PEMILIHAN FITUR MENGGUNAKAN ALGORITMA PENDEBUNGAAN BUNGA BAGI ANALISIS SENTIMEN DATA TWITTER

MUHAMMAD IQBAL BIN ABU LATIFFI

UNIVERSITI KEBANGSAAN MALAYSIA

PEMILIHAN FITUR MENGGUNAKAN ALGORITMA PENDEBUNGAN BUNGA BAGI ANALISIS SENTIMEN DATA TWITTER

MUHAMMAD IQBAL BIN ABU LATIFFI

DISERTASI YANG DIKEMUKAKAN UNTUK MEMENUHI
SEBAHAGIAN DARIPADA SYARAT MEMPEROLEHI
IJAZAH SARJANA SAINS KOMPUTER (KECERDASAN BUATAN)

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2022

PENGAKUAN

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

26 Jan 2022

MUHAMMAD IQBAL BIN ABU LATIFFI
P95662

PENGHARGAAN

Dengan nama Allah Yang Maha Pengasih lagi Maha Penyayang

Pertama sekali, Alhamdulillah, syukur ke hadrat Allah SWT serta selawat dan salam ke atas junjungan besar Nabi Muhammad S.A.W, ahli keluarga, dan sahabat baginda. kerana dengan rahmatnya dapat saya menyiapkan disertasi ini.

Saya ingin mengucapkan jutaan terima kasih yang tidak terhingga kepada satu-satunya penyelia saya iaitu Prof. Madya Dr. Mohd Ridzwan Yaakub di atas bimbingan, tunjuk ajar, teguran, masa, pengalaman dan kata-kata semangat kepada saya sepanjang menjalankan penyelidikan yang mana tidak mampu untuk saya balas. Tidak lupa juga penghargaan diberikan kepada pensyarah-pensyarah, rakan-rakan seperjuangan khususnya dari makmal Analisis Sentimen, dan kakitangan FTSM yang banyak memberikan kerjasama, komen, dan sokongan untuk memantapkan penyelidikan ini.

Tidak dilupakan sama sekali ribuan penghargaan dan terima kasih kepada kedua ibu bapa saya; En. Abu Latiffi Mohd Din dan Pn. Sabariah Abdul Ghani yang menjadi penaja pengajian saya serta menjadi tulang belakang kepada saya dalam menyiapkan penyelidikan ini. Penghargaan juga diucapkan kepada adik-beradik dan ahli keluarga di atas doa, masa dan pengorbanan yang diberikan tanpa henti.

Akhir kata, terima kasih diucapkan sekali lagi kepada semua yang telah membantu saya menyiapkan disertasi ini secara langsung atau tidak langsung. Semoga Allah merahmati anda semua.

ABSTRAK

Pemilihan fitur merupakan topik utama yang seringkali diperincikan oleh penyelidik-penyelidik dalam analisis sentimen. Ini adalah sejajar dengan kemajuan dalam dunia teknologi maklumat dan juga peningkatan penggunaan media sosial yang amat drastik. Masalah utama dalam proses pemilihan fitur bagi analisis sentimen ialah ia mempunyai ruang fitur yang berdimensi besar tambahan pula dengan data yang tidak berstruktur. Hal ini kerana, set data ulasan yang diperoleh daripada media sosial terutamanya di Twitter secara keseluruhannya mempunyai maklumat yang tidak relevan dan hingar untuk kegunaan pengelasan sentimen oleh kerana set data ulasan tersebut di tulis dalam bahasa pasar atau tidak formal seterusnya akan membawa kepada hasil prestasi pengelasan yang tidak memuaskan. Jadi keadaan ini telah memberi ruang kepada kajian untuk meningkatkan hasil prestasi pengelasan sentimen. Penemuan mendapati bahawa perlunya set data ulasan melalui proses pembersihan awal menggunakan kaedah prapemprosesan teks yang menggabungkan teknik pemprosesan bahasa tabii dengan teknik pemprosesan linguistik yang mana menghasilkan prestasi pengelasan yang tinggi. Selain itu, set data ulasan akan melalui proses pemilihan fitur yang berperanan penting dalam membantu mengurangkan saiz dimensi fitur dan juga berupaya memilih fitur-fitur yang bermakna sekaligus memberi hasil pengelasan yang tinggi. Antara teknik pemilihan fitur termasuklah teknik pembalut, teknik penapis dan juga teknik metaheuristik. Kelebihan pendekatan metaheuristik adalah berlakunya proses pencarian fitur secara menyeluruh dalam ruang untuk menghasilkan penyelesaian yang lebih baik dengan mengaplikasikan pengetahuan yang diperoleh daripada penyelesaian semula jadi terutamanya bagi data Twitter yang secara umumnya terkandung data teks yang tidak formal dan hingar. Oleh itu, kajian mencadangkan penggunaan Algoritma Pendebungaan Bunga (APB) sebagai algoritma pemilihan fitur yang berupaya memperoleh subset fitur optimum dan berkualiti. Keberkesanan kaedah yang telah dicadangkan diuji dengan menggunakan set data ulasan daripada Twitter dan bagi pengelasan sentiment pula menggunakan empat pengelas pembelajaran mesin iaitu Naïve Bayes, Mesin Vektor Sokongan, Pepohon Keputusan dan k-Jiran Terdekat. Hasil daripada eksperimen ini mendapati bahawa penggunaan teknik prapemprosesan teks serta teknik pemilihan fitur APB telah berjaya menghasilkan ketepatan prestasi pengelasan yang tertinggi iaitu 98.99% bagi pengelas Mesin Vektor Sokongan di mana terdapat kenaikan sebanyak 2.68% berbanding dengan teknik penanda aras iaitu algoritma Cuckoo Search. Kajian ini membuktikan penggabungan pada teknik prapemprosesan teks dan APB ini adalah lebih baik dan bermakna dalam penghasilan subset fitur yang optimum serta berkualiti bagi tujuan pengelasan sentimen.

FEATURE SELECTION USING FLOWER POLLINATION ALGORITHM (FPA) IN SENTIMENT ANALYSIS FOR TWITTER DATA

ABSTRACT

Feature selection is a major topic that is often detailed by researchers for sentiment analysis. This is in line with the advances in the world of information technology as well as the drastic increase in the use of social media. The main problem in the feature selection process for sentiment analysis is having a large dimensional feature space especially using the unstructured data. This is because, the reviews data set obtained from social media as a whole has noisy and irrelevant information for the use of sentiment classification. This problem will lead to unsatisfactory classification performance results. So, this situation has given space to studies to improve the results of sentiment classification performance. The findings found that the need for review data set through the initial cleaning process using text processing methods that combine natural language processing techniques with linguistic processing techniques which results in high classification performance. In addition, the review data set will go through the feature selection process which plays an important role in helping to reduce the size of the feature dimensions and also try to select meaningful features while providing high classification results. Among the feature selection techniques include wrapping technique, filtering technique as well as metaheuristic techniques. The advantage of the metaheuristic technique is that the overall feature search process in the solution space can solve this problem. Therefore, the study suggests that a combination of metaheuristic techniques namely Flower Pollination Algorithm (FPA) serves to evaluate and generate an optimal features subset. This combination is able to obtain an optimal feature subset and quality features. The effectiveness of the proposed method was tested using a review data set from Twitter and using four machine learning classifiers namely Naïve Bayes, Support Vector Machine, Decision Tree and k-Neighbors. The results of this experiment found that the use of text pre-processing techniques as well as a combination of FPA feature selection techniques have succeeded in producing the highest accuracy for sentiment classification performance which is 98.99% for Support Vector Machine where there a 2.68% increment compared to the baseline technique Cuckoo search technique. This study proves that the combination of text processing pre-processing techniques and FPA approach is better and more significance in producing a subset of optimal and quality features for the purpose of classifying sentiments.

KANDUNGAN

Halaman

SENARAI JADUAL

SENARAI RAJAH

SENARAI SINGKATAN

APB	Algortima Pendebungaan Bunga
BOW	*Bag of Words*
DT	*Decision Tree*
FN	*False Negative*
FP	*False Positive*
FPA	*Flower Pollination Algorithm*
GNU	*General Public License*
kNN	*k-Nearest Neighbor*
NB	Naïve Bayes
SVM	*Support Vector Machine*
TF-IDF	*Term Frequency-Inverse Document Frequency*
TN	*True Negative*
TP	*True Positive*
WEKA	*Waikato Environment for Knowledge Analysis*

BAB I

PENGENALAN

1.1 PENDAHULUAN

Pada bab ini akan menjelaskan secara umum keseluruhan kajian ini di mana ia merangkumi latar belakang kajian, pernyataan masalah, objektif kajian, skop kajian, metodologi kajian, ringkasan sumbangan kajian dan diakhiri dengan struktur kandungan tesis ini.

1.2 LATAR BELAKANG KAJIAN

Revolusi Perindustrian Keempat telah mengubah cara hidup, cara manusia bekerja dan cara manusia berkomunikasi. Teknologi maklumat kini telah menjadi wadah baharu dalam bidang perniagaan dan keusahawanan untuk terus berkembang dan meningkatkan mutu perkhidmatan mereka (Mohamad 2018). Kemajuan teknologi maklumat terutamanya dalam kemajuan capaian internet telah membuatkan pengguna mengambil keputusan untuk berurus niaga secara dalam talian dalam semua urusan harian mereka. Ini termasuklah urusan perbankan, urusan jual beli, urusan menempah perkhidmatan dan juga urusan hal ehwal akademik. Hal ini kerana, pengguna terutamanya boleh menggunakan capaian internet ini pada bila-bila masa dan di mana sahaja sekaligus memberi kesan yang besar kepada pengusaha yang datang dari pelbagai latar belakang dan industri.

Oleh yang demikian, kemunculan generasi kedua jaringan sejagat telah memacu kepada perkembangan laman sesawang yang lebih interaktif di mana pengguna-pengguna akan menjadi peranan penting apabila mereka akan berkongsi pendapat mahupun cadangan serta memberi ulasan menerusi perkhidmatan laman

sesawang tersebut. Interaksi yang berskala besar yang tidak terhad hanya kepada mengulas sesuatu berita, perkhidmatan dan juga barangan atau produk, tetapi juga boleh dijadikan medium bagi menyebar luaskan komunikasi di antara pengguna laman sesawang dalam bentuk rangkaian sosial. Hal ini kerana melalui rangkaian sosial, ia membenarkan pengguna menghubungi sesama sendiri, berinteraksi dan sehingga boleh berkolaberasi untuk menjadi pencipta kandungan dalam satu platform yang dikenali sebagai media sosial. Melalui media sosial mereka boleh memberi pandangan, ulasan serta pengalaman terhadap sesuatu isu (Dritsas et al. 2019; Priya & Devapriya 2019; Zhou et al. 2019).

Oleh kerana terdapat sejumlah bilangan besar pengguna dan juga berlaku pertukaran maklumat juga dalam jumlah yang besar dalam bentuk ulasan membuatkan proses menganalisis ulasan-ulasan tersebut secara manual adalah sukar dan akan mengambil masa yang lama bahkan ia adalah mustahil bagi sesetengah domain. Bagi menangani masalah ini, analisis sentimen telah diperkenalkan sebagai satu kaedah yang berkesan dalam mengenalpasti pengetahuan atau maklumat yang betul dan sahih (B. Liu 2012) daripada ulasan-ulasan pengguna kerana ia dilihat sebagai elemen penting kepada pengguna lain dan juga kepada pengusaha dan pengeluar produk atau perkhidmatan (Khan et al. 2016; Yaakub et al. 2013).

Analisis sentimen merupakan sejenis pengelasan teks yang melibatkan penyataan yang subjektif. Nama lain bagi analisis sentimen adalah perlombongan pendapat di mana suatu pendapat ataupun ulasan diproses untuk mengetahui persepsi pengguna terhadap sesuatu perkara ataupun isu. Analisis sentimen merupakan salah satu bidang perlombongan data dan perlombongan teks yang berada di bawah kategori teknik perlombongan kandungan web (Khan et al. 2016; Sohrabi & Hemmatian 2019). Fungsi utamanya adalah untuk mengelaskan maklumat yang terdapat di dalam dokumen teks seperti ulasan pengguna terhadap produk, perkhidmatan mahupun terhadap sesuatu isu kepada kumpulan sentimen positif atau sentimen negatif (Pang & Lee 2008). Maka, terdapat empat fasa penting dalam analisis sentimen iaitu bermula dengan pemprosesan teks, pemilihan fitur, pengekstrakan fitur dan pengelasan sentimen. Setiap fasa yang dinyatakan memainkan peranan penting dalam meningkatkan prestasi pengelasan sentimen.

Analisis sentimen mengandungi tiga peringkat iaitu peringkat dokumen, peringkat ayat dan peringkat aspek atau fitur. Peringkat dokumen dalam analisis sentimen lebih menumpukan kepada mengklasifikasikan dokumen ulasan sama ada keseluruhan dokumen tersebut adalah positif atau negatif sentimen (B. Liu 2012; Medhat et al. 2014). Bagi peringkat ayat menurut (Medhat et al. 2014) pula ia boleh takrifkan untuk mengklasifikasikan sentimen yang dinyatakan dalam setiap ayat. Ini dengan mengenal pasti sama ada ayat itu subjektif atau objektif. Sekiranya ayat itu bersifat subjektif, analisis sentimen peringkat ayat akan menentukan sama ada ayat itu menyatakan pendapat positif atau negatif. Manakala bagi peringkat aspek pula ia bertujuan untuk mengklasifikasikan sentimen berkenaan dengan aspek entiti tertentu. Tujuan utama bagi peringkat ini adalah untuk mengenalpasti entiti dan aspek mereka. Para penulis sentimen boleh memberikan pandangan yang berbeza untuk aspek yang berbeza dari entiti yang sama (Ahmad et al. 2016; Medhat et al. 2014). Hasil daripada analisis ini sentimen pada peringkat ini akan menghasilkan keputusan yang lebih terperinci berbanding dengan peringkat dokumen dan ayat (B. Liu 2012).

Proses pengelasan sentimen merupakan masalah utama dalam teknologi analisis sentimen kerana penglibatan data yang berbentuk teks (Ahmad et al. 2019a). Tambahan pula, pengelasan teks biasa dan pengelasan sentimen dalam perlombongan teks adalah dua perkara yang berbeza. Pengelasan teks dalam perlombongan teks ialah mengenal pasti topik yang terdapat dalam set data. Manakala bagi perlombongan ulasan pengguna, pengelasan dilakukan berdasarkan pengelasan iaitu jenis sentimen dalam ulasan pengguna (Seerat & Azam 2012). Prestasi pengelasan sentimen dapat ditambahbaik apabila terdapat kaedah untuk mengurangkan dimensi di dalam set data bersaiz besar (Pervaiz et al. 2018). Kaedah ini secara umumnya berfungsi untuk mengesan dan menghapuskan data yang tidak relevan dan bertindan supaya hasil pengelasan sentimen adalah bermakna. Jadi, pemilihan fitur telah menjadi penyelidikan yang aktif dalam bidang pembelajaran mesin, pengecaman corak dan pengesanan komuniti. Idea utama dalam pemilihan fitur ini adalah untuk memilih subset bagi set data asal dengan membuang data yang bertindan dan data yang mempunyai sedikit ataupun tidak langsung maklumat yang berguna. Tunjang utama teknik pemilihan fitur ini adalah data yang mengandungi fitur-fitur sama ada ia bertindan atau tidak relevan yang boleh dikeluarkan mahupun dibuang tanpa

menghilangkan maklumat penting di mana mengutamakan fitur-fitur yang bermakna seterusnya meningkatkan prestasi pengelasan (Deng et al. 2019). Secara umumnya terdapat tiga jenis teknik pemilihan fitur yang telah diperkenalkan iaitu teknik penapis, teknik pembalut dan teknik hibrid (Deng et al. 2019; P. Kumar et al. 2018). Selain daripada teknik-teknik ini, teknik metaheuristik seperti Algoritma Genetik (GA), Pengoptimum Koloni Semut (ACO), *Particle Swarm Optimisation* (PSO) dan Algoritma Kelawar (BA) telah menarik perhatian penyelidik-penyelidik (Ahmad et al. 2019b; Akshi & Renu 2017; Alyasseri et al. 2018; W. Liu & Wang 2019; Tubishat et al. 2019). Teknik-teknik ini telah cuba untuk mendapatkan penyelesaian yang lebih baik daripada teknik-teknik yang telah dinyatakan sebelumnya. Teknik ini dilihat sesuai bagi data yang mempunyai penyelesaian yang pelbagai kemungkinan.

1.3 PERMASALAHAN KAJIAN

Berdasarkan latar belakang kajian yang telah dikemukan, terdapat beberapa masalah yang telah dikenalpasti dalam proses analisis sentimen terutamanya dalam pengelasan sentimen. Menurut (Seerat & Azam 2012) masalah pengelasan teks merupakan cabaran utama dalam analisis sentimen kerana terdapatnya penggunaan data yang berbentuk teks. Pengelasan teks biasa dan perlombongan ulasan merupakan dua perkara yang berbeza. Pengelasan teks dalam perlombongan data bertujuan untuk mengenalpasti topik yang terdapat dalam set data. Tetapi bagi perlombongan ulasan pula, pengelasan berlaku berdasarkan jenis sentimen yang terdapat dalam ulasan. Dalam perlombongan data, pengelasan berperanan untuk mengelaskan fitur berdasarkan set data latihan yang telah diklasifikasikan dan ia dianggap penting (Seerat & Azam 2012).

Prestasi pengelasan sentimen dapat ditingkatkan dengan mengurangkan dimensi yang digunakan bagi mengurangkan saiz set data latihan yang bersaiz besar sekali gus dengan ini ia berfungsi untuk mengesan dan menghapuskan fitur yang tidak berkaitan dan bertindan (Kim & Ganesan 2011). Terdapat penyelidikan yang mengambil mudah akan proses prapemprosesan teks (Ganesan 2019). Didapati bahawa terdapat keputusan yang tidak konsisten daripada hasil keputusan pengelasan teks berkemungkinan yang berlaku adalah proses prapemprosesan teks ini tidak

berlaku dengan efektif ataupun menggunakan jenis prapemprosesan teks yang salah. Kebanyakan penyelidik hanya menggunakan sebahagian daripada kesuluruhan teknik-teknik prapemprosesan teks (Ahmad et al. 2019b; Akshi & Renu 2017; Tubishat et al. 2019). Di samping penambahbaikan pada fasa pemilihan fitur, fasa prapemprosesan teks juga memainkan peranan dalam meningkatkan prestasi pengelasan teks atau sentimen apabila pemilihan fitur-fitur penting dapat ditingkatkan (Krouska et al. 2016; Priya & Devapriya 2019). Menurut (Khader et al. 2019), kebiasaannya akan ada peningkatan pada hasil ketepatan pengelasan teks.

Isu seterusnya dalam analisis sentimen adalah saiz dimensi fitur yang besar digunakan bagi menghuraikan teks. Hal ini kerana konsep *Bag-of-Words* yang sering digunakan sebagai perwakilan bagi teks dokumen dalam pengelasan sentimen yang menggunakan kaedah pembelajaran mesin (B. Agarwal & Mittal 2013; Ahmad et al. 2015). Perkataan-perkataan yang terdapat dalam dokumen teks menghasilkan vektor fitur yang mempunyai dimensi bersaiz besar. Justeru, proses pemilihan fitur memfokuskan untuk memilih subset fitur yang optimum dari saiz fitur yang berdimensi besar. Ini dilakukan dengan menyingkirkan fitur yang bingar dan tidak relevan tanpa mengubah data asal (Ahmad et al. 2019a; P. Kumar et al. 2018; Trivedi & Dey 2018). Di dalam kajian tersebut kaedah pembelajaran mesin di mana ia digunakan untuk mengenal pasti perkataan fitur yang berkualiti dengan jumlah kuantiti fitur yang sesuai merupakan isu yang sering dibincangkan. Hal ini kerana, teknik pemilihan fitur penting dalam menentukan prestasi ketepatan pengelasan sentimen.

Oleh sebab itu beberapa kajian tentang teknik pemilihan fitur telah diutarakan pada bab kajian kesusasteraan di mana pemilihan subset fitur merupakan masalah polinomial yang tidak berketentuan dan memerlukan algoritma yang efisien seperti algoritma metaheuristik bagi menyelesaikan masalah pemilihan fitur. Kajian tentang pemilihan fitur mahupun pengelasan sentimen masih relevan untuk dilakukan untuk tujuan penyelidikan (Alsaffar & Omar 2015; Apte & Khetwat 2019; Chang et al. 2020; Meera & Sundar 2020).

Antara kaedah pemilihan fitur yang sering digunakan antaranya adalah kaedah pemeringkatan penapis di mana kaedah ini mengambil kira hubungan di antara setiap fitur dengan kelas yang sesuai. Kaedah ini hanya menilai setiap fitur yang ada secara berasingan berdasarkan kriteria tertentu dan seterusnya menapis keluar fitur-fitur yang berada pada kedudukan terendah. Tambahan lagi, kaedah ini tidak melibatkan kebergantungan interaksi antara fitur (Ansari et al. 2019; Bommert et al. 2020; Labani et al. 2018). Walaubagaimanapun, prestasi pengelasan sentimen dapat ditingkatkan apabila fitur-fitur yang berada pada kedudukan rendah atau dianggap lemah ini relevan terhadap kelas sasaran dengan sendirinya (Labani et al. 2018). Akan tetapi, fitur-fitur yang relevan secara individu ini boleh menjadi bertindih apabila digunakan bersama fitur-fitur yang lain. Jadi, set fitur yang optimum boleh hilang apabila berlakunya proses penghapusan atau pemilihan fitur.

Justeru itu, kajian ini akan menggunakan teknik-teknik tambahan dalam prapemprosesan teks berbanding pada model asas dan akan menggunakan empat (4) pengelas pembelajaran mesin yang lain daripada model asas bagi tujuan pengujian, penilaian dan analisis.

1.4 PERSOALAN KAJIAN

1. Apakah kesan pemilihan kaedah prapemprosesan teks daripada model pemprosesan bahasa tabii terhadap prestasi pengelasan sentimen?

2. Apakah teknik yang terbaik yang dapat melaksanakan proses pemilihan fitur data Twitter bagi meningkatkan prestasi pengelasan sentimen?

