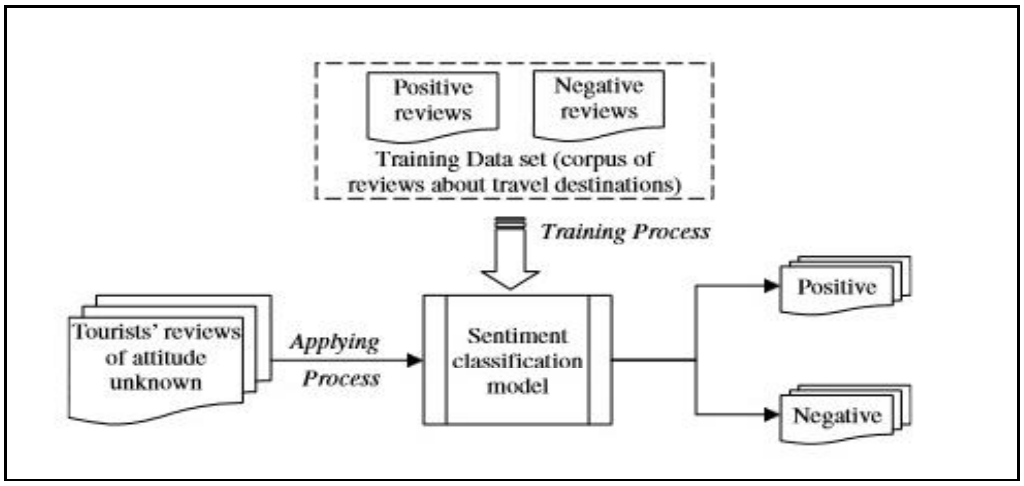
Kelebihan	Kekurangan
1. Ketepatan penilaian adalah tinggi.	1. Format pembentukan ayat perlu mengikut struktur bahasa tertentu. Ketepatan adalah rendah apabila kewujudan teks hingar adalah tinggi. (Garmon, 2004) 2. Perlombongan pendapat bagi bahasa selain dari bahasa Inggeris adalah sukar kerana kurang pangkalan data seperti WordNet dalam bahasa-bahasa lain (Tan, S., & Zhang 2008, Boiy & Moens 2009). 3. Kurang sesuai untuk data yang banyak kerana masa pemprosesan adalah tinggi. (Celikyilmaz, 2010 , Salvetti 2006, Sebastiani 2002,) .

Jadual 2.2 Kelebihan dan kekurangan melombong pendapat menggunakan kaedah pembelajaran mesin

Kelebihan	Kekurangan
1. Sesuai untuk data yang banyak (Salvetti 2006,Sebastiani 2002) 2. Sesuai untuk data yang menggunakan bahasa selain dari bahasa Inggeris (Tan & Zhang 2008, Ye et al. 2009)) 3. Sesuai untuk mesej yang mengandungi tahap teks hingar yang tinggi (Celikyilmaz et al. 2010, Gamon 2004).	1. Keputusan bergantung kepada data yang digunakan sewaktu latihan model pengelasan (Boiy & Moens 2009) . 2. Penggunaan model pengelasan tertentu mempengaruhi keputusan (Boiy 2007). 3. Fitur yang digunakan sewaktu pengelasan juga menentukan masa dan ketepatan pemprosesan (Konig & Brill 2006). 4. Perlu masa dan tenaga untuk melabel data latihan (Konig & Brill 2006)

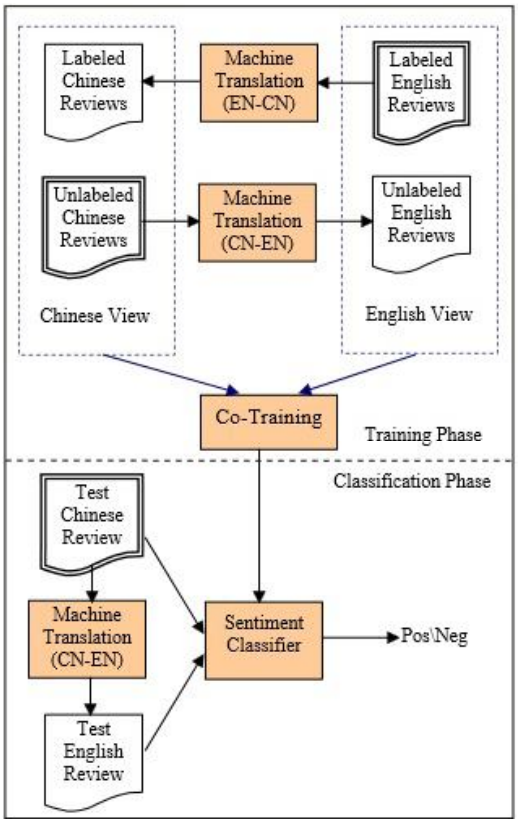
Kaedah perlombongan pendapat menggunakan pembelajaran mesin adalah lebih sesuai untuk melombong mesej dalam talian. Jadual 2.2 menyatakan kelebihan dan kekurangan melombong pendapat menggunakan kaedah pembelajaran mesin. Berdasarkan kepada kelebihan kaedah pembelajaran mesin, tumpuan kajian ini adalah terhadap proses perlombongan pendapat dengan menggunakan kaedah pembelajaran mesin.

Rajah 2.1 menunjukkan kerangka perlombongan pendapat yang digunakan oleh Ye et al. (2008) bagi melombong pendapat di mesej yang mengandungi maklum balas tentang tempat-tempat yang menarik untuk dilawati. Kaedah ini menggunakan pendekatan pembelajaran mesin di mana model pengelasan dibina dengan menggunakan data terdahulu. Model ini kemudiannya digunakan bagi meramal sentimen sesuatu dokumen sama ada positif atau negatif.



Rajah 2.1 Kerangka perlombongan pendapat Ye at el. (2008)

Rajah 2.2 pula menunjukkan kerangka perlombongan pendapat yang digunakan oleh Wan (2009) bagi melombong mesej yang ditulis dalam bahasa China. Kekurangan rujukan bagi mengenal pasti sentimen dalam bahasa China menyebabkan penyelidik menukar semua perkataan bahasa China kepada bahasa Inggeris sebelum proses perlombongan pendapat dilaksanakan.

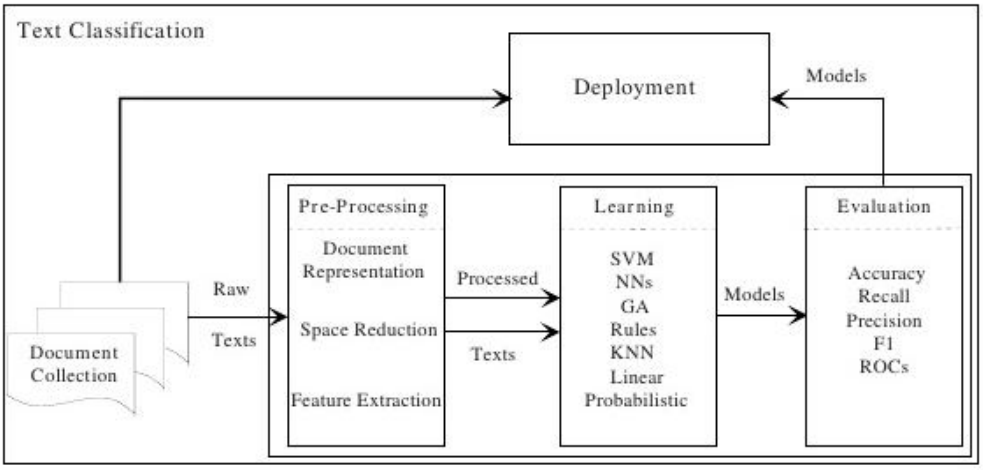


Rajah 2.2 Kerangka Perlombongan Pendapat oleh Wan (2009)

Kajian yang dilakukan oleh Ye at al. (2008) dan Wan (2009) adalah contoh kajian perlombongan pendapat yang menggunakan pendekatan pembelajaran mesin. Pendekatan ini menggunakan aktiviti-aktiviti yang sama seperti perlombongan teks. Hart & Timmis (2008) mendefinisikan kaedah 'pembelajaran' sebagai