1.5 OBJEKTIF KAJIAN

Matlamat kajian ini dijalankan untuk menganalisis maklumat yang bernilai dalam set data Twitter bagi menghasilkan pengelasan sentimen yang terbaik. Bagi merealisasikan matlamat tersebut, dua objektif telah dikemukakan:

1. Untuk mengenalpasti kesan pemilihan kaedah prapemprosesan teks daripada model pemprosesan bahasa tabii terhadap prestasi pengelasan sentimen.

2. Untuk mencadangkan algoritma APB bagi melaksanakan proses pemilihan fitur data Twitter bagi meningkatkan prestasi pengelasan sentimen.

1.6 SKOP KAJIAN

Kajian ini memfokuskan kepada pembangunan algoritma APB bagi melaksanakan proses pemilihan fitur. Tujuan utama pada proses ini adalah bertujuan untuk meningkatkan prestasi pengelasan sentimen di mana apabila subset fitur yang optimum dihasilkan untuk mewakili data daripada set data yang sebenar. Pemilihan algoritma ini adalah sebagai teknik pemilihan fitur berdasarkan implementasi algoritma ini pada pengelasan teks biasa. Tetapi terdapat perbezaannya kerana pada kajian ini lebih memfokuskan kepada pengelasan sentimen di mana ia akan mengelaskan fitur-fitur berdasarkan kategori sentimen iaitu positif dan negatif.

Kajian ini juga akan mengenalpasti kaedah pemprosesan teks daripada model pemprosesan bahasa tabii yang akan digabungkan dengan model pemprosesan linguistik untuk membersihkan dan menganalisis ulasan dari hingar, data yang tidak konsisten dan data yang tidak relevan. Empat jenis model gabungan kaedah pemprosesan teks akan diuji cuba dan perbandingan kadar ketepatan analisis sentimen akan dilaksanakan bagi memastikan kaedah pemprosesan yang terbaik untuk digunakan dalam kajian ini.

Pada fasa pengujian, teknik pemilihan fitur yang dicadangkan ini diuji dengan set data ulasan Twitter yang piawai di mana ia telah digunakan oleh penyelidik terdahulu (A. Kumar et al. 2018) yang boleh diperoleh daripada repositori *Kaggle*. Set data ini terdiri daripada 7086 ulasan Twitter yang berlabel positif dan negatif serta merupakan ulasan yang ditulis dalam Bahasa Inggeris. Pembangunan teknik pemilihan fitur diuji dengan algoritma asas perbandingan iaitu teknik TF-IDF dan *Binary Cuckoo Search*. Kajian ini tidak akan merangkumi jumlah masa yang diperlukan bagi fasa prapemprosesan teks, pemilihan fitur dan juga pengelasan. Maka, pengukuran terhadap kadar masa tidak akan dilakukan.

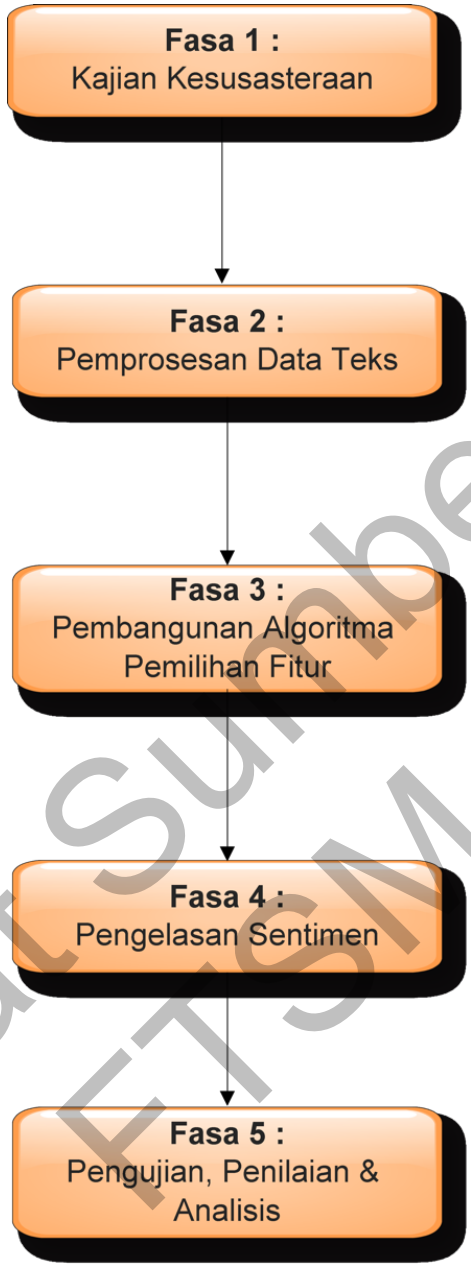
Selain daripada itu, bagi pengujian pengelasan sentimen menggunakan beberapa algoritma pengelas pembelajaran mesin seperti Mesin Vektor Sokongan

(SVM), Keputusan Pokok (DT), Naive Bayes (NB) dan k-Jiran Terdekat (kNN) di mana penilaian yang digunakan adalah berdasarkan tiga kriteria termasuklah ketepatan (A), kejituan (P) dan dapatan semula (R).

Bagi eksperimen prapemprosesan teks dan pemilihan fitur, ia dibangunkan dengan menggunakan pengaturcaraan Python yang melibatkan penggunaan perpustakaan *Natural Language Toolkit* (NLTK) dan mealpy masing-masing. Seterusnya, eksperimen pengelasan sentimen dilakukan dengan menggunakan perisian WEKA (Waikato Environment for Knowledge Analysis) versi 3.8.

1.7 METODOLOGI KAJIAN

Metodologi bagi kajian direka bentuk sebagai arah dan panduan bagi melaksanakan kajian ini seterusnya objektif-objektif kajian dapat dicapai. Selain daripada itu, kajian ini berdasarkan kepada penyelidikan eksperimental untuk kes perlombongan data berbentuk teks. Justeru itu, kajian ini mengandungi lima (5) fasa atau peringkat bermula dengan kajian kesusasteraan, pemprosesan teks, pembangunan algoritma pemilihan fitur, pengelasan sentimen dan pengujian, penilaian serta analisis. Ringkasan bagi fasa-fasa ini boleh dilihat pada Rajah 1.1. Setiap fasa yang terlibat ini akan diterangkan secara lebih mendalam pada Bab III.



Rajah 1.1 Metodologi Kajian Umum

1.8 RINGKASAN SUMBANGAN KAJIAN

Kajian ini dapat memberi sumbangan dari segi aspek kajian ilmiah dalam bidang analisis sentimen terutamanya data-data ulasan daripada blog mikro seperti Twitter khususnya dalam kajian permodelan analisis sentimen.

1. Cadangan kaedah prapemprosesan teks dari model pemprosesan bahasa tabii yang boleh meningkatkan prestasi ketepatan pengelasan sentimen.

2. Cadangan penggunaan algoritma teknik pembalut dan algoritma metaheuristik sebagai teknik pemilihan fitur dalam pengelasan sentimen. Cadangan ini telah menghasilkan APB yang dapat menghasilkan subset fitur yang optimum dan penting. Penggabungan algoritma-algoritma ini adalah bersesuaian dan boleh diaplikasikan kepada masalah pemilihan fitur yang lain.

1.9 STRUKTUR KANDUNGAN TESIS

Terdapat lima (5) bab secara keseluruhan di dalam tesis ini yang merangkumi;

Bab I membincangkan secara keseluruhan penyelidikan yang dilaksanakan termasuklah latar belakang kajian, permasalahan kajian, persoalan kajian, objektif kajian, skop kajian, metodologi kajian dan ringkasan sumbangan kajian.

Pada Bab II pula menerangkan tentang konsep asas analisis sentimen secara lebih mendalam seperti takrifan istilah, domain, kepentingan dan faktor keberkesanan analisis sentimen. Selain daripada itu, bab ini membincangkan secara lebih lanjut kajian-kajian lepas tentang teknik-teknik pemilihan fitur yang digunakan serta mengenalpasti kekurangan yang terdapat pada teknik-teknik sedia ada. Seterusnya, teknik-teknik pengelasan sentimen berdasarkan kajian-kajian lepas juga dibincangkan di dalam bab ini bagi mengenalpasti algoritma pengelas yang sering digunakan untuk pengelasan sentimen.

Bab III pula menerangkan metodologi kajian yang digunakan. Bab ini bermula dengan penyediaan dan pemprosesan data, implementasi, pembangunan algoritma bagi teknik pemilihan fitur, pengelasan sentimen dan diakhiri dengan kaedah pengujian, analisis serta penilaian.

Bab IV menerangkan dan merumuskan dapatan kajian yang diperolehi daripada metodologi kajian yang telah dilaksanakan. Ini merangkumi pembangunan dan implementasi algoritma-algoritma bagi menjalankan proses pemilihan fitur dalam analisis sentimen. Oleh itu, algoritma perbandingan asas daripada kajian terdahulu termasuklah algoritma APB digunakan bagi menguji kebolehan algoritma yang telah dicadangkan. Selain daripada itu, bagi mendapatkan hasil prestasi pengelasan

sentimen terdapat empat (4) algoritma pengelas pembelajaran mesin digunakan. Seterusnya ia akan membincangkan mengenai prestasi dan faktor-faktor yang menyumbang kepada keputusan yang diperoleh.

Bab V pula merumuskan secara keseluruhan tentang kajian yang telah dijalankan kemudian sumbangan kajian juga turut dibincangkan. Di samping itu, cadangan untuk perluasan kerja yang akan datang juga dikemukan.

BAB II

KAJIAN KESUSASTERAAN

2.1 PENGENALAN

Pada bab ini akan diterangkan hasil kajian kesusateraan yang merangkumi kajian-kajian lepas. Perkara yang berkaitan analisis sentimen termasuklah definisi istilah, domain, tujuan dan kepentingan analisis sentimen. Selain itu, proses-proses yang penting dan memainkan peranan besar dalam analisis sentimen iaitu pemilihan fitur, pengelasan sentimen dan penilaian juga akan diterangkan dalam bab ini. Bagi setiap proses ini, perbincangan secara mendalam akan dilakukan dengan mengkaji serta membandingkan kajian-kajian lepas secara komprehensif. Dengan cara ini, segala kelebihan dan kekurangan pendekatan serta teknik yang telah digunakan pada setiap kajian-kajian lepas yang dikaji ini dapat dikenalpasti.

2.2 DEFINISI ISTILAH

Di dalam penyelidikan ini, terdapat beberapa istilah berkaitan dengan analisis sentimen yang diguna pakai. Oleh hal yang demikian, istilah-istilah ini perlu difahami dengan terperinci sebelum penyelidikan secara mendalam dilakukan.

2.2.1 Analisis Sentimen

Menurut (B. Liu 2012), menyatakan bahawa sentimen analsis yang juga boleh dipanggil sebagai perlombongan pendapat merupakan kajian yang menganalsis pendapat, persespsi, penilaian, sikap dan emosi pengguna terhadap sesuatu entiti antaranya ialah produk, organisasi, perkhidmatan, peristiwa atau sifat-sifat entiti itu sendiri. Namun begitu, (Medhat et al. 2014) telah membezakan sentimen analisis dan perlombongan pendapat di mana sentimen analisis adalah mencari pendapat,

mengenal pasti sentimen yang diutarakan kemudian mengelaskan sentimen tersebut manakala bagi perlombongan pendapat merujuk kepada mengekstrak dan menganalisis pendapat pengguna terhadap sesuatu entiti. Pada pandangan (Zvarevashe & Olugbara 2018) analisis sentimen melibatkan penggunaan pemprosesan bahasa tabii dan linguistik komputeran untuk melakukan pengelasan sentimen secara automatik daripada ulasan. Justeru itu, tumpuan utama kajian tentang analisis sentimen ini adalah dengan mencari data iaitu pendapat atau ulasan pengguna, mengenalpasti sentimen yang terdapat di dalam pendapat atau ulasan pengguna tersebut dan akhir sekali pengelasan sentimen berlaku di mana akan terbahagi kepada positif, negatif atau neutral.

2.2.2 Model Analisis Sentimen

Di dalam model analisis sentimen, terdapat beberapa entiti yang sering digunakan bagi merujuk kepada sesuatu perkara antaranya termasuklah objek, sentimen terhadap fitur, penulis sentimen dan lain-lain. Oleh itu, perkara tersebut akan dibicangkan pada subseksyen di bawah.

a. Definisi Objek

(B. Liu 2012; Seerat & Azam 2012) menjelaskan bahawa objek A adalah entiti kepada e di mana e merupakan produk, perkhidmatan, organisasi mahupun orang. Ia boleh digambarkan seperti A: (K, S), di mana K merujuk kepada komponen dan sub komponen kepada A manakala S adalah atribut kepada A yang juga dikenali sebagai set fitur. Sebagai tambahan, setiap komponen mempunyai sub komponen dan atribut atau fitur sendiri. Di sini, kita boleh gambarkan dokumen t di mana mengandungi pendapat bagi objek A dan secara umumnya t mempunyai ayat-ayat seperti t = (p1, p2, p3... pn).

b. Definisi Sentimen terhadap Fitur

Sentimen atau persepsi terhadap fitur f bagi objek A yang diekstrak daripada dokumen t, yang juga mengandungi ayat-ayat daripada t yang terkandung beberapa pendapat

atau persepsi f. Dalam kata lain, satu-satu ayat boleh mengungkapkan pendapat atau persepsi untuk beberapa fitur produk (Seerat & Azam 2012).

c. Definisi Pengulas Sentimen

Individu yang mengeluarkan pendapat ataupun menulis ulasan terhadap sesuatu yang mengandungi unsur sentiment positif, negative mahupun neutral bagi sesuatu objek (B. Liu 2012)

d. Definisi Pengelasan Sentimen

Definisi bagi pengelasan sentimen menurut (Seerat & Azam 2012) ialah sekiranya satu set dokumen penilaian T diberikan, ia menentukan sama ada setiap dokumen t ϵ T menyatakan pendapat positif atau negatif pada sesuatu objek. Pengelasan sentimen pada dasarnya menentukan orientasi semantik pendapat yang dinyatakan pada objek A dalam setiap dokumen penilaian.

2.2.3 Definisi Fitur

Setiap domain dalam analisis sentimen mempunyai fitur atau ciri-ciri yang berbeza. Sebagai contoh dalam teks ulasan tentang pelancongan. “hotel” merupakan topik berpotensi yang mempunyai pelbagai ciri seperti harga, kebersihan, perkhidmatan pelanggan dan lokasi (Alaei et al. 2017).

2.2.4 Definisi Perkataan Sentimen

Bagi perkataan sentimen, ia merujuk kepada perkataan yang mengandungi sifat-sifat sentimen sama ada positif, negatif atau neutral mengenai fitur sesuatu objek. Contoh perkataan sentimen positif yang biasa ditulis oleh penulis dalam ulasan adalah *awesome, good* dan *excellent* manakala bagi perkataan sentimen negatif adalah *poor, terrible* dan *bad*.

2.2.5 Definisi Pemilihan Fitur

Terdapat pelbagai definisi pemilihan fitur yang ditafsirkan oleh para penyelidik daripada kajian terdahulu berdasarkan perspektif yang berbeza. Masalah pemilihan fitur sering dibincangkan dalam fungsi yang melibatkan pembelajaran mesin terselia dan tak terselia seperti pengelasan, pengelompokan regresi dan ramalan siri masa. Cadangan kajian ini, ialah proses pemilihan fitur produk bagi menyelesaikan masalah pemilihan fitur dalam pengelasan sentimen. Penekanan dalam pemilihan fitur iaitu untuk memperbaiki ketepatan ramalan atau mengurangkan saiz struktur dengan memilih satu subset fitur tanpa secara signifikan mengurangkan ketepatan ramalan pengelas yang dibina menggunakan fitur-fitur yang terpilih (P. Kumar et al. 2018). Walaubagaimanapun, takrifan (Basu & Murthy 2012) pemilihan fitur adalah satu set senarai fitur yang dipilih oleh sistem pengelasan bagi menghasilkan satu subset fitur yang terbaik. Bagi (Achhab & Lazaar 2018) masalah pemilihan fitur ialah memilih satu subset fitur yang bersaiz m daripada satu set fitur iaitu d yang menghasilkan ralat pengelasan yang kecil. Kebanyakan pendekatan dalam masalah ini memerlukan dua perkara iaitu: a) Mengkaji semua kemungkinan subset bagi saiz m, b) Memilih subset yang mempunyai jumlah yang besar bagi pengelasan.

Berbeza dengan takrifan oleh (H. Liu & Yu 2005) iaitu pemilihan fitur bermaksud proses memilih subset fitur yang minimum daripada senarai fitur yang asal berdasarkan kriteria penilaian yang tertentu. Begitu juga takrifan dari (Nicholls & Song 2010), pemilihan fitur ialah memilih fitur berdasarkan pengukuran metrik tertentu dan fitur yang tidak relevan perlu dihapuskan berdasarkan nilai ambang yang telah ditetapkan dalam kajian. Pengurangan fitur ini bukan hanya dapat meningkatkan kecekapan prosedur latihan dan pengujian tetapi membantu meningkatkan prestasi pengelasan.

Selain daripada itu juga, terdapat definisi pemilihan fitur yang merujuk kepada mengurangkan ruang dimensi fitur seperti (B. Agarwal & Mittal 2013; Koncz & Paralic 2011). Pemilihan fitur ini dapat mengurangkan ruang fitur yang bersaiz besar dengan menyingkirkan fitur-fitur yang kurang relevan bagi menghasilkan set fitur yang bersesuaian (Koncz & Paralic 2011). Manakala bagi (B. Agarwal & Mittal

2013), pemilihan fitur ialah memilih fitur-fitur yang penting dengan menghapuskan fitur-fitur yang tidak relevan. Pengurangan vektor fitur yang mengandungi hanya fitur-fitur yang relevan dapat membantu meningkatkan kelajuan proses pengiraan dan prestasi pengelasan kaedah pembelajaran mesin (Abbasi et al. 2011; B. Agarwal & Mittal 2013). Kesimpulannya, kajian ini medapati pemilihan fitur ialah satu proses mengenal pasti dan menghapuskan fitur yang berlebihan dan tidak relevan daripada senarai fitur bagi mengecilkan ruang dimensi fitur. Proses ini membantu meningkatkan prestasi pengelasan (Alijla et al. 2018; Zhang et al. 2018).

2.3 DOMAIN ANALISIS SENTIMEN

Analisis sentimen telah mendapat popularity dalam beberapa tahun kebelakangan ini dan telah digunakan di pelbagai aplikasi. Ia telah digunakan di pelbagai bidang seperti penjagaan kesihatan, sektor kewangan, sukan, politik, perhotelan dan pelancongan serta tingkah laku pengguna (Shayaa et al. 2018).

2.4 PERINGKAT DALAM ANALISIS SENTIMEN

Sentimen adalah ungkapan pendapat, perasaan atau emosi, atau penilaian yang dibuat oleh individu yang boleh menjadi positif atau negatif atau neutral. Polariti ini dikenali sebagai orientasi sentimen, orientasi pendapat atau polariti semantik. Polariti sedemikian boleh diklasifikasikan dan diramalkan menerusi analisis sentimen. Terdapat tiga peringkat dalam analisis sentiment: -

2.4.1 Peringkat Dokumen

Pada peringkat ini bertujuan untuk mengklasifikasikan dokumen pendapat yang dinyatakan sama ada positif atau negatif sentimen. Tahap ini terhad kepada dokumen yang tidak mengukur atau membandingkan pelbagai atribut kerana, pada tahap ini, seluruh dokumen mewakili sentimen ke arah atribut tunggal (Khan et al. 2016; B. Liu 2012; Trivedi & Dey 2018)

2.4.2 Peringkat Ayat

Ketika proses pengelasan sentimen, ayat digunakan untuk menentukan sentimen positif, negatif atau neutral terhadap produk atau perkhidmatan. Sentimen peringkat ayat menangani klasifikasi subjektif, dan membezakan klasifikasi sentimen subjektif dan objektif, di mana ayat subjektif mendedahkan pendapat atau sentimen, dan kalimat yang objektif menyampaikan maklumat yang benar (B. Liu 2012; Trivedi & Dey 2018).

2.4.3 Peringkat Aspek

Analisis sentimen pada peringkat ini berdasarkan pada ciri, atau sifat, teks di mana ciri, atau perkataan, diambil sama ada positif atau sentimen negatif. Ini adalah analisis yang lebih halus, di mana semua ciri-ciri, yang diambil bersama-sama, memberikan gambaran tentang pemberat sentimen keseluruhan pendapat. Analisis sentimen tahap aspek mendefinisikan pendapat sebagai positif, negatif atau neutral, berdasarkan kata-kata atau ciri pemberat sentimen (Khan et al. 2016; Trivedi & Dey 2018).

2.5 KEPENTINGAN ANALISIS SENTIMEN

Teknologi analisis sentimen memainkan peranan penting kepada pengguna ataupun organisasi. Teknologi ini mempunyai skop yang lebih meluas apabila dilaksanakan dalam aplikasi yang lebih praktikal di mana pengguna baru boleh mendapatkan maklumat seperti ringkasan ulasan pengguna yang telah membeli sesuatu produk. Maklumat ulasan ini membantu pengguna baru membuat penilaian dan keputusan untuk membeli atau sebaliknya. Selain daripada itu, pengguna juga boleh membuat perbandingan ulasan yang mengandungi kelebihan dan kekurangan bagi sesuatu produk. Manakala bagi organisasi atau perniagaan, teknologi ini membantu mereka membuat penilaian mengenai kelemahan dan kekurangan produk yang dihasilkan. Maklumat ini amat berguna bagi membantu bahagian pemasaran, penanda aras produk, reka bentuk dan pembangunan bagi memperbaiki mutu dan perkhidmatan produk supaya dapat memenuhi kepuasan dan kehendak pengguna (Shelke et al. 2012).

2.6 TEKNIK PEMPROSESAN TEKS

Antara langkah yang awal dan dianggap penting dalam proses pengelasan sentimen adalah pemprosesan teks. Pada fasa ini, pemprosesan teks berperanan dalam pembersihan dan penyediaan teks untuk tujuan pemilihan fitur dan seterusnya pengelasan sentimen (Haddi et al. 2013). Dengan kemajuan teknologi maklumat yang pesat, penggunaan media sosial kini menjadi pilihan utama. Justeru itu, menjadi cabaran utama dalam melakukan pemprosesan teks dari media sosial kerana data teks daripada medium ini mengandungi teks yang hingar dan teks yang tidak relevan seperti tag HTML, skrip dan pengiklanan. Walaubagaimanapun dengan melakukan teknik pemprosesan teks yang betul dapat menghapuskan teks yang hingar mahupun teks yang tidak relevan seterusnya akan dapat memperoleh keputusan pengelasan sentimen yang lebih tinggi (Pradha et al. 2019). Menurut (Pradha et al. 2019), berikut adalah senarai teknik pemprosesan teks yang boleh digunakan.

i. Penukaran Teks ke Huruf Kecil

Proses di mana menukar semua huruf besar kepada huruf kecil. Tujuan utamanya adalah supaya perkataan yang sama lebih mudah untuk dikenal pasti. Ia membantu dalam mengurangkan pertindanan perkataan yang sama dan sekali gus boleh membantu mengurangkan dimensi fitur.

Contoh: "*I hated the da Vinci code, the Movie was boring and made me sad"* ditukar kepada "*i hated the da vinci code, the movie was boring and made me sad*".

ii. Pembuangan Tanda @

Perbualan dalam laman web dan forum, pengguna biasanya menggunakan simbol @ dan diikuti oleh perkataan tertentu yang menggambarkan lokasi atau tempat ciapan ulasan itu dilakukan. Dalam media sosial, simbol @ juga boleh digunakan untuk merujuk nama pemilik akaun sosial media yang merujuk kepada individu atau nama organisasi dan ia yang berfungsi seakan-akan URL. Simbol @ dan perkataan di belakangnya akan dikenal pasti dan dibuang dari ulasan pengguna kerana ia tidak menyumbangkan kepada maklumat berkaitan mengenai sentimen ulasan berkenaan.

Contoh: "*i am impressed with the actor @CGJase*" ditukar kepada "*i am impressed with the actor*".

iii. Pembuangan Tanda Baca

Proses pembuangan tanda baca dari ulasan asal kerana ia juga tidak menyumbangkan kepada maklumat berkaitan sentimen ulasan berkenaan. ini merangkumi semua tanda baca yang terdapat dalam penggunaan bahasa antaranya ".", ",", "!", "-", ":", ";", "?", "/", "()", "{}" dan " ' ".

Contoh: "*it's too bad!! Do you agree?*" ditukar kepada "*its too bad Do you agree*".

iv. Pembuangan Tanda #

Tanda pagar atau simbol "#" kerap digunakan terutamanya dalam media sosial Facebook, Twitter dan Instagram pada masa sekarang. Tanda pagar ini biasanya diikuti oleh perkataan tertentu yang merujuk kepada tema atau hubungan kepada topik yang popular. Tanda pagar ini akan dikenal pasti dan dihapuskan daripada teks ulasan kerana ia tidak menyumbang kepada maklumat berkaitan sentimen ulasan.

Contoh : "*how can i watch the #davincicode freely*" ditukar kepada "*how can i watch the freely*".

v. Pembuangan Simbol

Simbol yang selain abjad atau nombor dihapuskan dari ulasan asal kerana ia juga tidak menyumbangkan kepada maklumat berkaitan sentimen ulasan berkenaan. Sebagai contoh, adalah simbol $.

Contoh: "*how much it will cost? $20 only*" ditukar kepada "*how much it will cost? 20 only*".

vi. ***Stemming***

Stemming merupakan penukaran perkataan kepada kata dasar sesuatu perkataan. Kaedah ini dilaksanakan dengan menghapuskan pengawalan dan pengakhiran sesuatu perkataan dan mencari kata dasar sebenar bagi perkataan tersebut. Ia membantu mengenal pasti perkataan yang mempunyai kata dasar yang sama bagi mengelakkan pertindanan. Secara tidak langsung ia membantu mengurangkan jumlah perkataan yang akan dipilih untuk dijadikan fitur dan masa pemprosesan dapat dikurangkan.

Contoh: "*it is was a disappointing experience and it will be shared to everyone*" ditukar kepada "*it is was a disappoint experi and it will be share to everyone".*

vii. ***Lemmatization***

Proses untuk mendapatkan asal usul sesuatu perkataan melalui analisis morfologi dikenali sebagai *lemmatization*. Ia melibatkan proses pembuangan pengawalan, pembuangan pengakhiran dan perbandingan dengan koleksi perkataan asal untuk melihat padanan dengan perkataan yang diproses. Proses ini membantu mengenal pasti perkataan yang mempunyai asal usul yang sama bagi mengelakkan pertindanan. Proses ini membantu mengurangkan saiz perkataan yang akan dipilih untuk dijadikan fitur dan sekali gus menjimatkan masa pemprosesan.

Contoh : "*yes i did hope for update soon and will have a meeting tomorrow*" ditukar kepada "*yes i do hope for update soon and will have a meet tomorrow*".

viii. Pembetulan Ejaan

Setiap ejaan ayat ulasan akan disemak untuk sebarang kesilapan. Jika terdapat kesilapan, pembetulan ejaan akan dilaksanakan.

Contoh : "*that was the bset movie*" ditukar kepada "*that was the best movie*".

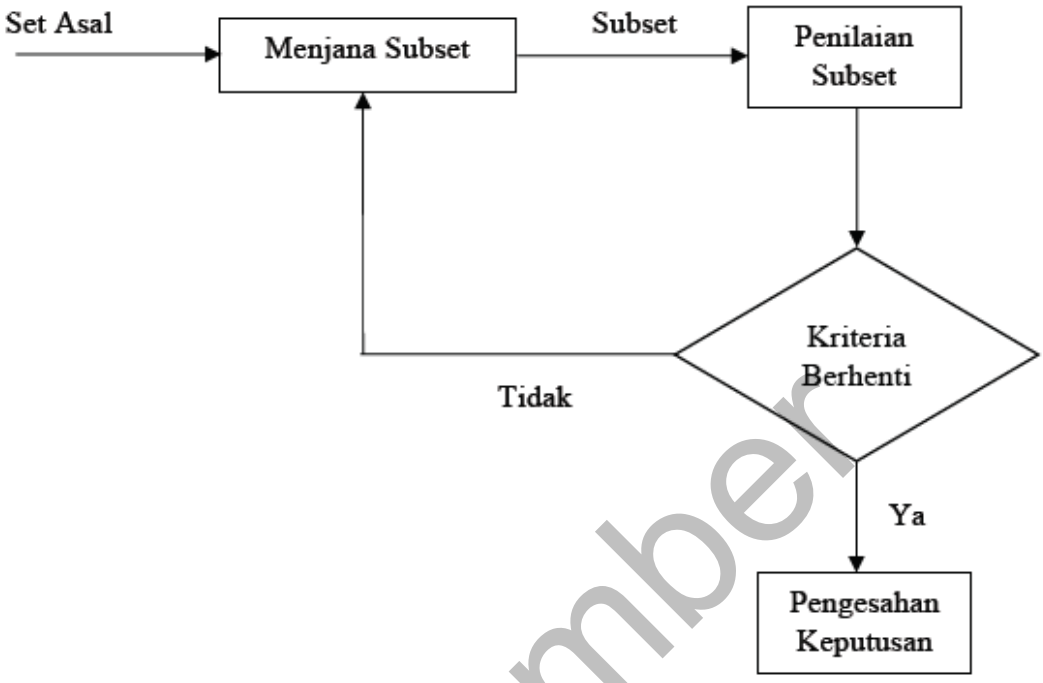
Teknik pemprosesan teks ini terbahagi kepada dua kategori utama iaitu pertama ia dikenali sebagai pemprosesan linguistik yang terdiri daripada proses penukaran ke huruf kecil, , penghapusan @, penghapusan tanda pagar, penghapusan simbol, penghapusan tanda baca. Manakala kategori kedua pula dikenali sebagai pemprosesan bahasa tabii yang terdiri daripada stemming, lemmatization, pembetulan ejaan (Khader et al. 2019).

2.7 TEKNIK PEMILIHAN FITUR

Fungsi teknik pemilihan fitur ialah memilih subset fitur optimum daripada set fitur asal berdasarkan kriteria penilaian tertentu. Proses pemilihan fitur yang bersesuaian dapat meningkatkan prestasi pengelasan, meningkatkan kecekapan pembelajaran dan membantu menghasilkan model yang lebih baik.

2.7.1 Proses Pemilihan Fitur

Terdapat empat langkah utama dalam algoritma pemilihan fitur. Rajah 2.1 menunjukkan langkah-langkah yang terlibat dalam pemilihan fitur yang terdiri daripada penjanaan subset, penilaian subset, kriteria pemberhentian dan pengesahan keputusan (H. Liu & Yu 2005; Stańczyk 2015).



Rajah 2.1 Langkah dalam proses pemilihan fitur

Sumber: (H. Liu & Yu 2005)

a. Penjanaan Subset

Senarai berikut merupakan langkah-langkah yang digunakan bagi melaksanakan penjanaan subset;

i. Carian Lengkap

Set fitur yang diperoleh adalah berasaskan kriteria penilaian apabila melalui kaedah carian ini. Melalui teknik ini akan menghasilkan keputusan dengan cepat dan mudah untuk dilaksanakan di mana fitur akan tambah dan dibuang satu persatu. Di antara kaedah pemilihan fitur yang menggunakan carian lengkap adalah seperti di gunakan oleh (H. Liu et al. 2010) dan (H. Liu & Zhao 2012). Walaupun carian dengan kaedah ini adalah secara menyeluruh, terdapat beberapa heuristik berbeza digunakan seperti carian pertama terbaik dan carian rasuk yang boleh digunakan.

ii. Carian Berturutan

Kaedah ini akan mengabaikan kesempurnaan dan terdedah dengan risiko untuk kehilangan subset yang berkualiti. Akan tetapi, teknik ini dapat menghasilkan keputusan dengan cepat dan mudah untuk dilaksanakan. Dalam teknik ini, fitur akan ditambah atau dibuang satu persatu. Ia merupakan variasi kepada pendakian bukit rakus seperti penghapusan dari belakang secara berturutan, pemilihan ke hadapan secara berturutan, pemilihan terapung dari belakang secara berturutan dan juga pemilihan terapung dari hadapan secara berturutan (Mao & Tsang 2013; Rahman & Fong 2018)

iii. Carian Rawak

Penyelidikan oleh (Dong et al. 2018) bermula dengan carian subset fitur yang dipilih secara rawak dan seterusnya dengan carian dari dua kaedah. Bagi kaedah yang pertama di mana merupakan carian berturutan dan fungsi rawak dimasukkan ke dalam carian berturutan piawai. Kaedah ini juga dikenali sebagai tidak-berketentuan. Kaedah yang kedua pula adalah dengan menghasilkan subset berturutan secara rawak. Kaedah ini juga dikenali sebagai algoritma Las Vegas.

Isu utama dalam proses pemilihan fitur ialah mempunyai ruang carian yang bersaiz besar dan kompleks. Oleh disebabkan hal ini, kaedah carian lengkap tidak sesuai digunakan kerana memerlukan sumber pemprosesan komputer yang tinggi dan memerlukan masa pemprosesan yang agak lama walaupun set data yang digunakan adalah kecil. Manakala bagi kaedah pemilihan berturutan ke hadapan dan juga pemilihan berturutan ke belakang pula mempunyai masalah sendiri kerana fitur yang dihapuskan atau yang telah dipilih tidak boleh dipilih semula untuk langkah seterusnya (Xue et al. 2016). Untuk mengatasi masalah-masalah ini, carian rawak digunakan oleh kebanyakan metaheuristik berasaskan populasi untuk menghasilkan lebih dari satu penyelesaian dengan mudah (Aghdam & Heidari 2015; Xue et al. 2016).

b. Penilaian Subset

Penilaian akan dilaksanakan dengan menggunakan kriteria penilaian sebaik sahaja subset fitur dijana. Terdapat dua jenis kriteria penilaian yang wujud iaitu pertama merupakan kriteria yang bersandar dan yang kedua adalah yang tidak bersandar. Kriteria bersandar digunakan dalam teknik pembalut. Algoritma pengelasan diperlukan dalam kriteria ini. Subset fitur yang memberikan prestasi pengelasan yang tertinggi akan dipilih sebagai subset fitur yang terbaik. Kriteria tidak bersandar biasanya digunakan dalam teknik penapis. Kualiti subset adalah berdasarkan ciri-ciri data latihan. Empat ciri yang biasa digunakan dalam kajian adalah termasuk ukuran maklumat, ukuran jarak, ukuran konsistensi dan ukuran sandaran.

c. Kriteria Pemberhentian

Kriteria ini berperanan untuk mengambil keputusan sama ada untuk memberhentikan proses penilaian fitur atau meneruskan proses pemilihan fitur. Terdapat beberapa kaedah yang biasa digunakan sebagai kriteria berhenti untuk mengenal pasti sama ada proses carian boleh dihentikan atau tidak. Antara kedah yang biasa digunakan adalah dengan melihat sama ada nilai sasaran telah dicapai (contoh bilangan generasi atau bilangan minimum atau maksimum fitur yang diperlukan). Kaedah yang lain adalah dengan melihat jika proses penambahan atau penghapusan fitur seterusnya tidak dapat menghasilkan subset yang lebih baik atau subset fitur yang lebih baik tidak dapat dijana.

d. Penilaian Keputusan

Kriteria ini digunakan untuk menilai sama ada subset yang dijana adalah sah, sama ada dinilai berdasarkan kajian yang dilaksanakan sebelum ini atau pemerhatian secara tidak langsung dalam prestasi pemilihan. Berdasarkan kajian semasa, penyelidikan menggunkan penilaian prestasi ke atas teknik pemilihan fitur melalui keputusan ketepatan pengelasan. Sekiranya ketepatan pengelasan dikekalkan atau dipertingkatkan dengan fitur yang dipilih, ia dianggap sebagai sah. Kaedah ini banyak di aplikasikan dalam kajian-kajian sebelum ini seperti (Das et al. 2018; Shunmugapriya & Kanmani 2017)

2.7.2 Kaedah Terselia

Pemilihan fitur boleh dikategorikan sebagai kaedah terselia, tidak terselia dan separuh terselia. Kaedah pemilihan fitur terselia juga boleh dikategorikan dengan lebih lanjut kepada model penapis, model pembalut dan model terbenam. Dengan maklumat terselia, fitur biasanya dinilai dengan menganggarkan kadar hubungan di antara fitur dan juga kelas sasaran.

Fasa latihan dalam pengelasan model bergantung sepenuhnya kepada fitur yang dipilih. Kebiasaannya, pengelasan dilatih berasaskan subset fitur yang dipilih dari fitur terselia yang terpilih. Pemilihan fitur boleh dilakukan dengan tidak berasaskan algoritma pembelajaran (kaedah penapis) atau ia bergantung kepada algoritma pembelajaran di mana ia menggunakan kelebihan model pengelasan untuk menilai kualiti fitur yang dipilih (kaedah pembalut), atau di benam bersama fitur yang dipilih dalam algoritma pembelajaran (kaedah terbenam). Akhir sekali, pengelasan yang telah dipilih digunakan untuk mengklasifikasikan item-item yang tersembunyi dalam set ujian dengan menggunakan fitur yang dipilih. Fokus utama dalam kajian ini merupakan teknik terselia untuk masalah klasifikasi.

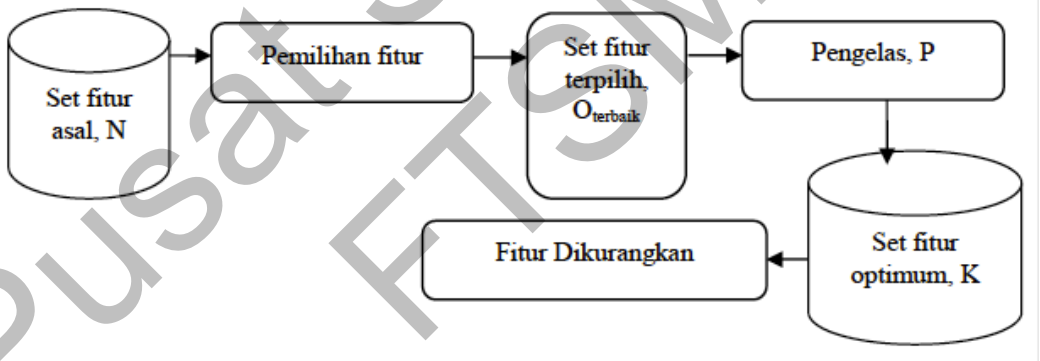
2.7.3 Kaedah Strategi Pemilihan

Teknik pemilihan fitur secara umumnya boleh dikategorikan kepada tiga iaitu penapis, pembalut dan terbenam. Bagi teknik penapis ia menilai subset yang dipilih dengan menggunakan ciri-ciri umum data, manakala teknik pembalut menggunakan prestasi pengelasan untuk menilai subset fitur. Fokus utama teknik pembalut adalah untuk meningkatkan prestasi algoritma pengelasan dan secara umumnya memerlukan sumber pengkomputeran yang lebih tinggi daripada model penapis. Seterusnya bagi teknik terbenam pula, pemilihan fitur terbenam dalam model pembelajaran. Terdapat juga beberapa kajian mengenai kaedah pemilihan fitur menyatakan terdapat satu lagi tambahan dalam strategi pemilihan teknik iaitu teknik pemilihan fitur hibrid (Ang et al. 2016; Shen et al. 2018). Teknik hibrid merupakan gabungan daripada pelbagai teknik pemilihan fitur seperti penapis, pembalut dan terbenam. Teknik ini memanipulasi kelebihan teknik-teknik pemilihan fitur yang dipilih dengan mengaplikasikan penilaian kriteria yang berlainan pada tahap yang berbeza (H. Liu &

Yu 2005). Teknik pembalut dan penapis merupakan teknik yang kerap digunakan dalam kajian-kajian sebelum ini.

2.7.4 Teknik Penapis

Teknik penapis merupakan salah satu teknik pemilihan fitur yang paling awal dalam kaedah mesin pembelajaran (Chen et al. 2020; Stańczyk 2015). Teknik ini menggunakan heuristik berdasarkan ciri-ciri umum data bagi menilai subset fitur. Kebiasaannya proses teknik penapis lebih cepat berbanding dengan teknik pembalut. Selain daripada itu, proses pengiraannya mudah, pantas dan tidak bergantung kepada algoritma pengelasan (Saeys et al. 2007). Oleh itu, teknik penapis lebih praktikal untuk digunakan bagi data yang berdimensi tinggi (Chen et al. 2020; Stańczyk 2015). Kelemahan teknik ini ialah tidak berinteraksi dengan fitur atau pun algoritma pengelas. Terbahagi kepada dua kategori iaitu univariat dan multivariat. Rajah 2.2 menunjukkan proses pemilihan fitur bagi teknik penapis.



Rajah 2.2 Proses pemilihan fitur bagi teknik penapis

Sumber: (Chen et al. 2020)

Teknik penapis univariat menilai fitur berdasarkan penilaian kedudukan fitur secara individu. Manakala teknik multivariat menilai subset fitur dalan bentuk berkumpulan berdasarkan kriteria penilaian tertentu. Oleh kerana, univariat menilai fitur secara individu maka proses penilaian lebih pantas berbanding dengan teknik multivariat. Kaedah ini banyak digunakan dalam pelbagai bidang (Manek et al. 2016; Okun 2011; Wu et al. 2017). Bagaimanapun terdapat kelemahan teknik univariat iaitu ia tidak mengambil kira kebergantungan antara fitur. Ini berbeza dengan teknik multivariat iaitu mengambil kira kebergantungan antara fitur-fitur. Keadaaan ini

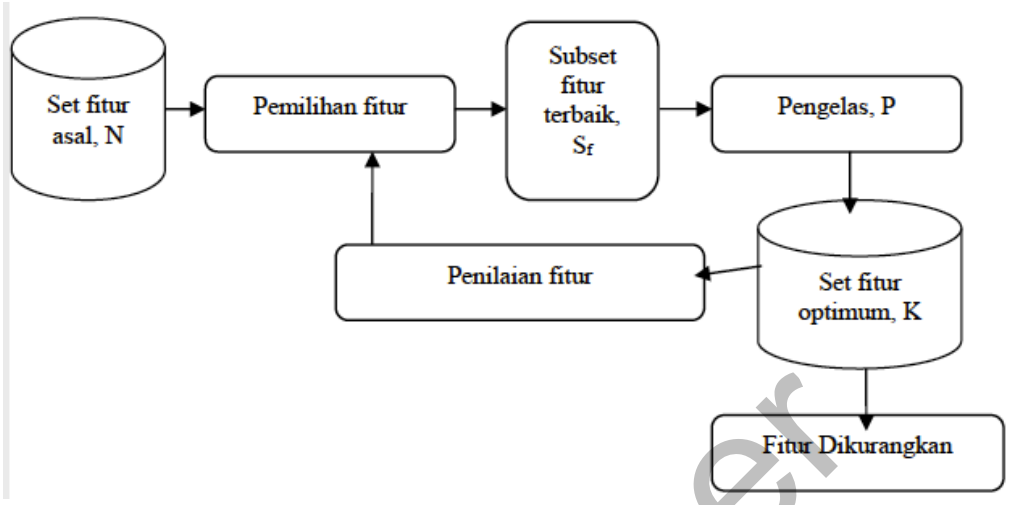
menyebabkan proses pemilihan oleh teknik multivariat menjadi perlahan berbanding dengan teknik univariat. Antara contoh penapis univariat adalah kotak-chi (Sumaiya Thaseen & Aswani Kumar 2017), jarak euclidean (Sharif et al. 2017), capaian maklumat (Zhu et al. 2017) dan kadar capaian (Nagpal & Gaur 2015). Antara contoh penapis multivariate pula adalah seperti pemilihan fitur berasaskan hubungan (Jain et al. 2018), penapis litupan-Markov (Yu et al. 2017) dan penapis berasaskan hubungan pantas (Egea et al. 2018).

Pendekatan teknik penapis adalah ia memisahkan pemilihan fitur daripada pengelas pembelajaran daripada mempengaruhi algoritma pembelajaran yang tidak mempunyai kesan terhadap pengaruh algoritma pemilihan fitur.