> *" process of acquiring knowledge from experience and being able to re-apply that knowledge to previously unseen problem instances."*

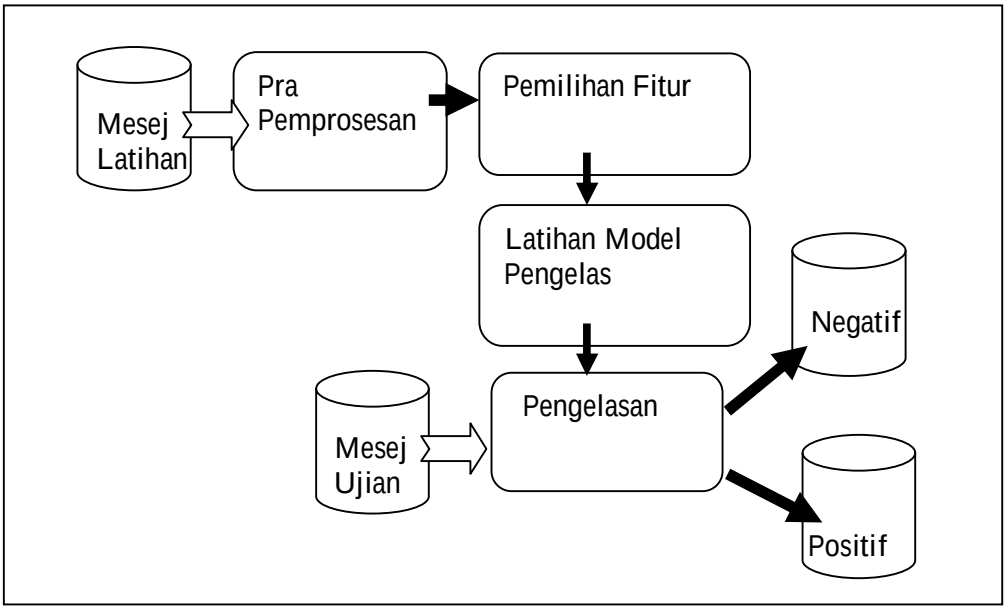
Oleh itu, terdapat aktiviti-aktiviti yang membolehkan sesuatu model pengelas mempelajari sesuatu paten dari data-data yang lepas. Model tersebut digunakan untuk mengelas data yang baru. Rajah 2.3 memaparkan kerangka perlombongan teks oleh Silva (2010) yang menjelaskan secara umum aktiviti-aktiviti yang dilaksanakan bergantung pada keperluan sewaktu menyelesaikan masalah iaitu aktiviti pra pemprosesan, model pembelajaran, penilaian dan pengelasan.



Rajah 2.3 Kerangka perlombongan teks (Silva 2010).

2.2.2 Aktiviti Utama Perlombongan Pendapat Menggunakan Kaedah Pembelajaran Mesin

Secara umumnya terdapat empat aktiviti utama bagi melombong pendapat menggunakan pendekatan ini iaitu Pra Pemprosesan, Pemilihan Fitur, Latihan Model Pengelas dan Pengelasan sebagaimana yang ditunjukkan dalam Rajah 2.4. Fasa-fasa ini dijelaskan secara terperinci dalam sub seksyen berikut.