2.7.5 Teknik Pembalut

Teknik pembalut amat bergantung kepada prestasi pengelasan algoritma yang dipilih untuk menilai kualiti subset fitur yang dihasilkan. Dengan menggunakan algoritma pengelasan pilihan, teknik pembalut dilakukan dalam dua langkah iaitu penjanaan dan penilaian subset fitur. Dua langkah ini diulang sehingga kriteria berhenti dicapai. Ia bermula dengan komponen carian akan menjana subset fitur dan kemudiannya algoritma pengelasan akan menilai kualiti fitur yang dipilih berdasarkan prestasi pengelasan. Keseluruhan proses ini berlaku secara berulang-ulang sehingga prestasi pengelasan yang terbaik dicapai atau bilangan jumlah fitur yang diperlukan telah diperoleh. Kemudian subset fitur yang memberikan keputusan pengelasan yang tertinggi akan dipilih sebagai fitur bersaiz minimum.

Akan tetapi antara masalah yang dihadapi oleh teknik ini ialah ia memerlukan sumber pengkomputeran yang tinggi dan tidak praktikal apabila dimensi adalah bersaiz besar. Disebabkan itu terdapat kaedah strategi yang lain diwujudkan seperti carian berturutan (Guyon & Elisseeff 2003; Stańczyk 2015), carian panjat-bukit, carian pertama-terbaik (Arai et al. 2016) dan algoritma genetik (Sastry et al. 2005) diperkenalkan untuk mendapatkan prestasi pengelasan. Akan tetapi, ruang carian masih lagi besar untuk set data yang mempunyai dimensi besar. Disebabkan itu teknik pembalut ini agak kurang digunakan. Rajah 2.3 menunjukkan proses teknik pembalut.



Rajah 2.3 Proses pemilihan fitur bagi teknik pembalut

Sumber: (Chen et al. 2020)

2.8 TEKNIK PEMILIHAN FITUR DALAM PENGELASAN SENTIMEN

Dalam bidang pengelasan sentimen, pemilihan fitur yang efektif adalah penting untuk memastikan proses pembelajaran lebih efisien dan lebih tepat (Adachi et al. 2016). Terdapat banyak algoritma yang diperkenalkan oleh kajian-kajian sebelum ini, akan tetapi hanya sedikit sahaja kajian yang melibatkan pengelasan teks. Bahagian ini akan menerangkan kajian yang terkini yang telah dilaksanakan untuk pemilihan fitur untuk pengelasan teks.

Kajian ini boleh dikategorikan kepada beberapa cara mengikut pelbagai kriteria. Sebagai contoh, kajian ini boleh dikategorikan daripada perspektif penyeliaan iaitu terselia dan tidak terselia atau berdasarkan kaedah peringkat seperti kaedah peringkat tunggal atau kaedah peringkat subset. Daripada perspektif lain, kajian ini juga boleh dikategorikan berasaskan strategi carian yang digunakan kepada teknik pembalut atau teknik penapis. Teknik penapis boleh dibahagikan kepada dua iaitu berasaskan peringkat subset, sama ada ia akan memulangkan nilai peringkat atau subset yang juga dikenali sebagai univariat dan multivariat.

2.8.1 Teknik Penapis Univariat dan Multivariat

Kaedah peringkat fitur tunggal atau dikenali sebagai univariate merupakan kaedah ringkas untuk pemilihan fitur. Dalam kaedah ini, skor menggambarkan kepentingan fitur tunggal yang diukur dengan ciri-ciri yang telah dikenal pasti. Semua fitur akan disusun mengikut skor dan pemilihan fitur dilaksanakan dengan memilih sekumpulan fitur yang mempunyai fitur yang terbaik. Biasanya, jumlah fitur yang terbaik ini adalah berdasarkan keperluan.

Berasaskan pendekatan ini, (Wang et al. 2015) mencadangkan kaedah dua peringkat pemilihan fitur. Dalam peringkat pertama, ia telah memilih fitur dengan menggunakan kaedah pemilihan fitur yang dinamakan sebagai darjah sumbangan fitur. Kaedah ini digunakan untuk mengurangkan jumlah fitur dengan memilih fitur yang mempunyai darjah sumbangan yang tinggi ke atas pengelasan. Dalam peringkat kedua, ia menggunakan kaedah pengindeksan semantik terpendam untuk membangunkan ruang vektor konseptual yang baru. Terdapat kajian yang mengkaji dan membandingkan pelbagai dimensi kaedah pengurangan dalam fasa awal pemprosesan (Imambi & Sudha 2011). Ia juga mencadangkan skema pemberat fitur yang dinamakan sebagai pemberat berkaitan global. Seni binanya terdiri daripada lapisan awal pemprosesan dan lapisan pemilihan fitur.

Dalam kajian yang dilaksanakan oleh (Pinheiro et al. 2012), ia mencadangkan kaedah penapisan yang dinamakan sebagai sekurang-kurang satu fitur. Kaedah ini memfokuskan fitur tertentu untuk memastikan setiap dokumen dalam set latihan digambarkan sekurang-kurangnya oleh satu fitur. Selain itu terdapat kaedah pemilihan fitur penapis berasaskan kebarangkalian yang dinamakan sebagai pemilih fitur pembezaan untuk pengelasan teks. Kaedah ini memilih fitur tersendiri dan menghapuskan fitur yang tidak berinformasi berdasarkan kriteria keperluan (Uysal & Gunal 2012).

Seterusnya kajian oleh (Basu & Murthy 2012) pula mencadangkan kaedah pemilihan fitur berasaskan persamaan di antara terma dan kelas untuk pengelasan teks. Daripada kaedah yang dicadangkan, setiap terma perbendaharaan kata akan diberikan skor berdasarkan kesamaannya terhadap semua kelas. Semua terma akan

disusun berdasarkan skor individu. Kemudian, beberapa terma yang mempunyai skor yang tertinggi akan dipilih sebagai fitur penting. Sekiranya terma tidak pernah ditemui di dalam kelas, ia akan diberikan nilai skor negatif yang menggambarkan terma ini tidak ada kaitan dengan kelas tersebut.

2.8.2 Teknik Metaheuristik

Teknik pengoptimuman berasaskan metaheuristik adalah kaedah yang meningkatkan prestasi penyelesaian secara berulang-ulang (Asghari & Navimipour 2015). Sesetengah daripada teknik metaheuristik ini adalah berdasarkan pemerhatian ke atas fenomena alam semula jadi seperti pengoptimuman koloni semut, algoritma kelawar dan pengoptimuman pendebungaan bunga. Metaheuristik ini kemudiannya digunakan dalam kaedah pemilihan fitur pembalut. Kaedah ini semakin mendapat perhatian kebelakangan ini (Alijla et al. 2018) kerana ia cuba untuk menghasilkan penyelesaian yang lebih baik dengan mengaplikasikan pengetahuan yang diperoleh daripada penyelesaian semula jadi. Pendekatan pembalut untuk pemilihan fitur bergantung kepada prestasi klasifikasi untuk menilai kualiti fitur yang telah dipilih. Kaedah pembalut mempunyai dua langkah utama iaitu: (1) mencari subset fitur dan (2) menilai fitur yang dipilih. Kedua-dua langkah ini akan diulang sehingga kriteria berhenti dipenuhi. Ia bermula dengan penghasilan subset, kemudian pengelasan akan menilai subset yang dihasilkan.

Terdapat kajian terdahulu yang mencadangkan kaedah pemilihan fitur berasaskan pengoptimuman pengelompokan partikel dan pengelasan k-jiran terdekat (Aghdam & Heidari 2015). Selain itu, berdasarkan kajian dari (Ghareb et al. 2016) pula mencadangkan pendekatan hibrid pemilihan fitur berasaskan algoritma genetik untuk pengelasan teks Arab menggunakan model pembalut. Dalam langkah pertama, enam kaedah penilaian fitur digunakan dalam masa yang sama untuk memilih subset fitur. Kemudian algoritma genetik yang telah dipertingkatkan digunakan untuk mengoptimumkan subset yang telah dipilih.

Teknik pemilihan subset fitur yang optimum menggunakan algoritma kunang-kunang (*fireflies*) bagi masalah analisis sentimen juga telah membuktikan kemampuan algoritma metaheuristik ini. (Akshi & Renu 2017). Manakala dalam kajian yang

dilakukan oleh (Tubishat et al. 2019) di mana penyelidik menambah baik algoritma pengoptimuman paus bagi tujuan pemilihan fitur dua jenis data analisis sentimen iaitu dalam Bahasa Arab dan Bahasa Inggeris. Selain itu, penghibridiban di antara algoritma pengoptimuman koloni semut dan k-Jiran terdekat telah berjaya meningkatkan hasil pengelasan sentimen kerana telah diaplikasikan pada proses pemilihan fitur (Ahmad et al. 2019b). Kajian-kajian ini melaporkan bahawa kaedah mereka adalah lebih efektif berbanding dengan penggunaan kaedah penapis.

2.9 ALGORITMA PENDEBUNGAAN BUNGA

Sebilangan besar tanaman adalah tanaman berbunga dan terdapat lebih dari 250 000 spesies tanaman berbunga di seluruh dunia, di mana pendebungaan merupakan strategi penghasilan utama tanaman (Kawasaki & Bell 1991). Pendebungaan adalah proses memindahkan debunga dari satu bunga ke bunga lain oleh angin atau ejen pendebunga seperti serangga, rama-rama, lebah, burung dan kelawar. Tumbuhan berbunga telah berkembang untuk menghasilkan nektar atau madu untuk menarik ejen pendebunga dan memastikan pendebungaan berlaku (Glover 2008). Di samping itu, beberapa ejen pendebunga dan spesies tumbuhan seperti burung adu dan tanaman berbunga ornithophilous membentuk beberapa pemalaran bunga evolusi bersama (Kawasaki & Bell 1991). Berdasarkan ciri utama pendebungaan, algoritma pendebungaan bunga telah dikembangkan (X. S. Yang 2012).

Seterusnya merupakan penerangan dengan lebih mendalam tentang proses asas dalam pendebungaan bunga di mana pendebungaan mengambil dua bentuk utama iaitu biotik dan abiotik. Pendebungaan biotik merupakan bentuk utama pendebungaan juga disebut pendebungaan silang, dilakukan oleh ejen pendebunga seperti serangga, burung dan lain-lain. Hampir 90% tanaman berbunga menggunakan pendebungaan bentuk ini. Ketika ejen pendebunga bergerak dan bahkan terbang dengan pelbagai kelajuan, pergerakan debunga agak jauh. Pendebungaan sedemikian juga boleh dianggap sebagai pendebungaan global dengan pengaplikasian sifat penerbangan Lévy (Ozsoydan & Baykasoglu 2019; Pavlyukevich 2007; X. S. Yang 2012). Sekiranya debunga dikodkan sebagai vektor penyelesaian, tindakan ini adalah setara dengan carian global.

Pendebungaan abiotik juga disebut pendebungaan kendiri, yang tidak memerlukan ejen pendebungaan. Dianggarkan bahawa sekitar 10% tanaman bunga mengambil bentuk pendebungaan ini. Oleh kerana pendebungaan bentuk ini lebih cenderung bersifat lokal dan pendebungaan kendiri, ia dapat dicapai dengan penyebaran oleh angin (Glover 2008; Kawasaki & Bell 1991). Jarak yang dilalui oleh gerakan tempatan kebiasaanya lebih pendek, maka ia boleh dianggap sebagai pencarian tempatan.

Kemalaran bunga adalah baik bagi tanaman dan ejen pendebunga seperti burung madu untuk menjalin kerjasama bagi menjimatkan tenaga sekaligus menghasilkan kejayaan. Oleh itu, kemalaran bunga telah berkembang dan bagi kes ini, ejen pendebunga hanya mengunjungi sekumpulan jenis bunga yang tetap tanpa membuang tenaga untuk meneroka jenis bunga baru, sementara tanaman bunga berkembang agar dapat menyediakan nektar atau madu yang mencukupi kepada ejen pendebunga sehingga mendorong kunjungan yang kerap oleh ejen pendebunga dan dengan demikian memaksimumkan kejayaan pembiakan mereka (Glover 2008; Kawasaki & Bell 1991).

Ciri-ciri di atas telah digunakan untuk merancang pembangunan algoritma pengoptimuman, yang disebut algoritma pendebungaan bunga (APB) (X. S. Yang 2012). Ciri-ciri utama dan komponen algoritma APB dapat diringkaskan dalam Jadual 2.1, yang menunjukkan hubungan atau kesetaraan antara istilah pengoptimuman dan konteks bunga. Dengan komponen dan ciri ini, penerangan tentang algoritma pendebungaan bunga akan dilakukan secara terperinci.

Jadual 2.1 Konteks pendebungaan bunga dan komponen pengoptimuman

Pendebungaan Bunga	**Komponen Penoptimuman**
Pendebunga (Serangga, Angin)	Pergerakan/pengubahsuaian pemboleh ubah
Biotik	Carian Global
Abiotik	Carian Tempatan
Penerbangan Levy	Saiz Langkah
Debunga	Vektor Penyelesaian
Pemalaran Bunga	Persamaan dengan vektor penyelesaian
Evolusi bunga	Evolusi iterasi penyelesaian
Pembiakan semula optimum bunga	Set penyelesaian optimum

Algoritma APB merupakan algoritma yang diilhamkan daripada alam semula jadi yang meniru tingkah laku pendebungaan utama bagi tanaman berbunga. Empat peraturan yang diidealisasikan oleh (X. S. Yang 2012). **Peraturan 1**, pendebungaan global melibatkan biotik dan pendebungaan silang di mana ejen pendebunga akan membawa debunga berdasarkan sifat penerbangan Lévy. **Peraturan 2**, pendebungaan tempatan melibatkan abiotik dan pendebungaan kendiri. Seterusnya bagi **Peraturan 3** pula, kemalaran bunga boleh dianggap sebagai kebarangkalian pembiakan yang berkadar dengan persamaan bagi kedua-dua bunga. Akhir sekali, **Peraturan 4**, kebarangkalian pertukaran $p \in [0, 1]$ dapat dikawal antara pendebungaan tempatan dan pendebungaan global kerana beberapa faktor luaran, seperti angin. Pendebungaan tempatan mempunyai pecahan p yang signifikan dalam keseluruhan aktiviti pendebungaan.

2.9.1 Carian Global bagi APB (Biotik)

Ejen pendebunga seperti burung dan kelawar dapat memindahkan debunga pada jarak jauh semasa pendebungaan biotik, ini akan dapat memastikan pendebungaan yang kepelbagai dan sesuai untuk pembiakan. Oleh itu, peraturan APB yang pertama (**Peraturan 1**) dan yang ketiga (**Peraturan 3**) dapat dirumuskan seperti berikut:

$$x_i^{t+1} = x_i^t + L\,(g^* - x_i^t) \qquad \ldots(2.1)$$

di mana x^t_i adalah debunga atau vektor penyelesaian pada lelaran t dan g^* adalah penyelesaian terbaik dijumpai di antara semua penyelesaian pada lelaran semasa. Parameter L adalah kekuatan bagi pendebungaan, yang pada dasarnya adalah ukuran langkah. Oleh kerana ejen pendebunga bergerak jarak jauh dengan pelbagai jarak, penerbangan Lévy dapat menjadi simulator yang cekap untuk ciri ini (X. S. Yang 2012); iaitu, L dapat diambil dari taburan Lévy seperti berikut:

$$L \sim \frac{\lambda \Gamma(\lambda) \sin\left(\frac{\pi\lambda}{2}\right)}{\pi} \frac{1}{s^{1+\lambda}}, \qquad (s \ll 0) \qquad \ldots(2.2)$$

di mana $\boldsymbol{\Gamma}$ (λ) menunjukkan fungsi gamma asas dan taburan ini sah bagi langkah yang besar $s > 0$. Pada kebiasaannya, disarankan agar menggunakan λ = 1.5 (X. S. Yang 2012).

2.9.2 Carian Tempatan bagi APB (Abiotik)

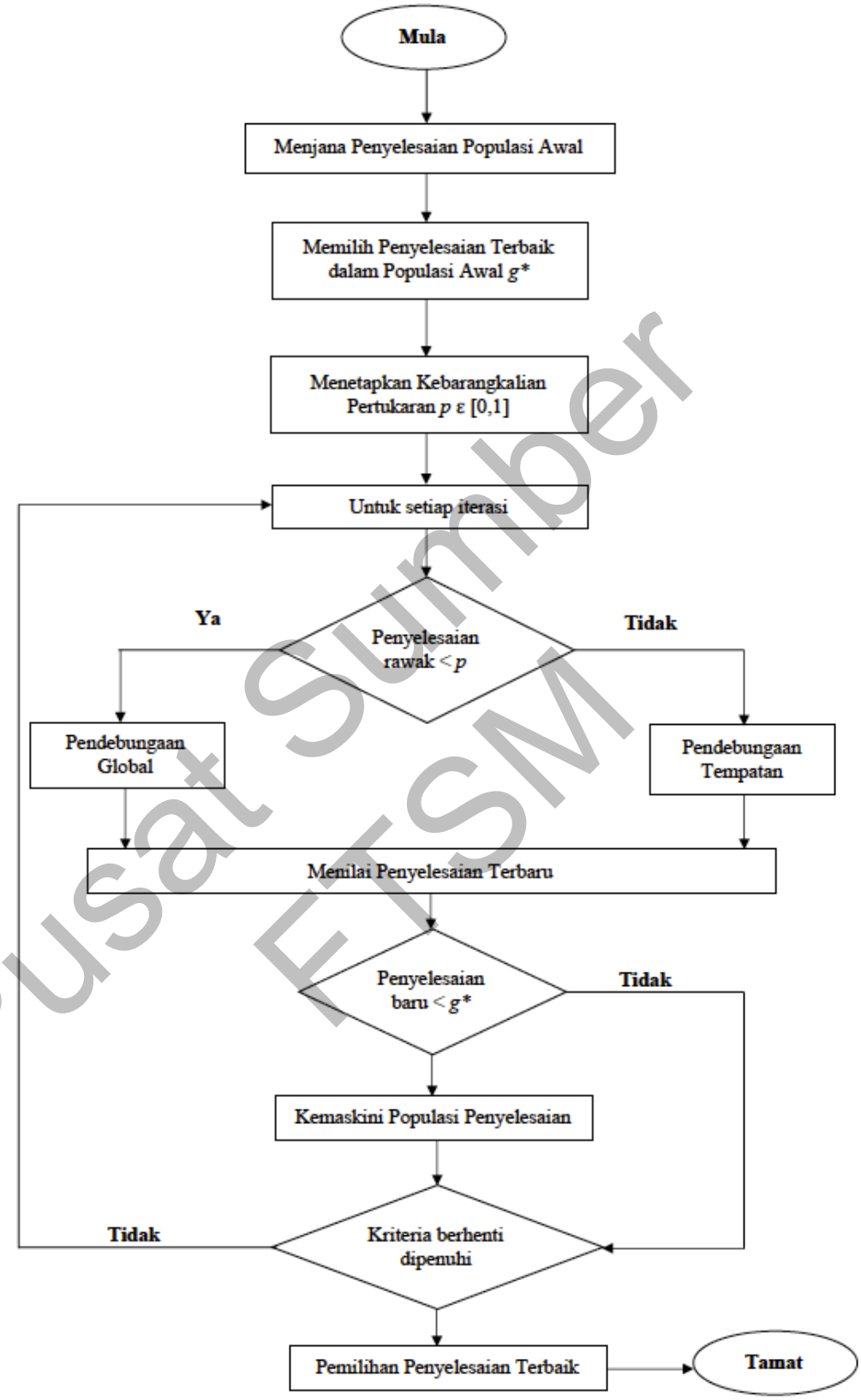
Oleh kerana pendebungaan abiotik berlaku disebabkan oleh angin atau penyebaran tanpa ejen pendebunga, pendebungaan tempatan (**Peraturan 2**) dan kemalaran bunga (**Peraturan 3**) dinyatakan seperti berikut:

$$x_i^{t+1} = x_i^t + \varepsilon\,(x_j^t - x_k^t) \qquad \ldots(2.3)$$

di mana x^t_j dan x^k_j adalah debunga dari bunga yang berbeza dari jenis tumbuhan yang sama. Persamaan ini secara dasarnya meniru kemalaran bunga di kawasan yang terhad. Secara matematik, jika x^t_j dan x^k_j adalah dari spesies yang sama yang boleh dipilih dari populasi yang sama, jadi persamaan akan menjadi jalan rawak tempatan jika ε ditarik dari sebaran seragam [0, 1], dan vektor penyelesaian baru yang dihasilkan tidak akan menjadi terlalu jauh dari penyelesaian yang ada.

2.9.3 Keberangkalian Pertukaran bagi APB

Walaupun pendebungaan biotik dan abiotik telah disimulasikan, peratusan dan frekuensi setiap jenis pendebungaan masih belum dipertimbangkan. Bagi meniru ciri ini, kebarangkalian pertukaran (**Peraturan 4**) digunakan, di mana nilai p menentukan sama ada pengubahsuaian penyelesaian mengikuti pendebungaan tempatan atau global. Walaupun nilai naif $p = 0.5$ digunakan, nilai $p = 0.8$ yang lebih realistik dan berkesan dalam memberikan prestasi yang lebih baik (daripada $p = 0.5$) untuk kebanyakan aplikasi (X. S. Yang 2012). Rajah 2.4 menunjukkan carta alir bagi APB.



Rajah 2.4 Carta alir bagi algoritma pendebungaan bunga

2.9.4 Penyelidikan ke atas Algoritma APB

APB berjaya disesuaikan untuk beberapa domain masalah pengoptimuman (Alyasseri et al. 2018). Bagi domain sistem elektrik dan kuasa, penyelidikan oleh (Sarijiya & Saputra 2016) telah mencadangkan APB yang diubah (MFPA), di mana menggunakan kebarangkalian beralih dinamik, penerapan GA berkod sebenar (RCGA) sebagai mutasi untuk carian global dan tempatan, dan pembezaan antara pencarian tempatan sementara dan penyelesaian optimum. MFPA kemudian dinilai untuk 10 penanda aras sistem kuasa, dan hasil eksperimen mereka menunjukkan kos bahan bakar yang lebih rendah daripada yang ditemukan oleh APB asas. Dalam kajian lain, (Sakti et al. 2017) mencadangkan MFPA untuk mennilai pembiayaan bahan bakar dan masa yang diperlukan untuk mendapatkan penyelesaian optimum global, dan MFPA yang diuji di bawah sistem ujian IEEE 30-bas seterusnya menunjukkan hasil yang unggul berbanding APB asas dan algoritma pengoptimuman metaheuristik yang lain.