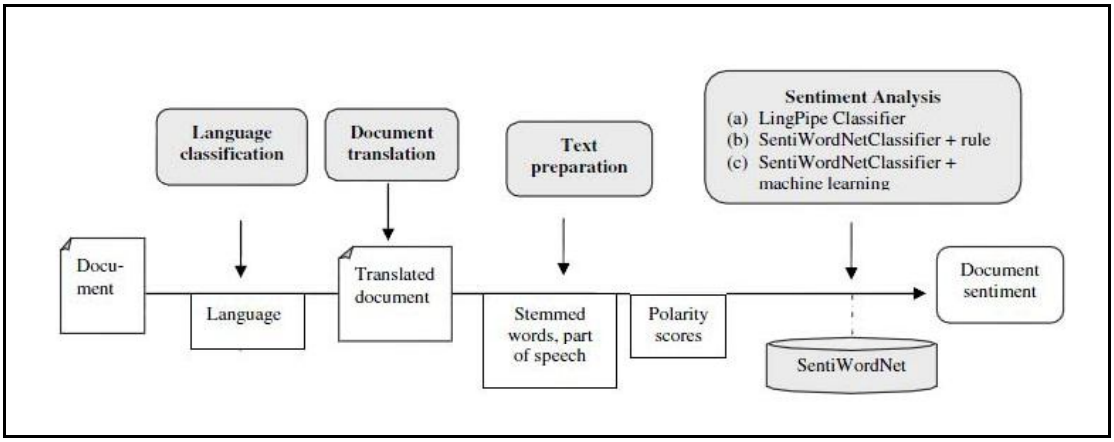


Rajah 2.4 Kerangka Perlombongan Pendapat Menggunakan Kaedah Pembelajaran Mesin

Pra Pemprosesan

Sama seperti perlombongan pendapat menggunakan pemprosesan bahasa, aktiviti pra pemprosesan bertujuan untuk mengurangkan bilangan fitur yang terlibat dalam proses seterusnya. Beberapa aktiviti dilaksanakan dalam fasa ini seperti di senarai berikut.

a) Penggunaan proses pemprosesan bahasa seperti POS, pemangkasan dan perubahan perkataan terbitan kepada perkataan dasar. Garmon (2004) menggunakan perisian NLPWin untuk memilih frasa yang mengandungi struktur *Noun-Verb-Subject* sebagai fitur. Menandakan perkataan berdasarkan nahu ayat juga digunakan oleh penyelidik seperti Boiy & Moens (2007), Boiy & Moens (2009), Ghorpade & Ragha (2012) dan Salvetti et al. (2006)). Walau bagaimanapun teknik ini kurang berkesan apabila digunakan dengan data seperti mesej dalam talian yang mana kandungan teks hingar adalah tinggi di samping pembentukan ayat yang tidak mengikut peraturan nahu sesuatu bahasa. Rajah 2.5 menunjukkan kerangka perlombongan pendapat oleh Deneche (2008) yang menggunakan proses pemprosesan bahasa sebagai proses pra pemprosesan.



Rajah 2.5 Kerangka perlombongan pendapat oleh Deneche (2008)

b) Menggunakan elemen-elemen lain sebagai sebahagian dari kaedah untuk mengenal pasti sentimen. Barbosa & Lee (2005) pula menggunakan beberapa maklumat seperti kewujudan perkataan sentimen, ikon dan penggunaan huruf besar/kecil bagi mengenal pasti sentimen sesuatu mesej Twitter. Kajiannya mendapati 95% mesej positif mengandungi perkataan *"awesome"," rock","love"* dan"*beat"*. Selain itu, 96% mesej Twitter negatif mengandungi perkataan *"hate","suck","wtf","piss","stupid"* dan "*fail"*. Walau bagaimanapun, kajian ini memerlukan tenaga manusia yang tinggi untuk mengenal pasti sama ada sesuatu

mesej mengandungi sentimen atau tidak. Akhir sekali, Konig & Brill (2006) pula melihat kewujudan pola frasa tertentu bagi mengenal pasti sentimen sesuatu mesej. Sekiranya tidak wujud pola ini, mesej akan melalui proses perlombongan pendapat bagi mengenal pasti pola mesej tersebut.

c) Menghapuskan perkataan – perkataan yang tidak berkaitan dengan pengelasan seperti perkataan *'a', 'is', 'are'* dan *'hello'* dalam bahasa Inggeris. Perkataan-perkataan ini juga dikenali sebagai *'stop word'*. Istilah seperti *'saya', 'kamu', 'ini'* dan *'yang'* adalah contoh istilah bahasa Melayu yang tidak berkaitan dengan pengelasan dan boleh dikeluarkan dari mesej.

Pemilihan Fitur

Pemilihan fitur bertujuan untuk mengurangkan bilangan fitur sewaktu proses perlombongan pendapat dilaksanakan.

a) Kaedah statistik seperti kekerapan perkataan, IG dan CHI SQUARE untuk memberi pemberat tertentu kepada setiap fitur. Kemudian fitur-fitur disusun mengikut turutan menaik / menurun bagi memudahkan pengguna memilih fitur yang bersesuaian.
b) Menggunakan kaedah statistik untuk mengenal pasti perkataan subjektif. Pang & Lee (2004) menggunakan kaedah statistik untuk mengenal pasti sama ada sesuatu ayat mengandungi sentimen atau tidak sebelum proses perlombongan pendapat dilaksanakan. Hanya ayat yang menunjukkan sentimen sahaja yang dipilih sebagai fitur. Ayat yang berbentuk fakta dihapuskan dari mesej latihan.

Latihan Model Pengelasan dan Pengelasan Mesej

Pendekatan pembelajaran mesin bermaksud, menggunakan mesej-mesej lalu untuk mempelajari paten atau peraturan yang akan digunakan bagi meramal kelas mesej yang baru. Mesej-mesej terdahulu ini perlu di label dengan kelas yang berkenaan terlebih dahulu. Di fasa ini, model pengelasan yang sesuai dibina dengan menggunakan pengelas seperti NB, kNN dan SMO/ SVM. Pengujian keberkesanan sesuatu model kebiasaannya dilakukan dalam aktiviti Pengelasan Mesej. Dalam fasa ini model yang dihasilkan di fasa Latihan Model Pengelasan digunakan untuk meramal mesej-mesej baru kepada kelas positif atau kelas negatif.

2.2.3 Melombong Pendapat Menggunakan Bahasa Rojak / Bahasa Melayu

Sehingga 2011, tiada aktiviti perlombongan pendapat yang menggunakan mesej dalam talian yang diwujudkan oleh rakyat Malaysia. Beberapa kajian yang melibatkan perlombongan teks dalam bahasa Melayu telah dikenal pasti. Yasukawa et al, (2009) menggunakan perlombongan teks bagi menguji keberkesanan kaedah pemangkasan perkataan. Zamin & Ghani (2010) menggunakan kaedah statistik untuk menghasilkan ringkasan bagi sesuatu dokumen. Noah & Ismail pula menggunakan model pengelas NB untuk mengenal pasti peribahasa dalam himpunan ayat. Kajian perlombongan pendapat yang pertama menggunakan data bahasa rojak dilakukan oleh Norlela et al. (2011). Beliau menjalankan kajian menggunakan 500 maklum balas positif dan 500 maklum balas negatif berkaitan tayangan filem menggunakan kerangka seperti yang dinyatakan dalam rajah 2.5. Keputusan ketepatan perlombongan pendapat yang didapati adalah lebih rendah dari ketepatan perlombongan pendapat dengan menggunakan data yang serupa dalam bahasa Inggeris. Keputusan perlombongan pendapat dengan menggunakan model SVM adalah 62.3 bagi mesej dalam bahasa Melayu berbanding 76.6 bagi mesej dalam bahasa Inggeris. Keputusan perlombongan pendapat dengan menggunakan model pengelas NB juga rendah iaitu 62.8. Pengkaji berpendapat kewujudan teks hingar yang tinggi adalah di antara sebab keputusan ketepatan perlombongan pendapat mesej yang rendah. Penyelidik bersetuju dengan kenyataan Pang et al. (2002) yang menyatakan, perlombongan pendapat menggunakan kaedah pembelajaran mesin memerlukan aktiviti tambahan bagi mendapatkan keputusan ketepatan perlombongan pendapat yang baik. Oleh yang demikian, kaedah penormalan teks hingar dan kaedah pemilihan fitur yang sesuai untuk memilih fitur yang relevan dengan pendapat adalah perlu dalam kerangka perlombongan pendapat dalam bahasa rojak.

2.3 PENORMALAN TEKS HINGAR

Sesuatu perkataan diklasifikasikan sebagai teks hingar apabila ia berlainan dari perkataan yang sebenarnya bagi menyatakan sesuatu maksud (Subramaniam et al. 2009). Peratus kewujudan teks hingar adalah tinggi di mesej dalam talian. Berikut adalah sebab-sebab ia berlaku:

a) Komunikasi yang dilaksanakan secara dalam talian adalah **komunikasi tidak formal** yang melibatkan rakan-rakan dan kaum keluarga terdekat. Oleh yang demikian, mod komunikasi yang digunakan lebih kepada mod perbualan. Ini