Seterusnya, penyelidikan yang melibatkan APB dilakukan pada domain pemprosesan isyarat dan imej. Algoritma pendebungaan bunga binari (APBB) telah diaplikasikan dalam mengatasi masalah pengurangan bilangan sensor yang diperlukan untuk pengenalan orang berdasarkan isyarat EEG (Douglas Rodrigues et al. 2016). APBB digunakan untuk memilih subset saluran yang optimum yang memberikan ketepatan tertinggi. Hasil eksperimen APBB menunjukkan kadar pengiktirafan hingga 87% berdasarkan pengelasan Hutan Jalur Optimum. Selain itu, (Dahi et al. 2016) telah menjalankan kajian secara sistematik dalam menilai prestasi APBB dalam penyelesaian domain Masalah Penentuan Antena (APP). APBB diuji menggunakan data yang realistik, sintetik, dan rawak dengan dimensi yang berbeza dan dibandingkan dengan Pembelajaran Berasaskan Populasi (PBIL) dan algoritma Pembezaan Evolusi (DE), yang merupakan dua algoritma yang cekap dalam domain APP. APBB mencapai penemuan teknikal yang lebih kompetitif daripada PBIL dan DE dalam domain APP.

APB juga tidak ketinggalan diaplikasikan kepada domain pengelompokan dan pengelasan. Prestasi APB yang diubah diuji melalui aplikasi pengelompokan. Algoritma ini dinilai dengan beberapa algoritma pengoptimuman yang berbeza

antaranya algoritma kelawar, algoritma kunang-kunang, dan APB konvensional pada 10 kumpulan data kelompok. Dari 10 set data, 8 dihasilkan dari pengecaman corak dan 2 dihasilkan secara buatan. Hasil pengelompokan dikira dari segi nilai fungsi objektif dan waktu yang diambil oleh CPU pada setiap larian. Graf taburan panjang menggambarkan tingkah laku penumpuan algoritma. Hasil menunjukkan bahawa cadangan APB yang diubahsuai melebihi algoritma setandingnya dari segi pencapaian nilai kecergasan terbaik dan mengurangkan pemprosesan masa CPU (P. Agarwal & Mehta 2016).

Algoritma pendebungaan bunga binari (APBB) diaplikasi kepada masalah pemilihan fitur dan menguji APBB pada enam set data, dan memberikan hasil yang lebih baik daripada *Particle Swarm Optimization* (PSO), *search harmoni* (HS), dan *firefly algorithm* (FA) (Douglas et al. 2015). (Sayed et al. 2016) memperkenalkan algoritma hibrid yang disebut BCFA, yang menggabungkan Algoritma Pemilihan Klonal (CSA) dengan APB untuk menyelesaikan masalah pemilihan fitur. Penggunaan pengelasa Hutan Jalur Optimum sebagai fungsi objektif, dan algoritma hibrid yang dicadangkan (BCFA) telah menghasilkan prestasi yang lebih baik daripada algoritma metahuristik yang lain. Model baru untuk pemilihan ciri pelbagai objektif berdasarkan kombinasi APB dan teori set kasar untuk mencari set ciri optimum untuk klasifikasi (Zawbaa et al. 2016). Model ini mengeksploitasi ciri-ciri pemilihan fitur penapis dan pembalut. Kaedah penapis berfungsi sebagai teknik berorientasi data, manakala kaedah pembalut adalah untuk teknik pengelasan. Prestasi kaedah yang dicadangkan ini disahkan dengan menggunakan lapan set data UCI dan mendapati bahawa kaedah yang ini sangat kompetitif jika dibandingkan dengan FPA asas, PSO, dan algoritma genetik. Penggabungan APB dengan algoritma Ada-Boost bagi tujuan meningkatkan ketepatan dalam pengelasan dokumen teks, di mana fasa pertama digunakan untuk pemilihan fitur, sementara yang terakhir digunakan untuk mengelaskan dokumen teks. Prestasi algoritma yang dicadangkan dinilai dengan menggunakan tiga set data standard seperti Reuters-21578, WEBKB, dan CADE 12. Hasil eksperimen menunjukkan bahawa algoritma yang dicadangkan menunjukkan prestasi yang lebih baik daripada algoritma Ada-Boost dan algoritma lain (Majidpour & Soleimanian 2018).

Hasil kajian-kajian lepas, algoritma APB dipilih untuk diaplikasikan dalam kajian ini kerana ia mampu beroperasi secara optimum tanpa memerlukan sebarang pelarasan parameter spesifik berbanding algoritma lain. Kelebihan lain algoritma ini ialah ia telah diterapkan dalam banyak masalah termasuk beberapa masalah pengoptimuman kombinasi dengan keputusan yang lebih baik serta mempunyai kadar penumpuan yang cepat. Algoritma ini juga mudah di implementasikan berbanding dengan algoritma yang lain dan juga ia hanya memerlukan sumber pemprosesan yang minimum untuk melaksanakan pengiraan.

2.10 TEKNIK PENGELASAN SENTIMEN

Di dalam kecerdasan buatan, pembelajaran mesin merupakan bidang penyelidikan utama. Ini termasuklah dengan reka bentuk, analisis, pelaksanaan dan aplikasi sesuatu program yang mampu belajar di dalam lingkungan (Alpaydin 2011; Carbonell et al. 2013). Sistem pembelajaran mesin ini dapat meningkatkan prestasi dalam melaksanakan tugas tertentu kerana memperoleh pengalaman yang lebih banyak.

Algoritma pembelajaran mesin ini umumnya menggunakan mekanisme maklum balas untuk saling bertukar tingkah laku mereka (belajar). Bergantung kepada jenis maklum balas, algoritma pembelajaran mesin boleh di kategorikan kepada tiga kategori utama; pembelajaran mesin berselia, pembelajaran mesin tidak berselia dan pembelajaran mesin separa selia (Biagetti et al. 2018). Bagi pembelajaran mesin berselia, mesin belajar dengan mengenal objek atau perkara yang berlabel. Jadi hasil yang diinginkan bagi sesuatu masalah akan diketahui terlebih dahulu. Tujuannya adalah untuk mempelajari fungsi pemetaan input ke output yang diingini. Pengelasan adalah teknik daripada pembelajaran mesin berselia. Sebagai contoh, diberi satu set perkara yang diwakili oleh atribut atau fitur dan dipadankan dengan label kelas yang sesuai, proses pengelasan melibatkan pembelajaran model untuk meramal setiap label kelas bagi sesuatu objek dengan betul (Breiman et al. 2017).

Pengelasan merupakan salah satu proses utama di dalam pembelajaran mesin yang merujuk kepada proses menentukan data input yang diberikan kepada salah satu kategori atau kelas yang telah ditentukan (Breiman et al. 2017). Pengelas akan belajar dengan algoritma pembelajaran atau algoritma pengelasan yang juga dikenali sebagai

pemula pengelasan di mana merupakan algoritma pembelajaran mesin berselia. Satu set contoh digunakan oleh algoritma pembelajaran mesin untuk belajar mengelaskan label kelas sesuatu perkara yang belum dilihat atau yang belum dipelajari. Pengelas yang telah belajar akan mengambil nilai fitur atau atribut sesuatu objek sebagai input dan label kelas yang telah ditentukan sebagai output. Sesuatu set label kelas ditakrifkan sebagai sebahagian daripada masalah oleh pengguna.

2.10.1 Naïve Bayes

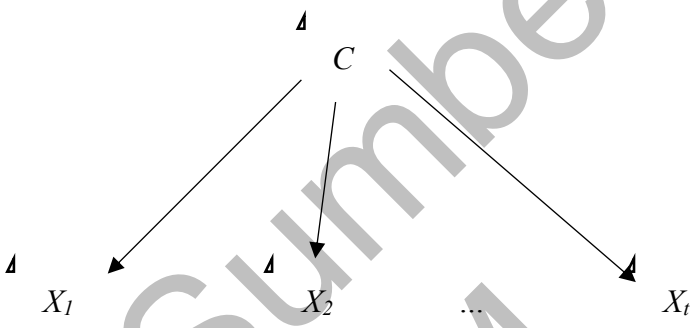
Naïve Bayes (NB) merupakan algoritma pembelajaran mesin popular yang berasaskan statistik. Algoritma NB ini merupakan algoritma yang mudah akan tetapi mempunyai keupayaan yang mengagumkan untuk permodelan ramalan. Algoritma ini merupakan kumpulan teknik pengelasan berdasarkan teori kebarangkalian Bayesian.

Pada dasarnya algoritma pengelasan Naive Bayes ini menghasilkan kebarangkalian untuk setiap kes. Kemudian ia meramalkan hasil kebarangkalian tertinggi. Andaian naif di buat kerana fitur-fitur adalah tidak bersandar, maka pengelasan setiap pasangan fitur saling bergantung antara satu sama lain. Pengelas ini mampu menangani sebarang sebilangan pemboleh ubah berterusan dan kategori dengan cekap (Han et al. 2012). Dipertimbangkan satu set pemboleh ubah, x = {x_1, x_2, x_3,..., x_t}; diperlukan untuk mengetahui kebarangkalian posterior untuk peristiwa C_j dari ruang ruang sampel C = {c_1, c_2, c_3,..., c_t} seperti ditunjukkan pada Rajah 2.5. Secara ringkasnya, peramal adalah X dan C adalah kumpulan tahap kategori yang terdapat dalam pemboleh ubah bersandar. Penggunaan peraturan Bayes seperti berikut:

$$P\,(C_j \mid x_1, x_2, x_3, \ldots, x_t) \cdot P\,(x_1, x_2, x_3, \ldots, x_t \mid C_j)\,P\,(C_j)$$

di mana $P\ (C_j \mid x_1, x_2, x_3, \ldots, x_t)$ adalah kebarangkalian posterior iaitu kebarangkalian peristiwa X miliki oleh C_j ditunjukkan. Pada Naive Bayes, terdapat anggapan bahawa kebarangkalian bersyarat dari pemboleh ubah tidak bersandar mempunyai kebebasan statistik. Dengan menggunakan peraturan Bayes, kes baru X dilabel dengan tahap kelas C_j yang mencapai kebarangkalian posterior tertinggi. Walaupun andaian naif ini menyatakan bahawa pemboleh ubah peramal adalah tidak

bersandar antara satu sama lain tidak selalu tepat. Andaian ini menjadikan proses pengelasan lebih mudah, kerana ia membolehkan ketumpatan bersyarat kelas P (x_d | C_j) dikira untuk setiap pemboleh ubah secara berasingan dan dengan itu tugas multidimensi dikurangkan kepada beberapa tugas satu dimensi. Lebih tepat lagi, ia menukar tugas anggaran kepadatan dimensi tinggi kepada anggaran kepadatan kernel satu dimensi. Tugas pengelasan kekal tidak terpengaruh kerana andaian ini tidak banyak mempengaruhi kebarangkalian posterior, terutama di kawasan yang terletak berdekatan dengan batas keputusan.



Rajah 2.5 Pengelas Naïve Bayes

Sumber: (Kaviani & Dhotre 2017)

Algoritma Naive Bayes tidak memerlukan set data yang besar, malah untuk penetapan anggaran parameter, set data latihan yang bersaiz kecil sudah mencukupi. Ia juga melakukan pengiraan kebarangkalian eksplisit untuk hipotesis. Prospek berguna untuk memahami pelbagai algoritma pembelajaran juga disediakan dengan kaedah ini. Sebagai perbandingan dengan kaedah halus lain, pengelasan ini juga sangat pantas. Ia dapat menyelesaikan masalah diagnostik dengan cekap (Sen et al. 2020).

Telah terdapat beberapa penyelidik terdahulu menggunakan algoritma Naive Bayes dalam kajian mereka. Algortima ini telah diaplikasikan kepada domain analsis sentimen. Kajian (Srikanth et al. 2021) memfokuskan pada aliran kerja peringkat tinggi hingga akhir untuk menyelesaikan masalah pengelasan teks menggunakan algoritma pembelajaran mesin Naive Bayes bagi masalah pengelasan teks yang bertujuan melombong pendapat dan ulasan pengguna dari laman Amazon.

Perbandingan hasil nilai ketepatan yang diperoleh menggunakan kaedah Naïve Bayes dengan kaedah lain untuk melihat keberkesanan kaedah yang digunakan (Ying et al. 2021). Menurut kajian ini, nilai ketepatan yang dihasilkan belum mencapai maksimum dan dengan demikian masih dapat disusun semula dan dinilai semula menjadi model yang lebih baik. Dalam hal ini, kajian tersebut berusaha meningkatkan nilai ketepatan yang dihasilkan agar mesin ini dapat meramalkan berita yang mengandungi sarkasme melalui pengubahsuaian nilai ambang yang ditetapkan. Apabila nilai ambang diubah menjadi 0.3, nilai ketepatan meningkat kepada 77% dan nilai kesalahan dalam ramalan tajuk positif palsu juga meningkat dengan ketara dan hanya menghasilkan 89 kesalahan dalam meramalkan sarkasme.

Kajian seterusnya dalah pembangunan aplikasi pengangkutan bernama Trafi untuk mendapatkan pendapat atau komen mengenai aplikasi dari orang yang telah menggunakan aplikasi tersebut. Satu set data positif dan negatif dari komen atau pendapat menerusi aplikasi tersebut yang akan. Untuk pengelasan, penyelidikan menggunakan Naïve Bayes (NB), di mana NB adalah salah satu algoritma yang paling popular untuk pengecaman corak. Selain itu, pengelas NB adalah teknik pembelajaran mesin yang popular untuk pengelasan teks, *Particle Swarm Optimization* (PSO) digunakan sebagai medium pemilihan fitur digabungkan dengan NB untuk meningkatkan prestasi. Sebelum penggunaan PSO ketepatan set data yang diperoleh adalah 69.50% dan dan setelah gabungan PSO dan NB nilai ketepatan adalah 72.34% (Yulia & Solecha 2021). Sehubungan dengan itu, NB dipilih sebagai algoritma pengelasan dalam kajian ini di samping algoritma ini juga digunakan oleh model asal perbandingan.

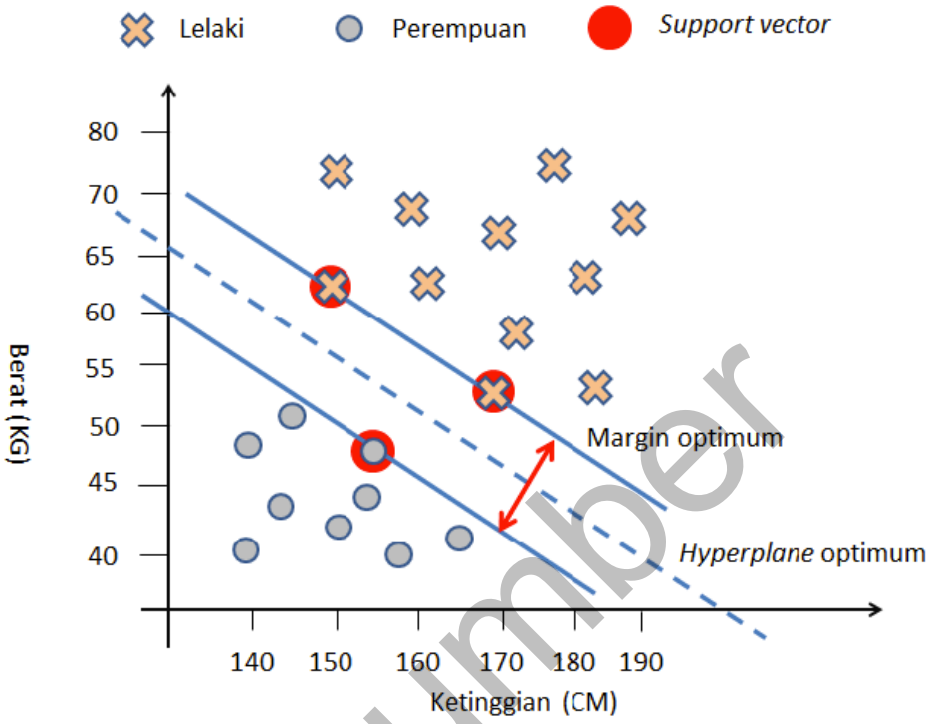
2.10.2 Mesin Vektor Sokongan (SVM)

SVM juga dikenali sebagai rangkaian vektor sokongan telah dipelopori oleh (Cortes & Vapnik 1995). Pada peringkat awal, SVM dibangunkan untuk pengelasan binari iaitu pengelasan dua dimensi sahaja (Duan & Keerthi 2005). Sejajar dengan perkembangan teknik pengelasan, prestasi dan strategi pengelasan, SVM ini telah menjadi terkenal dalam bidang pengelasan sentimen. SVM mampu mencipta garisan sempadan atau hyperplane yang memisahkan data mengikut kelas. *Support vector* merupakan data

yang terhampir dengan hyperplane. Manakala margin pula merupakan jarak di antara permukaan vektor yang paling terhampir dengan hyperplane.

Terdapat banyak kajian yang telah mencadangkan sistem SVM yang kompleks untuk pengelasan bukan hanya untuk dua jenis kelas sahaja, tetapi untuk pelbagai kelas sentimen. Bagaimanapun, terdapat juga kajian yang mencadangkan pelbagai pendekatan untuk mengaplikasikan pengelasan pelbagai kelas dengan menggunakan teknik SVM yang ringkas. Kebanyakan pendekatan yang dicadangkan adalah dengan menggabungkan beberapa pengelasan SVM untuk mengelaskan masalah pelbagai kelas. Antara pendekatan yang popular adalah "satu lawan satu", "satu lawan semua" dan kod output pembetulan-ralat (Li et al. 2005; Platt et al. 1999). Kajian empirikal ke atas sistem pengelasan pelbagai kelas SVM yang dijalankan oleh Duan & Keerthi (2005) mencadangkan kombinasi SVM "satu lawan satu" mengatasi kaedah pengelasan yang lain. Tambahan pula, kajian perbandingan yang dijalankan oleh Hsu & Lin (2005) juga memberikan sokongan yang kuat bahawa kombinasi "satu lawan satu" adalah lebih baik dari pendekatan kombinasi yang lain, terutama apabila berhadapan dengan masalah set data latihan yang tidak mencukupi. SVM juga mempunyai fungsi yang dipanggil kernels yang menggunakan ruang kelas berdimensi rendah dan menukarkannya kepada kelas berdimensi tinggi.

Rajah 2.6 menunjukkan contoh pengelasan SVM dua dimensi. Contoh pengelasan SVM ini adalah dengan menggunakan set data yang memaparkan maklumat ketinggian dan berat mengikut jantina. Terdapat dua jenis kelas yang digunakan iaitu kelas lelaki dan kelas perempuan. Berdasarkan contoh ini kita dapat mengelaskan data kepada kelas masing-masing berdasarkan maklumat berat dan ketinggian yang dibekalkan. Contohnya, jika berat seorang individu 70 kg dan mempunyai ketinggian 175 cm, maka jantina individu tersebut tergolong dalam kelas lelaki. Sebaliknya, jika berat seorang individu 50 kg dan mempunyai ketinggian tersebut 155 cm, maka individu tersebut tergolong dalam kelas perempuan.



Rajah 2.6 Pengelasan SVM bagi kelas dua dimensi

Sumber: (Vohra & Teraiya 2013)

Pengelas SVM terselia memerlukan set data latihan untuk mengira dan menghasilkan hyperplane optima yang memisahkan dan mengelaskan data input kepada dua kategori. Apabila SVM ini digunakan untuk pelbagai kelas yang melebihi dua dimensi kelas, ruang dua dimensi ini digantikan dengan ruang pelbagai dimensi yang beraras tinggi, iaitu dengan melaksanakan penggabungan beberapa pengelasan SVM dua dimensi. Pelaksanaan penggabungan ini dilakukan melalui fungsi kernels. Dengan fungsi ini, ia mampu menukarkan masalah kelas yang tidak boleh dipisahkan kepada boleh dipisahkan. Ia amat penting terutama untuk pemisahan kelas bukan linear. Mengikut kajian perbandingan yang dijalankan oleh (Vohra & Teraiya 2013), SVM menghasilkan kadar ketepatan pengelasan sentimen yang tinggi berbanding dengan pengelasan yang lain. Sehubungan dengan itu, SVM dipilih sebagai algoritma pengelasan dalam kajian ini di samping algoritma ini juga digunakan oleh model asal perbandingan.

2.10.3 Pokok Keputusan (DT)

Algoritma berasaskan logik menangani masalah dengan aliran data langkah demi langkah dengan berfungsi logik dalam setiap langkah. Di sini pokok keputusan telah dilihat sebagai contoh klasik algoritma berasaskan logik. (Murthy 1998) telah memberikan gambaran keseluruhan kerja di DT dan contoh kebergunaannya kepada pendatang baru dan juga pengamal dalam bidang pembelajaran mesin. DT adalah pokok yang mengklasifikasikan keadaan dengan menyusunnya berdasarkan nilai fitur. Setiap nod di DT mewakili fitur dalam objek yang akan diklasifikasikan, dan setiap cabang mewakili nilai yang dapat diandaikan oleh nod. Data akan diklasifikasikan bermula pada nod akar dan disusun berdasarkan nilai fitur seperti di Rajah 2.7.

Rajah 2.7 Keputusan pokok membuat sesuatu keputusan

Sumber: (Sen et al. 2020)

Seperti SVM, DT juga berfungsi dengan baik untuk kedua-dua pengkategorian dan kebergantungan objek. Kaedah ini sebenarnya membina model berasaskan aliran keputusan berurutan pada nilai sebenar fitur dalam set data. Keputusan dibahagikan kepada struktur seperti pokok. Keputusan di buat di setiap nod pokok melainkan ramalan di buat untuk data input tertentu. DT dilatih untuk kegunaan data bagi masalah pengelasan. Algoritma ini berfungsi sangat pantas dengan menghasilkan ketepatan yang memuaskan. Ini adalah teknik kegemaran dan banyak digunakan dalam pembelajaran mesin dan juga berfungsi dengan cekap pada jumlah set data

jumlahnya lebih sedikit. Algoritma ini membahagikan kumpulan item data menjadi dua atau lebih set homogen berdasarkan atribut yang paling signifikan untuk kumpulan yang berbeza. DT merupakan kaedah yang mudah, senang untuk difahami dan divisualisasikan. Selain itu, DT juga tidak memerlukan masa yang banyak untuk diproses dan boleh digunakan untuk data jenis berangka dan kategori.

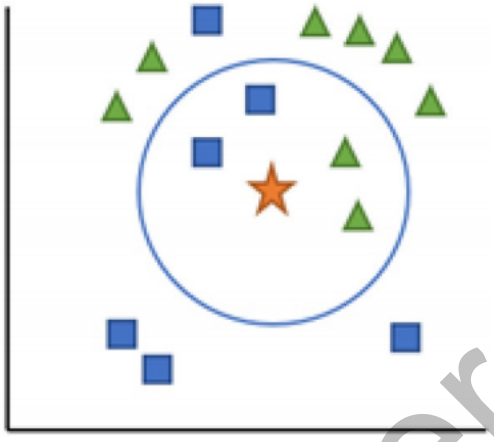
Antara kerja-kerja penyelidikan yang menggunakan algortima pengelas DT adalah algoritma pengelas keputusan pokok dalam menganalisis skor guru pendidikan jasmani di kolej. Tujuan penyelidikan ini adalah untuk mengkaji penerapan algoritma klasifikasi berdasarkan keputusan pokok dalam analisis prestasi guru pendidikan jasmani universiti. Makalah ini memperkenalkan algoritma perlombongan data keputusan pokok untuk analisis dan penilaian prestasi guru. Berdasarkan algoritma keputusan pokok, model keputusan pokok bagi menganalisis dan penilaian prestasi guru pendidikan jasmani telah bangunkan, dan peraturan yang sesuai diekstrak. Tujuannya adalah untuk mendapatkan indeks yang diperlukan untuk mempengaruhi prestasi guru pendidikan jasmani melalui algoritma keputusan pokok. Hasil eksperimen menunjukkan bahawa penyelidikan ini dapat meletakkan asas dan memberikan rujukan untuk peningkatan kesan pengajaran. Berdasarkan penilaian guru pendidikan jasmani di sekolah, penyelidikan ini mendapati bahawa penilaian guru sekolah pada dasarnya tertumpu pada tahap "sederhana", menyumbang kepada nilai 70% (Sun 2021).

Seterusnya, penyelidikan terhadap teknologi analisis sentimen juga telah dilakukan oleh (Adnan et al. 2019). Hal ini di mana dalam persaingan perniagaan restoran yang semakin sengit, banyak restoran bersaing untuk memberikan kualiti terbaik kepada pengguna. Kualiti restoran merangkumi makanan dan minuman, persekitaran, tempat dan perkhidmatan. Keadaan itu mempengaruhi imej jenama restoran yang ditandakan sama ada pengguna berpuas hati atau tidak. Salah satu laman web restoran ulasan yang sering disebut adalah TripAdvisor yang dipilih sebagai sumber data kerana mengandungi ciri Kandungan yang Dihasilkan Pengguna (UGC) Kelebihan menggunakan UGC adalah memastikan keaslian data komen pengguna. Pengelasan teks bahasa Inggeris digunakan dalam kajian ini untuk menentukan ketidakpuasan (negatif) dan kepuasan (positif) pengguna berdasarkan komen atau

ulasan mereka. Penyelidikan ini menggunakan keputusan pokok sebagai kaedah pengelasan. Pengumpulan data dilakukan dengan mengekstrak data menggunakan WebHarvy. Prestasi keseluruhan kaedah keputusan pokok telah mengatasi kaedah pengelasan yang lain. Hasil daripada pengelasan ini adalah sebagai cadangan untuk pengguna memilih restoran yang terbaik. Sehubungan dengan itu, DT dipilih sebagai algoritma pengelasan dalam kajian ini di samping algoritma ini juga digunaka oleh model asal perbandingan.

2.10.4 k-Jiran Terdekat (kNN)

Pengelas k-NN adalah salah satu algoritma yang paling mudah dan paling banyak digunakan dalam kategori algoritma pengelasan. k-NN dicadangkan pada tahun 1951 oleh (Fix & Hodges 1989). Teknik ini boleh digunakan untuk pengelasan dan regresi (Mohammed et al. 2017). Konsep utama bagi k-NN adalah ia bergantung pada pengiraan jarak antara sampel yang diuji, dan data latihan untuk mengenal pasti jiran terdekatnya. Sampel yang diuji kemudiannya hanya diberikan kepada kelas jiran terdekatnya (Larose 2015). Dalam k-NN, nilai k mewakili bilangan jiran terdekat. Nilai ini adalah faktor penentu inti bagi pengelasan ini kerana nilai k menentukan berapa jiran yang mempengaruhi pengelasan tersebut. Apabila k = 4 maka objek data baru hanya diberikan kepada kelas jiran terdekatnya seperti pada Rajah 2.8. Jiran diambil dari satu set objek data latihan untuk mengetahui pengelasan yang betul sudah diketahui. k-NN berfungsi secara semula jadi dengan data berangka. Pelbagai ukuran berangka telah digunakan seperti jarak Euclidean, Manhattan, Minkowsky, City-block, dan Chebyshev. Euclidean merupakan fungsi jarak yang paling banyak digunakan dengan k-NN (Larose & Larose 2014).



Rajah 2.8 k-Jiran Terdekat bagi nilai k=4

Berikut merupakan langkah utama algoritma k-NN; yang pertama adalah menentukan bilangan jiran terdekat (nilai K). Kedua adalah menghitung jarak antara sampel ujian dan semua sampel latihan. Ketiga adalah mengisih jarak dan tentukan jiran terdekat berdasarkan jarak minimum K. Keempat pula adalah menghimpunkan kategori jiran terdekat. Akhir sekali, menggunakan sebahagian besar kategori jiran terdekat sebagai nilai ramalan objek data baru. Pengelasan k-NN dapat digunakan untuk mengklasifikasikan objek data baru hanya menggunakan jaraknya ke sampel yang dilabel (Weinshall et al. 1998). Walau bagaimanapun, beberapa penyelidikan mempertimbangkan sebarang ukuran metrik atau bukan metrik yang digunakan dengan pengelasan ini: hal ini di mana beberapa kajian telah dilakukan untuk menilai pengelasan k-NN menggunakan ukuran metrik dan bukan metrik yang berbeza seperti kajian yang dikemukakan dalam (Hu et al. 2016; Lopes et al. 2015; Todeschini et al. 2016). Sehubungan dengan itu, k-NN dipilih sebagai algoritma pengelasan dalam kajian ini di samping algoritma ini juga digunaka oleh model asal perbandingan.

2.11 ALAT PENGUKURAN

Terdapat dua jenis alat pengukuran yang telah digunakan dalam eksperimen pemilihan fitur analisis sentimen bagi kajian ini. Kedua-dua teknik tersebut adalah *Term Frequency-Inverse Document Frequency* (A. Kumar et al. 2018; A. Kumar & Jaiswal 2019) dan *Binary Cuckoo Search* (A. Kumar et al. 2018).

2.11.1 *Term Frequency-Inverse Document Frequency* (TF-IDF)

TF-IDF merujuk kepada frekuensi terma-songsangan frekuensi dokumen yang digunakan untuk mengukur kepentingan terma dalam sesuatu dokumen. Ia terdiri daripada gabungan dua ayat yang berlainan iaitu frekuensi terma dan songsangan frekuensi dokumen. Frekuensi terma merujuk kepada berapa kali sesuatu terma berlaku dalam dokumen yang dipilih (Qaiser & Ali 2018). Seperti diketahui bahawa kepanjangan dokumen adalah berbeza di antara satu sama lain, maka kemungkinan satu terma berlaku lebih kerap di dalam dokumen yang besar berbanding dengan dokumen yang lebih kecil.

Untuk mengatasi masalah ini, jumlah terma yang ditemui dibahagi dengan jumlah keseluruhan terma yang wujud di dalam dokumen tersebut untuk mendapatkan nilai frekuensi. Sebagai contoh, terdapat sebuah dokumen *T1* yang mengandungi 10,000 perkataan dan perkataan "*landing*" berlaku sebanyak 90 kali. Berdasarkan contoh ini, frekuensi perkataan "*landing*" dalam dokumen *T1* adalah 90 maka TF akan dikira seperti berikut:

TF= 90/10000 = 0.009

Apabila pengiraan terma frekuensi sesuatu dokumen dilakukan, semua perkataan akan dilayan secara sama rata walaupun perkataan tersebut tidak mempunyai sebarang kepentingan seperti kata henti "*the*". Dalam sesuatu dokumen, setiap perkataan mempunyai kepentingan sendiri. Andaikan perkataan "*the*" wujud sebanyak 2,000 kali dalam dokumen ini tetapi ia tidak mempunyai kepentingan dan tidak mempunyai impak ke atas pengelasan, ini akan menyebabkan kadar ketepatan pengelasan berkurangan. Oleh itu frekuensi songsangan dokumen digunakan untuk memberi pemberat rendah kepada perkataan yang berfrekuensi tinggi dan pemberat yang tinggi kepada perkataan yang mempunyai frekuensi rendah. Sebagai contoh, jika terdapat 10 buah dokumen dan terma "*crew*" berlaku ditemui di dalam 5 buah dokumen, maka frekuensi songsangan dokumen dikira seperti berikut:

IDF = log_e (10/5) = 0.310

Daripada pengiraan di atas, di dapati bahawa perkataan yang paling banyak ditemui dalam dokumen akan mempunyai frekuensi paling tinggi, sebaliknya perkataan yang paling kurang ditemui pula akan mempunyai kepentingan yang tinggi. Maka TF-IDF merupakan hasil darab di antara frekuensi terma (TF) dan juga frekuensi songsangan dokumen (IDF).

$$TF - IDF = TF * IDF \quad \ldots(2.4)$$

2.11.2 *Binary Cuckoo Search* (BCS)

Pada tahun-tahun kebelakangan ini, algoritma meta-heuristik berdasarkan tingkah laku biologi dan sistem fizikal dalam alam semula jadi telah dicadangkan untuk menyelesaikan masalah pengoptimuman. Antara cadangan yang pertama bagi versi binari Pengoptimuman Particle Swarm (PSO) yang terkenal dipanggil BPSO, di mana algoritma PSO asal telah diubah suai untuk menangani masalah pengoptimuman binari (Kennedy & Eberhart 1997). Seterusnya, BPSO ini telah diaplikasikan kepada masalah pemilihan fitur (C. S. Yang et al. 2008).

Beberapa tahun kemudian, (Rashedi et al. 2010) mencadangkan versi binari Algoritma Carian Gravitasional (GSA) yang dipanggil BGSA untuk pemilihan ciri, dan (Ramos et al. 2011) pula mempersembahkan versi Carian Harmony (HS) untuk tujuan yang sama dalam konteks pengesanan kecurian dalam sistem pengagihan kuasa. Kemudian, (Nakamura et al. 2012) memperkenalkan versi Algoritma Kelawar (BA) untuk masalah pengoptimuman binari, yang dipanggil BBA.

Baru-baru ini, (S. Yang et al. 2014) mencadangkan kaedah metaheuristik baharu untuk pengoptimuman berterusan iaitu Cuckoo Search (CS), yang berdasarkan strategi pembiakan burung cuckoo yang menarik. Beberapa spesies terlibat dengan parasitisme brood yang bertelur di sarang burung tuan rumah yang lain. Pendekatan sedemikian telah menunjukkan untuk mengatasi beberapa teknik pengoptimuman yang diilhamkan oleh alam semula jadi yang terkenal, seperti PSO dan Algoritma Genetik.

Tingkah laku parasit sesetengah spesies Cuckoo sangat menarik. Burung ini boleh meletakkan telurnya di dalam sarang perumah, dan meniru ciri luaran telur perumah seperti warna dan bintik-bintik. Sekiranya strategi ini tidak berjaya, hos boleh membuang telur cuckoo, atau hanya meninggalkan sarangnya, membuat sarang baharu di tempat lain. Berdasarkan konteks ini, pembangunan algoritma pengoptimuman evolusi baru yang dinamakan sebagai Carian Cuckoo (CS), dan ia telah meringkaskan CS menggunakan tiga peraturan (S. Yang et al. 2014), seperti berikut:

i. Setiap cuckoo memilih sarang secara rawak untuk bertelur.

ii. Bilangan sarang perumah yang tersedia adalah tetap, dan sarang dengan telur berkualiti tinggi akan dibawa ke generasi seterusnya.

iii. Sekiranya burung perumah menemui telur cuckoo, ia boleh membuang telur itu atau meninggalkan sarang, dan membina sarang yang baru sepenuhnya.

Dalam CS asal, penyelesaian dikemas kini dalam ruang carian ke arah kedudukan bernilai berterusan. Tidak seperti CS, bagi pemilihan fitur untuk BCS pula, ruang carian dimodelkan sebagai kekisi boolean n-dimensi, di mana penyelesaian dikemas kini merentasi penjuru hiperkubus. Di samping itu, kerana masalahnya adalah untuk memilih atau tidak ciri yang diberikan, vektor binari penyelesaian digunakan, di mana 1 sepadan sama ada ciri akan dipilih untuk mengarang set data baharu dan 0 sebaliknya. Untuk membina vektor binari ini, (Pereira et al. 2014; D. Rodrigues et al. 2013) telah menggunakan persamaan 2.5, yang hanya boleh memberikan nilai binari dalam kekisi boolean yang mengehadkan penyelesaian baharu kepada nilai binari sahaja:

$$x^{j}_{i}(\mathrm{t}+1) = \begin{cases} 1 & \text{jika } S(x^{j}_{i}(t)) > \sigma, \\ 0 & \text{sebaliknya} \end{cases} \qquad \ldots(2.5)$$

di mana $\sigma \sim U(0, 1)$ dan $x^{j}_{i}(t)$ menandakan nilai telur baru pada masa t.

2.12 PERBINCANGAN

Proses menganalisis dan mengenal pasti jenis sentimen daripada ulasan pengguna yang berbentuk teks boleh ditakrifkan sebagai analisis sentimen. Ia digunakan untuk menganalisis ulasan pengguna yang diperoleh dari laman sosial, laman web, forum dan lain-lain lagi. Masalah utama yang dihadapi oleh analisis sentimen ini adalah ulasan pengguna atau set data ini biasanya mengandungi hingar dan data yang tidak relevan. Seterusnya saiz fitur dalam analisis sentimen ini biasanya mempunyai dimensi fitur yang amat besar dan memerlukan masa pemprosesan yang agak panjang dan sumber pengkomputeran yang agar tinggi.

2.12.1 Rumusan Masalah Pemprosesan Teks

Bagi mengatasi masalah set data yang mengandungi hingar dan data tidak relevan ini, proses pemprosesan teks perlu dilaksanakan. Pemprosesan teks merupakan kaedah pembersihan set data dari hingar dan data tidak relevan. Seterusnya pemprosesan teks ini juga membantu dalam penyediaan set data kepada set fitur untuk proses pemilihan fitur. Berdasarkan kajian kesusasteraan, terdapat dua kategori pemprosesan teks dikenal pasti iaitu pemprosesan linguistik dan pemprosesan bahasa tabii. Antara teknik pemprosesan dalam kategori pemprosesan linguistik ini adalah menukar huruf kecil, penghapusan @, pembuangan tanda baca, penghapusan tanda pagar # dan penghapusan simbol. Akan tetapi, bagi teknik pemprosesan bahasa tabii pula akan menggunakan teknik *lemmatize*, *stemming* dan pembetulan ejaan.

2.12.2 Rumusan Masalah Pemilihan Fitur

Kemudian, bagi mengatasi masalah dimensi fitur yang agak besar, kaedah pemilihan fitur dicadangkan. Dalam kajian ini terdapat beberapa kajian lepas membincangkan kajian dan eksperimen mengenai teknik pemilihan fitur dalam analisis sentimen. Pada seksyen-seksyen dalam bab ini telah dibincangkan teknik pemilihan fitur secara khusus bagi kajian-kajian terdahulu. Set data dalam bentuk dokumen teks mewujudkan set vektor fitur yang berdimensi besar. Keadaan ini memerlukan satu pendekatan yang mampu menerokai ruang pencarian secara menyeluruh iaitu pendekatan metaheuristik yang berkeupayaan memperolehi penyelesaian yang

optimum. Pemilihan subset fitur ialah merupakan masalah polinomial bukan deterministik iaitu memerlukan algoritma yang berkesan bagi menyelesaikan masalah pemilihan fitur iaitu algoritma metaheuristik. Sehubungan dengan itu, pemilihan fitur algoritma APB digunakan untuk memilih subset fitur yang berkualiti. Algoritma APB ini merupakan adaptasi daripada proses pendebungan bunga. Pendebungaan adalah proses memindahkan debunga dari satu bunga ke bunga lain oleh angin atau ejen pendebunga seperti serangga, rama-rama, lebah, burung dan kelawar. Tumbuhan berbunga telah berkembang untuk menghasilkan nektar atau madu untuk menarik ejen pendebunga dan memastikan pendebungaan berlaku. Ia terdiri daripada dua jenis pendebungaan iaitu pendebungaan global dan pendebungaan tempatan. Antara kelebihan algoritma APB ini berbanding algoritma pemilihan fitur yang lain ialah ia tidak memerlukan sebarang parameter kawalan spesifik sendiri, sebaliknya hanya memerlukan parameter kawalan umum seperti saiz populasi dan jumlah generasi. Algoritma ini juga agak cepat dan mudah dilaksanakan dan tidak memerlukan sumber pengkomputeran yang tinggi untuk melaksanakan pengiraan. Oleh itu algoritma pendebungaan bunga akan digunakan dalam kajian ini.

Untuk algoritma pengelasan pula, Mesin Vektor Sokongan, Keputusan Pokok, Naive Bayes dan k-Jiran Terdekat dipilih sebagai algoritma pengelas. Keempat-empat algoritma pengelas ini dipilih kerana ia juga digunakan oleh penyelidik terdahulu dan penyelidikan ini dijadikan sebagai model penanda aras (A. Kumar et al. 2018; A. Kumar & Jaiswal 2019).

2.13 KESIMPULAN

Bagi kajian kesusasteraan yang dijalankan dalam bab ini adalah meliputi skop utama kajian iaitu pemprosesan teks dan pemilihan fitur dalam analisis sentimen. Tujuan utama kajian ini adalah untuk mengkaji berkenaan pemilihan fitur oleh algoritma APB dan seterusnya melihat bagaimana ia boleh diaplikasikan sebagai algoritma pemilihan fitur dalam pengelasan sentimen. Algoritma APB ini akan berfungsi sebagai alat untuk menghasilkan subset fitur bersaiz kecil dan berkualiti yang akan digunakan oleh algoritma pengelas bagi tujuan pengelasan ulasan pengguna. Di samping itu, tujuan kajian ini adalah untuk mengkaji kesan pemprosesan teks ke atas ketepatan

pengelasan sentimen. Kajian dilakukan dengan menyenaraikan teknik pemprosesan yang dijalankan oleh kajian-kajian lepas. Seterusnya kajian ini melihat kemungkinan teknik-teknik pemprosesan teks yang boleh digabungkan untuk meningkatkan ketepatan pengelasan. Oleh itu semua isu yang berkaitan dengan pemilihan fitur, algoritma APB dan pemprosesan teks diberi perhatian khusus semasa kajian kesusasteraan ini dilakukan. Bab yang seterusnya akan menerangkan mengenai metodologi kajian yang akan digunakan dalam penyelidikan ini.

BAB III

METODOLOGI KAJIAN

3.1 PENGENALAN

Bab ini menerangkan metodologi kajian yang digunakan untuk menjalankan penyelidikan ini. Penyelidikan eksperimental merupakan asas kepada kajian ini di mana ia berasaskan kepada lima fasa penyelidikan perlombongan pendapat digunakan iaitu bermula dengan kajian kesusasteraan, prapemprosesan data teks, pembangunan teknik pemilihan fitur, pengelasan sentimen dan diakhiri dengan pengujian, penilaian dan analisis. Permulaan bagi bab ini adalah dengan menerangkan kerangka metodologi kajian dan reka bentuk eksperimen yang merupakan tulang belakang bagi kajian ini. Selain itu, bab ini juga menerangkan secara terperinci teknik pemprosesan data teks, pembangunan teknik pemilihan fitur, pengelasan sentimen serta pengujian, penilaian dan analisis. Tidak ketinggalan, maklumat berkaitan dengan data yang digunakan untuk eksperimen ini, kaedah pengujian dan penilaian algoritma turut dibincangkan.

3.2 KERANGKA METODOLOGI KAJIAN

Kajian ini merupakan kajian terhadap teknologi analisis sentimen yang menghasilkan output seperti subset fitur yang optimum dan hasil prestasi pengelasan. Untuk tujuan tersebut, pendekatan kajian yang digunakan di dalam penyelidikan ini ialah berasaskan kepada penyelidikan eksperimental bagi perlombongan teks khususnya di dalam pemilihan fitur dan pengelasan sentimen. Metodologi kajian ini mempunyai lima fasa iaitu kajian kesusasteraan, pemprosesan teks, pembangunan teknik pemilihan fitur, pengelasan sentimen, dan akhir sekali pengujian, penilaian dan

analisis prestasi pembangunan teknik dan model seperti yang diringkaskan di dalam Rajah 3.1.

Fasa kajian kesusasteraan melibatkan kajian ke atas penyelidikan semasa dan yang lepas melalui penilaian penting terhadap buku, artikel, prosiding, laporan penyelidikan dan lain-lain sumber akademik yang lain. Kajian ini merangkumi pemahaman mendalam terhadap isu-isu semasa dalam analisis sentimen, konsep asas, teknik prapemprosesan teks, teknik pemilihan fitur sedia ada, pengelasan sentimen dan faktor-faktor yang mempengaruhi dalam analisis sentimen. Tujuan fasa ini ialah untuk mengkaji isu dan cabaran yang dihadapi di dalam analisis sentimen yang mempunyai potensi untuk dibuat penambahbaikan. Setiap dapatan daripada tinjauan literatur akan diringkaskan ke dalam bentuk yang tersusun agar jurang penyelidikan dapat dikenalpasti dengan mudah serta dapat menentukan sumbangan penyelidikan ini. Maklumat terperinci tentang kajian awalan dipersembahkan di dalam Bab II.

Fasa pemprosesan teks merupakan proses pembangunan model untuk pembersihan, penukaran dan penyediaan set data yang dipilih ke dalam bentuk yang boleh diproses digunakan dengan algoritma pemilihan fitur. Fasa ini melibatkan proses pembetulan ayat dari segi tanda bacaan, kesilapan ejaan, penggunaan huruf besar, penghapusan karakter berulang dan simbol. Input kepada model ini merupakan set data yang telah dikenal pasti dan menghasilkan senarai set fitur sebagai output.

Fasa pembangunan algoritma pemilihan fitur iaitu pembangunan algoritma APB. Output daripada algoritma ini ialah subset fitur optimum. Terdapat eksperimen dijalankan bagi menguji keberkesanan algoritma dalam memilih subset fitur optimum. Bab IV menjelaskan mengenai hasil eksperimen algoritma pemilihan fitur.

Fasa pengelasan sentimen pula berperanan dalam menghasilkan model-model pengelasan berdasarkan empat (4) algoritma pengelas pembelajaran mesin di mana mengelaskan sentimen berdasarkan set fitur yang dipilih dari fasa ketiga sebagai input. Hasil daripada fasa ini merupakan model pengelasan sentimen ulasan pengguna yang akan digunakan dalam fasa seterusnya.

Fasa pengujian, penilaian dan analisis bagi setiap algoritma dan model yang dibina dalam fasa-fasa sebelum ini dijelaskan dalam fasa ini. Fasa ini melibatkan beberapa siri pengujian yang menggunakan set data Twitter. Proses penilaian ke atas teknik-teknik dan model yang dibangunkan diuji dengan teknik dan model perbandingan yang setanding. Di dalam fasa ini juga, beberapa penilaian prestasi dilakukan untuk menguji teknik dan model yang dibangunkan.

FASA	AKTIVITI	OUTPUT
Kajian Kesusasteraan	- Menilai penyelidikan semasa dan penyelidikan lepas - Mengenal pasti masalah dan jurang penyelidikan	- Jurang penyelidikan dikenalpasti - Cadangan penyelidikan dilaporkan
Pemprosesan Teks	- Penyediaan set data untuk eksperimen	- Set data yang bersih
Pembangunan Algoritma Pemilihan Fitur	- Pembangunan algoritma pendebungaan bunga binari	- Set fitur yang optimum
Pengelasan Sentimen	- Mengelaskan sentimen berdasarkan set fitur pada fasa pemilihan fitur	- Prestasi pengelasan sentimen set data
Pengujian, Penilaian dan Analisis	- Menilai algoritma dengan teknik perbandingan setanding dan set data tanda aras - Menggunakan beberapa penilaian prestasi	- Analisis penuh keputusan - Hasil perbandingan

Rajah 3.1 Metodologi Kajian

3.3 REKA BENTUK KAJIAN

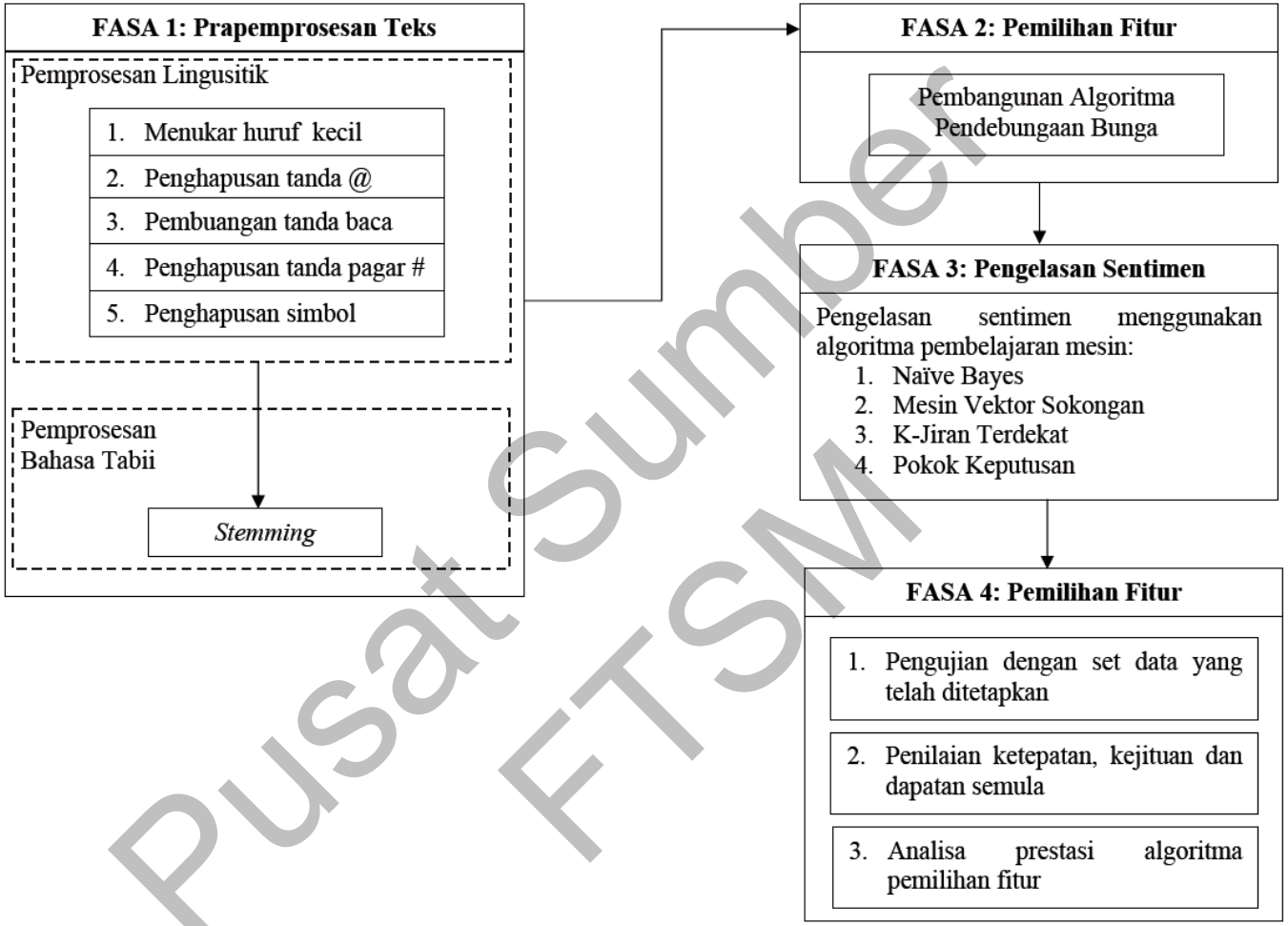
Reka bentuk eksperimen dalam kajian ini mempunyai empat fasa yang penting iaitu fasa pemprosesan teks, fasa pembangunan algoritma pemilihan fitur, fasa pengelasan sentimen dan fasa pengujian, penilaian dan analisis. Rajah 3.2 menunjukkan reka bentuk eksperimen yang diaplikasikan di dalam kajian ini dan penerangan terperinci berkaitan dengan setiap fasa akan dilakukan di dalam seksyen berikutnya. Secara ringkasnya, fasa reka bentuk eksperimen yang terlibat adalah seperti berikut:

A. Fasa 1: Prapemprosesan teks melibatkan pemilihan data yang sesuai untuk tujuan kajian, menjalankan pembersihan data serta mentransformasikan set data. Fasa ini akan menghasilkan senarai set fitur untuk proses pada fasa berikutnya. Bagi Rajah 3.3, menunjukkan reka bentuk eksperimen di mana terdapat perubahan penggunaan teknik pemprosesan bahasi tabii iaitu penambahan dua teknik lain.

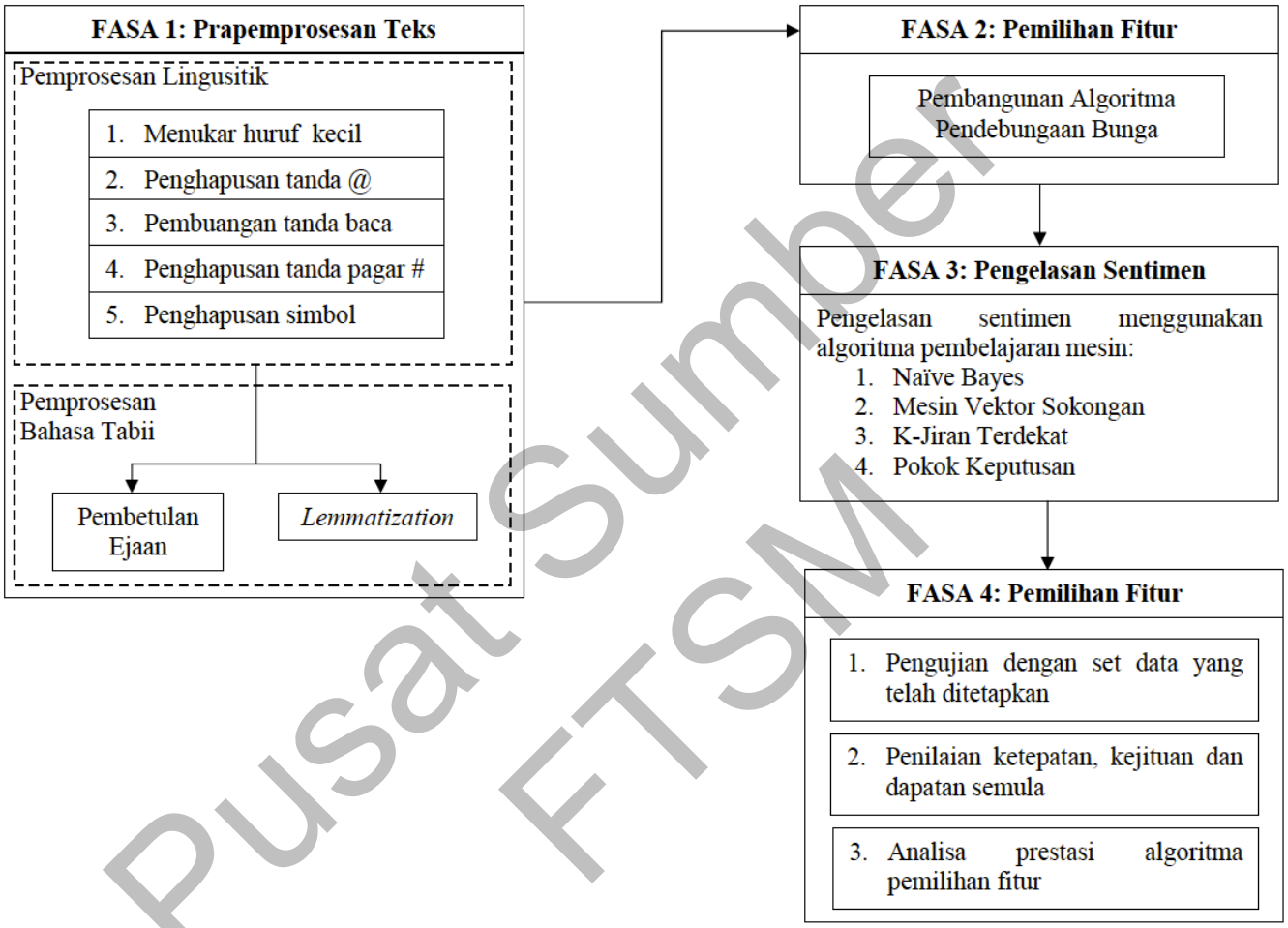
B. Fasa 2: Pembangunan algoritma pemilihan fitur merupakan proses yang membangunkan APB bagi memilih subset fitur yang optimum. Hasil daripada fasa ini ialah teknik yang dibangunkan mengikut langkah-langkah yang terdapat dalam analisis sentimen.

C. Fasa 3: Pengelasan sentimen akan menghasilkan prestasi pengelasan sentimen berdasarkan empat (4) algoritma pengelas pembelajaran mesin menggunakan set fitur yang dipilih dari fasa ketiga sebagai input. Algoritma pengelas pembelajaran mesin yang terbaik akan dapat diperoleh.

D. Fasa 4: Pengujian, penilaian dan analisis ke atas hasil eksperimen dengan menggunakan beberapa kaedah pengujian, penilaian dan analisis yang telah dikenal pasti. Maklumat terperinci mengenai langkah-langkah yang terlibat dalam setiap langkah ini akan diterangkan dalam bahagian yang seterusnya.



Rajah 3.2 Reka bentuk kajian



Rajah 3.3 Reka bentuk kajian 2

3.4 PERISIAN DAN PLATFORM

Dalam kajian ini, beberapa platfrom dan perisian telah digunakan untuk penyediaan data teks, prapemprosesan data teks dan juga pengujian data teks. Bagi penyediaan data mentah, Perisian Microsoft Excel 2019 dan pangkalan data MySQL telah digunakan kerana perisian ini memudahkan pengurusan data dari segi pengiraan, unjuran, analisis dan juga kemaskini.

Bagi prapemprosesan teks dan pemilihan fitur, penggunaan perisian Python 3.6.4 di platform Notebook Jupyter dengan pemasangannya di unit pemprosesan grafik (GPU) kerana saiz data yang besar telah dipilih. Pengaturcaraan Python telah digunakan bagi membantu proses pengaturcaraan dengan bahasa yang lebih mudah di samping mempercepatkan proses pengujian data teks yang berkapasiti besar. Tambahan pula, Python mempunyai banyak modul perpustakaan yang boleh digunakan bagi setiap fasa eksperimen termasuklah fasa prapemprosesan teks dan juga fasa pemilihan fitur. Dari segi persembahan visualisasi, Python menawarkan Matplotlib yang mana merupakan pangkalan untuk pembangunan perpustakaan seperti *seaborn, pandas plotting* dan *ggplot*.

Seterusnya, bagi pengelasan sentimen pula, Pengelasan sentimen ini dilakukan dengan menggunakan perisian WEKA (Waikato Environment for Knowledge Analysis) versi 3.8. Perisian ini merupakan perisian sumber terbuka yang ditulis dalam bahasa pengaturcaraan Java dan diedarkan di bawah terma GNU (General Public License). Ia boleh digunakan dalam pelbagai platform dan telah digunapakai dalam pelbagai sistem operasi komputer seperti Windows, Linux dan Macintosh. Dalam perisian WEKA, algoritma SVM, Naïve Bayes, k-Jiran terdekat dan Pokok Keputusan masing-masing dikenali sebagai LibSVM, Naïve Bayes, IBk dan J48.

3.5 FASA I: PRAPEMPROSESAN DATA TEKS

Proses ini melibatkan pembersihan dan penyediaan set data yang dipilih ke dalam bentuk yang membolehkan set data ini diproses pada fasa pemilihan fitur. Input kepada fasa ini merupakan set data tanda aras ulasan pengguna dari laman media sosial Twitter yang telah dilabelkan dengan maklumat sentimen. Set fitur yang

dihasilkan akan digunakan untuk kegunaan fasa pembangunan algoritma pemilihan fitur.

3.5.1 Pengumpulan Set Data

Set data tanda aras yang digunakan dalam kajian ini diperoleh daripada laman web Kaggle yang boleh dicapai melalui https://www.kaggle.com/c/si650winter11/data dan telah digunakan (A. Kumar et al. 2018). Set data ini merupakan ulasan pengguna tentang pelbagai isu dan perkara di mana ia telah dikumpulkan pada tahun 2010 dari laman blog mikro Twiter dan disimpan dalam format fail teks (txt). Set data ini juga pernah digunakan oleh (A. Kumar & Jaiswal 2019) bagi kajian terhadap prestasi pengelasan sentimen menggunakan algoritma metaheuristik.

Set data ini mengandungi 7086 ulasan yang ditulis dengan menggunakan bahasa Inggeris. Sebanyak 4005 daripada jumlah keseluruhan data merupakan ulasan yang telah dilabelkan sebagai positif dan baki sebanyak 3081 adalah sebagai negatif. Pada Rajah 3.4 pula merupakan sebahagian contoh set data daripada set data ulasan yang mengandungi ayat ulasan sebelum berlakunya proses prapemprosesan teks. Pada rajah ini juga memaparkan sebahagian contoh set data ulasan Twitter.

Label Sentimen	**Ulasan Twitter**
1	I still like Tom Cruise.
0	well, i had a piece of crap toyota celica but it died in portland and i got a ford ranger..
1	i love angelina jolie.
1	I still like Tom Cruise.
1	UCLA is beautiful.
0	I think Angelina Jolie is so much more beautiful than Jennifer Anniston, who, by the way, is majorly OVERRATED.
1	Angelina Jolie is beautiful.
1	and honda's are awesome:).
1	I love Harvard.
1	i love tom cruise!..

Rajah 3.4 Sebahagian Contoh Set Data Ulasan Twitter

3.5.2 Reka Bentuk Kaedah Pemprosesan Teks

Dalam proses ini, setiap ayat ulasan pengguna yang terdapat dalam fail txt akan di ekstrak dan disimpan di dalam pangkalan data awan terlebih dahulu. Setiap ayat

ulasan yang di ekstrak akan disimpan sebagai dokumen di dalam pangkalan data. Set data ini kemudiannya akan melalui dua jenis pemprosesan teks iaitu pemprosesan linguistik dan pemprosesan bahasa tabii.

1. Kaedah pemprosesan teks bagi tanda aras

Teknik pemprosesan linguistik yang digunakan dalam eksperimen ini terdiri daripada lima teknik pemprosesan linguistik iaitu menukar huruf kecil, penghapusan @, pembuangan tanda baca, penghapusan tanda pagar # dan penghapusan simbol. Manakala bagi teknik pemprosesan bahasa tabii pula akan melalui *stemming*.

2. Kaedah pemprosesan teks bagi tanda aras dan teknik tambahan

Bagi eksperimen ini pula, teknik pemprosesan linguistik yang digunakan adalah sama seperti pada kaedah pemprosesan teks bagi tanda aras. Hal ini di mana, terdapat lima teknik pemprosesan linguistik iaitu menukar huruf kecil, penghapusan @, pembuangan tanda baca, penghapusan tanda pagar # dan penghapusan simbol. Akan tetapi, bagi teknik pemprosesan bahasa tabii pula akan menggunakan dua teknik yang lain iaitu *lemmatize*, dan pembetulan ejaan.

Maka, secara keseluruhannya terdapat empat model pemprosesan teks yang akan diuji dan dinilai seperti diterangkan di dalam Jadual 3.1. Model-model ini terdiri daripada gabungan kaedah pemprosesan linguistik dan pemprosesan bahasa tabii. Model A iaitu model berdasarkan model tanda aras akan digunakan terlebih dahulu dan kemudian akan melalui fasa pemilihan fitur.

Jadual 3.1 Senarai model pemprosesan teks

Model	**Teknik Pemprosesan Teks**
A (Tanda Aras)	Pemprosesan Linguistik + *Stemming*
B	Pemprosesan Linguistik + *Lemmatize*
C	Pemprosesan Linguistik + Pembetulan Ejaan
D	Pemprosesan Linguistik sahaja

Seksyen yang berikutnya akan menerangkan secara lebih mendalam mengenai proses yang berlaku semasa pemprosesan teks dilaksanakan.

3.5.3 Teknik Pemprosesan Linguistik

Pemprosesan linguistik yang dipilih untuk eksperimen ini terdiri dari lima teknik pemprosesan teks iaitu menukar huruf kecil, penghapusan @, pembuangan tanda baca, penghapusan tanda pagar # dan penghapusan simbol. Sebagai contoh bagi pemprosesan teks, data ulasan di bawah akan gunakan: "*@Piwi_47 I hated the da Vinci code, the movie witha passion, it was boring and made me sad!! that a movie blasphying the name of Jesus is being played world wide $. #davincicode"*

i. Menukar Huruf Kecil

Ini merangkumi penukaran semua huruf besar kepada huruf kecil seperti ditunjukkan dalam Jadual 3.2 .Dalam jadual ini menunjukkan perkataan yang ditulis dengan huruf besar seperti "*Piwi*", "*I*", "*Vinci*" dan "*Jesus"* telah ditukar kepada huruf kecil iaitu "*piwi*", "*i*", "*vinci*" dan "*jesus"*.

Jadual 3.2 Proses menukar ke huruf kecil

Sebelum	Selepas
@Piwi_47 I hated the da Vinci code, the movie witha passion, it was boring and made me sad!! that a movie blasphying the name of Jesus is being played world wide $. #davincicode	*@piwi_47 i hated the da vinci code, the movie witha passion, it was boring and made me sad!! that a movie blasphying the name of jesus is being played world wide $. #davincicode*

ii. Penghapusan @

Proses menghapuskan @ yang terdapat dalam set data seperti seperti ditunjukkan dalam Jadual 3.3. Perkataan yang mengandungi @ akan dihapuskan bersama-sama dengan teks yang bersambung dengannya. Dalam rajah ini telah menunjukkan bahawa *"@piwi_47"* dihapuskan dari data.

Jadual 3.3 Proses Penghapusan @

Sebelum	Selepas
@piwi_47 i hated the da vinci code, the movie witha passion, it was boring and made me sad!! that a movie blasphying the name of jesus is being played world wide $. #davincicode	*i hated the da vinci code, the movie witha passion, it was boring and made me sad!! that a movie blasphying the name of jesus is being played world wide $. #davincicode*

iii. Pembuangan Tanda Baca

Proses pembuangan tanda baca ini merangkumi semua tanda baca yang terdapat dalam penggunaan bahasa antaranya “.”, “,”, “!”, “-”, “:”, “;”, “?”, “/”, “()”, “{}” dan “ ’ ”. seperti yang ditunjukkan pada Jadual 3.4, tanda baca “!” dan “.” akan dihapuskan daripada data tersebut.

Jadual 3.4 Proses Pembuangan Tanda Baca

Sebelum	Selepas
i hated the da vinci code, the movie witha passion, it was boring and made me sad!! that a movie blasphying the name of jesus is being played world wide $. #davincicode	*i hated the da vinci code, the movie witha passion, it was boring and made me sad that a movie blasphying the name of jesus is being played world wide $ #davincicode*

iv. Penghapusan Tanda Pagar #

Proses menghapuskan tanda pagar iaitu “#” yang terdapat dalam set data seperti ditunjukkan dalam Jadual 3.4. Maklumat tanda pagar akan dihapuskan bersama-sama dengan teks yang bersambung dengannya. Dalam jadual ini menunjukkan tanda pagar *“#davincicode”* dihapuskan daripada data.

Jadual 3.5 Proses penghapusan tanda pagar #

Sebelum	Selepas
i hated the da vinci code, the movie witha passion, it was boring and made me sad that a movie blasphying the name of jesus is being played world wide $ #davincicode	*i hated the da vinci code, the movie witha passion, it was boring and made me sad that a movie blasphying the name of jesus is being played world wide $*

v. Penghapusan Simbol

Ia merupakan proses menghapuskan simbol yang terdapat dalam set data seperti ditunjukkan dalam Jadual 3.6. Dalam jadual ini menunjukkan simbol “$” dihapuskan daripada data tersebut.

Jadual 3.6 Proses penghapusan simbol

Sebelum	Selepas
i hated the da vinci code, the movie witha passion, it was boring and made me sad that a movie blasphying the name of jesus is being played world wide $	*i hated the da vinci code, the movie witha passion, it was boring and made me sad that a movie blasphying the name of jesus is being played world wide*

3.5.4 Teknik Pemprosesan Bahasa Tabii

Seterusnya set data yang diperoleh daripada hasil pemprosesan linguistik tadi akan melalui pemprosesan bahasa tabii. Terdapat empat teknik pemprosesan bahasa tabii yang terlibat di dalam eksperimen ini iaitu *stemming*, *lemmatize,* pembetulan ejaan dan pembuangan kata henti di mana ia akan dijelaskan secara terperinci pada seksyen berikutnya. Untuk tujuan ini, data *"it was really ironic that he spent the first part of class talking about his own professot at Harvard who was a pompous arrogant ass"* akan digunakan sebagai contoh bagi setiap proses ini.

i. *Stemming*

Stemming adalah satu proses untuk adalah penukaran perkataan kepada kata dasar perkataan berkenaan set data seperti ditunjukkan dalam Jadual 3.7. Dalam jadual ini menunjukkan perkataan yang ditukar strukturnya.

Jadual 3.7 Proses *stemming*

Sebelum	Selepas
*it was **really ironic** that he spent the first part of class **talking** about his own professot at Harvard who was a pompous **arrogant** ass*	*it was **realli iron** that he spent the first part of class **talk** about his own professot at Harvard who was a pompous **arrog** ass*

ii. Pembetulan Ejaan

Ia merupakan proses membetulkan ejaan set data seperti ditunjukkan dalam Jadual 3.8. Dalam jadual ini menunjukkan perkataan "*professot*" dibetulkan kepada "*professor*".

Jadual 3.8 Proses pembetulan ejaan

Sebelum	Selepas
*it was really ironic that he spent the first part of class talking about his own **professot** at Harvard who was a pompous arrogant ass*	*it was really ironic that he spent the first part of class talking about his own **professor** at Harvard who was a pompous arrogant ass*

iii. *Lemmeatize*

Lemmatization merupakan satu proses bagi mendapatkan asal usul sesuatu perkataan melalui analisis morfologikal set data seperti ditunjukkan dalam Jadual 3.9. Dalam jadual ini menunjukkan perkataan "*talking*" telah ditukar kepada perkataan asalnya.

Jadual 3.9 Proses *lemmatize*

Sebelum	Selepas
*it was really ironic that he spent the first part of class **talking** about his own professot at Harvard who was a pompous arrogant ass*	*it was really ironic that he spent the first part of class **talk** about his own professot at Harvard who was a pompous arrogant ass*

3.6 FASA II: PEMBANGUNAN ALGORITMA PEMILIHAN FITUR

Fasa 2 ini pula merupakan pembangunan algoritma bagi pemilihan fitur daripada set fitur yang dihasilkan dalam Fasa 1. Ia terdiri daripada pembangunan algoritma APB dengan menggunakan pengaturcaraan Python dan pangkalan data awan.

3.6.1 Pembangunan algoritma APB

Pembangunan algoritma APB adalah berdasarkan kod pseudo yang ditunjukkan dalam Rajah 3.6. Kod pseudo ini akan diterangkan secara terperinci dalam bahagian pembangunan algoritma yang diterangkan dalam bahagian seterusnya. Secara umumnya, pembangunan algoritma APB mempunyai daripada tiga bahagian utama iaitu pengisytiharan parameter, permulaan dan pencarian.

Seperti yang ditunjukkan pada Rajah 3.5, set fitur diwakilkan ke dalam bentuk susunan yang akan memudahkan pemproses. Antara contoh senarai fitur termasuklah *"actual", "good", "movie", "awesome", "book", "disappointed", "pretty", "interest", "would", "normal"* yang diperoleh daripada fasa pemprosesan teks

diwakili oleh sel tatasususan. Sebagai contoh, fitur *"movie"* diwakili oleh sel F1, manakala fitur *"book"* diwakili oleh F2 dan seterusnya.

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
movie	*book*	*normal*	*would*	*interest*	*actual*	*good*	*awesome*	*disappointed*	*pretty*

Rajah 3.5 Perwakilan fitur dalam bentuk tatasusunan

```
    Fasa Pengistiharaan Parameter
1   Pengistiharaan saiz populasi
2   Pengistiharaan kriteria perberhentian (jumlah iterasi)
3   Menetapkan kebarangkalian pertukaran
4   Menetapkan penyelesaian terbaik dalam populasi
    Fasa Pengawalan
5   Menjana populasi penyelesaian awal secara rawak
6   Menetapkan nilai penyelesaian awal mengikut nilai kebarangkalian pertukaran
7   Setiap individu dinilai menggunakan fungsi kelayakan
    Fasa Permulaan
8   Semak jika kriteria pemberhentian belum dipenuhi
9   {
        Pendebungaan Global
10      Pilih penyelesaian yang terbaik
11      Penyelesaian secara rawak dipilih dan lebih kecil daripada kebarangkalian pertukaran
12      Proses pendebungaan global berlaku
13      Penilaian penyelesaian baru dihasilkan
14      Jika penyelesaian baru lebih baik
15          Masukkan penyelesaian ke dalam populasi penyelesaian
16      Sebaliknya
17          Penyelesaian baru ditolak
        Pendebungaan Tempatan
18      Pilih dua penyelesaian secara rawak
19      Proses pendebungaan tempatan berlaku
20      Penilaian penyelesaian baru dihasilkan
21      Jika penyelesaian baru lebih baik
22          Masukkan penyelesaian ke dalam populasi penyelesaian
23      Sebaliknya
24          Penyelesaian baru ditolak
25      Langkah 6 diulang
26  }
27  Pulangkan subset fitur yang mempunyai skor terbaik sebagai set fitur berkualiti.
```

Rajah 3.6 Kod pseudo algoritma APB

Sumber: (Alyasseri et al. 2018)

a. Fasa Pengistiharan Parameter

Dalam bahagian pengisytiharan parameter umum, terdapat dua jenis parameter yang perlu ditetapkan iaitu saiz populasi dan juga bilangan iterasi. Untuk eksperimen ini penetapan saiz parameter populasi adalah dibuat berdasarkan kajian yang dilakukan oleh (X. S. Yang et al. 2014), iaitu saiz populasi yang paling sesuai untuk mendapatkan keputusan yang terbaik ialah 25. Untuk bilangan generasi pula, bilangan generasi ditetapkan berdasarkan kajian yang dilaksanakan oleh (X. S. Yang 2012), iaitu bilangan generasi yang paling sesuai ialah 100. Bilangan generasi ini akan menjadi kriteria pemberhentian untuk algoritma ini. Penetapan parameter ini telah dilaksanakan pada kajian domain elektrik dan system kuasa, domain teknologi rangkaian tanpa wayar, dan domain pengelompokan dan pengelasan data (Alyasseri et al. 2018).

b. Fasa Permulaan

Dalam fasa permulaan ini, subset penyelesaian akan dijana secara rawak dan disimpan dalam bentuk tatasusunan satu dimensi seperti contoh yang ditunjukkan dalam Rajah 3.7. Dalam rajah ini menggambarkan satu subset penyelesaian yang terdiri dari sepuluh atribut fitur yang dilabelkan daripada F1 sehingga F10. Sel yang bernilai 1 mewakili atribut fitur yang dipilih dan sel bernilai kosong menggambarkan atribut fitur tersebut tidak dipilih. Dalam contoh rajah ini, atribut fitur yang dipilih adalah F1 (*movie*), F3 (*normal*), F5 (*interest*), F6 (*actuall*), F7 (*good*) dan F8 (*awesome*). Sebanyak dua puluh lima subset penyelesaian awal akan dijana pada peringkat ini dan dipanggil sebagai satu populasi.

Seterusnya semua subset penyelesaian awal yang dijana ini akan dinilai berdasarkan prestasi penilaian ketepatan pengelasan sentimen dan disusun berdasarkan nilai skor ketepatan pengelasan yang diperoleh. Subset penyelesaian permulaan ini akan dijadikan menjana subset generasi seterusnya.

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
1	0	1	0	1	1	1	1	0	0

Rajah 3.7 Contoh tatasusunan subset penyelesaian

c. Fasa Pencarian

Bagi fasa pencarian, algoritma ini terbahagi kepada dua iaitu pendebungaan global dan pendebungaan tempatan. Sebelum meneruskan ke fasa pencarian, kriteria pemberhentian akan disemak terlebih dahulu. Sekiranya kriteria pemberhentian tidak dipenuhi, fasa pencarian akan diteruskan. Sebaliknya jika kriteria pemberhentian telah dipenuhi, subset fitur yang mempunyai skor pengelasan sentimen yang terbaik akan dipulangkan dan digunakan sebagai senarai fitur untuk pengelasan sentimen.

i. Pendebungaan Global

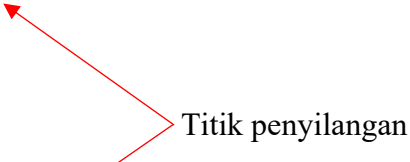
Pendebungaan global ini bermula apabila penyelesaian secara rawak dijana dan jika nilainya adalah lebih kecil berbanding dengan kebarangkalian pertukaran, penyelesaian baharu akan dijana oleh peraturan taburan Levy. Penyelesaian sedia ada akan dipilih bersama penyelesaian baharu ini untuk mendapatkan skor penyelesaian lebih baik. Pada pendebungaan global ini berlakunya kaedah penyilangan seperti dalam Rajah 3.8.

Sesi pendebungan global ini akan menghasilkan dua penyelesaian baru dan kedua-duanya akan melalui proses penilaian prestasi pengelasan sentimen. Berdasarkan keputusan penilaian ini, hanya penyelesaian baru yang dapat menghasilkan skor yang lebih baik daripada senarai skor penyelesaian sedia ada sahaja yang diterima ke dalam populasi dan kedudukannya dalam populasi akan ditentukan berdasarkan skor yang diperoleh. Sebaliknya, penyelesaian baru ini tidak akan diterima jika mendapat skor yang kurang dari skor populasi sedia ada.

INPUT:

Penyelesaian rawak 1:

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
1	0	1	0	1	1	1	1	0	0



Penyelesaian rawak 2:

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
1	1	0	0	1	0	0	1	0	1

OUTPUT:

Penyelesaian baharu 1:

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
1	0	1	0	1	0	0	1	0	1

Penyelesaian baharu 2:

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
1	1	0	0	1	1	1	1	0	0

Rajah 3.8 Sesi pendebungaan global

ii. Pendebungaan Tempatan

Dalam sesi pendebungaan tempatan pula, ia berlaku apabila penyelesaian secara rawak dijana dan jika nilainya adalah lebih besar berbanding dengan kebarangkalian pertukaran. Oleh itu, dua penyelesaian yang telah dijana secara rawak dipilih untuk melalui pendebungaan tempatan. Dalam algoritma ini, proses pembelajaran di antara kedua-dua pelajar berlaku melalui kaedah penyilangan seperti ditunjukkan dalam Rajah 3.9.

Hasil daripada pendebungaan tempatan ini akan menghasilkan dua penyelesaian baru dan kemudian ianya dinilai berdasarkan prestasi pengelasan sentimen. Berdasarkan skor penilaian sentimen yang diperoleh, sekiranya penyelesaian baru ini lebih baik daripada penyelesaian sedia ada, ia akan diterima dan kedudukannya dalam populasi akan dikemas kini. Sebaliknya, penyelesaian ini tidak akan diterima sekiranya ia mendapat skor yang rendah daripada populasi penyelesaian sedia ada.

INPUT:

Penyelesaian rawak 1:

F1	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**
0	0	1	1	1	0	1	1	0	0

Titik penyilangan

Penyelesaian rawak 2:

F1	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**
1	0	0	1	1	0	1	0	1	1

OUTPUT:

Penyelesaian baharu 1:

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
0	0	1	1	1	0	1	0	1	1

Penyelesaian baharu 2:

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
1	0	0	1	1	0	1	1	0	0

Rajah 3.9 Sesi pendebungaan tempatan

Kedua-dua sesi pendebungaan global dan pendebungaan tempatan ini akan diulang sehingga kriteria perbehentian dipenuhi. Sekiranya kriteria berhenti tidak dipenuhi, proses pembaikan ini akan diulang semula. Output kepada algoritma ini adalah subset fitur berkualiti dan bersaiz kecil yang akan digunakan dalam fasa pengelasan sentimen seperti contoh yang ditunjukkan dalam Rajah 3.10. Dari rajah ini ditunjukkan bahawa fitur yang dipilih sebagai subset fitur yang dipilih ialah F1(*movie*), F3(*normal*), F5(*interest*), F6(*actual*) dan F7(*good*).

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
1	0	1	0	1	1	1	0	0	0

Rajah 3.10 Subset fitur yang dipilih

3.7 FASA III: PENGELASAN SENTIMEN

Pada fasa ketiga ini melibatkan empat algoritma pengelas pembelajaran mesin untuk menjalankan proses pengelasan sentimen dengan memadankan subset fitur yang dihasilkan dengan maklumat fitur yang terdapat dalam ulasan pengguna.. Algoritma yang digunakan adalah Mesin Vektor Sokongan (SVM), Naive Bayes (NB), k-Jiran Terdekat (kNN) dan Keputusan Pokok (DT).

Pengelasan sentimen ini dilakukan dengan menggunakan perisian WEKA (Waikato Environment for Knowledge Analysis) versi 3.8. Perisian ini merupakan perisian sumber terbuka yang ditulis dalam bahasa pengaturcaraan Java dan diedarkan di bawah terma GNU (General Public License). Ia boleh digunakan dalam pelbagai platform dan telah digunapakai dalam pelbagai sistem operasi komputer seperti Windows, Linux dan Macintosh.

Bagi fasa pengelasan ini, kaedah 10 lipatan pengesahan silang digunakan dalam kajian ini. Pengesahan silang digunakan terutamanya dalam pembelajaran mesin untuk menganggar kemahiran model pembelajaran mesin pada data yang tidak kelihatan. Hal ini di mana untuk menggunakan sampel yang terhad bagi tujuan menganggarkan prestasi model secara umum apabila digunakan untuk membuat ramalan pada data yang tidak digunakan semasa model latihan. Ia adalah kaedah yang popular kerana ia mudah difahami dan kerana ia biasanya menghasilkan anggaran kemahiran model yang kurang berat sebelah atau kurang optimis berbanding kaedah lain, seperti pendekatan pemisahan latihan/ujian.

3.8 FASA IV: PENGUJIAN, PENILAIAN DAN ANALISIS

Pada fasa keempat ini merangkumi proses pengujian, penilaian dan analisis hasil daripada eksperimen yang telah dijalankan dengan menggunakan beberapa kaedah yang telah dikenal pasti. Pengujian dan penilaian ini dilakukan dengan berdasarkan keputusan eksperimen pengelasan sentimen menggunakan subset fitur yang telah dipilih oleh algoritma pemilihan fitur.

Eksperimen prestasi pengelasan yang digunakan diuji dengan menggunakan matriks kekeliruan. Matriks kekeliruan berperanan dalam menunjukkan maklumat mengenai jumlah sebenar sesuatu kelas dan jumlah ramalan yang dijana oleh algoritma pengelasan. Positif benar (TP) ialah keadaan di mana kes positif berjaya kelaskan sebagai positif. Negatif benar (TN) ialah keadaan di mana kes negatif berjaya dikelaskan sebagai negatif. Positif palsu (FP) ialah kes negatif tetapi disalah kelaskan sebagai positif. Negatif palsu (FN) ialah kes positif tetapi disalah kelaskan sebagai kes negatif seperti ditunjukkan dalam Jadual 3.10.

Jadual 3.10 Matriks Kekeliruan

		Kelas Sebenar	
		Ya	**Tidak**
Keputusan Pengelasan	**Ya**	TP	FN
	Tidak	FP	TN

Terdapat tiga kriteria pengujian prestasi yang digunakan iaitu ketepatan (*a*), kejituan (*p*) dan dapatan semula (*r*). Nilai-nilai pengukuran ini digunakan untuk menilai prestasi pengelasan output yang di hasilkan daripada teknik yang digunakan. Persamaan (3.1) berikut adalah persamaan yang digunakan untuk nilai ketepatan di mana jumlah nilai ramalan yang tepat (TP) dan (TN) dibahagikan dengan jumlah keseluruhan ramalan (TP), (TN), (FP) dan (FN):

$$a = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad \ldots(3.1)$$

Kejituan adalah ukuran yang menunjukkan bilangan positif benar (TP) dibahagikan dengan semua ramalan yang positif. Ketepatan juga dipanggil Nilai Prediktif Positif. Ia juga adalah ukuran ketepatan pengelas. Kejituan rendah menunjukkan bilangan positif palsu adalah tinggi. Persamaan (3.2) menunjukkan formula pengiraan kejituan:

$$p = \frac{TP}{TP + FP} \quad \ldots(3.2)$$

Dapatan semula adalah bilangan positif benar (TP) dibahagikan dengan bilangan nilai positif dalam set data ujian. Dapatan semula juga dikenali sebagai

Sensitiviti atau Kadar Positif Benar di mana ia adalah ukuran kesempurnaan pengelas. Jika bilangan dapatan semula adalah rendah ia menunjukkan bilangan negatif adalah tinggi. Formula kiraan parameter dapatan semula adalah seperti persamaan (3.3):

$$r = \frac{TP}{TP + FN} \quad \text{…(3.3)}$$

Bagi tujuan perbandingan dan penilaian ke atas algoritma yang digunakan, pengujian dengan beberapa kajian yang dipilih. Kriteria kajian yang dipilih untuk perbandingan dan penilaian ialah kajian yang menggunakan algoritma pemilihan fitur yang berlainan, tetapi menggunakan set data ulasan pengguna yang sama dan juga menggunakan kaedah pengelasan sentimen yang sama. Akhir sekali, analisis dilakukan ke atas keputusan eksperimen bagi melihat prestasi algoritma APB untuk pemilihan fitur dalam pengelasan sentimen.

3.9 KESIMPULAN

Secara keseluruhannya, bab metodologi kajian ini berfungsi untuk menerangkan fasa-fasa penting yang dijalankan dalam kajian ini. Ia terdiri dari lima fasa utama iaitu kajian kesusasteraan, pemprosesan teks, pembangunan algoritma pemilihan fitur, fasa pengelasan sentimen serta fasa pengujian, penilaian dan analisis.

Pada fasa pemprosesan data set data ulasan pengguna akan melalui proses pembersihan dan transformasi bagi menyediakan set data untuk proses fasa seterusnya. Dalam fasa ini juga, eksperimen akan dilaksanakan bagi mendapatkan gabungan kaedah pemprosesan yang terbaik daripada segi ketepatan pengelasan sentimen.

Kemudian ia diikuti dengan fasa pembangunan algoritma pemilihan fitur pula, APB dibangunkan untuk memproses set fitur yang telah dihasilkan. Pembangunan APB berfungsi untuk melakukan proses untuk memilih fitur berkualiti untuk pengelasan sentimen. Subset fitur yang dihasilkan ini kemudiannya akan dihantar ke fasa seterusnya.

Seterusnya, fasa pengelasan sentimen pula akan menggunakan pengelas pembelajaran mesin untuk mengelaskan sentimen set data ulasan pengguna berdasarkan subset fitur yang telah diperoleh. Output kepada fasa ini merupakan keputusan pengelasan yang telah dilakukan untuk diproses ke dalam fasa seterusnya.

Fasa pengujian, penilaian dan analisis ialah fasa terakhir dalam kajian ini. Pengujian dan penilaian ke atas keputusan fasa sebelum ini dijalankan dengan menggunakan beberapa teknik dan kajian lepas yang telah dikenal pasti. Seterusnya analisis daripada pengujian dan penilaian ini akan diterangkan dan dijelaskan dengan lebih lanjut.

BAB IV

DAPATAN KAJIAN

4.1 PENGENALAN

Di dalam bab ini akan dibincangkan tentang hasil eksperimen yang telah dijalankan. Terdapat dua eksperimen utama yang telah dijalankan dalam penyelidikan ini iaitu eksperimen tentang pemilihan fitur menggunakan teknik APB bagi tujuan pengelasan sentimen serta eksperimen tentang pemilihan teknik pra-pemprosesan teks yang betul dan sesuai. Bagi mencapai matlamat penyelidikan ini, eksperimen-eksperimen ini telah dibangunkan dengan menggunakan bahasa pengaturcaraan Python.

Subseksyen yang berikutnya akan menerangkan hasil prestasi algoritma pemilihan fitur yang menggunakan pendekatan APB dengan lebih terperinci. Eksperimen yang dijalankan akan menguji keberkesanan algoritma pemilihan fitur yang dicadangkan di mana keputusan yang diperoleh dianalisis dan dibincangkan. Bagi penilaian pemilihan teknik prapemprosesan teks pula, eksperimen dijalankan dan hasil keputusan eksperimen akan dibandingkan dengan model asas.

Prestasi eksperimen ini diukur dengan menggunakan metrik pengukuran ketepatan, kejituan dan dapatan semula pengelasan. Metrik ketepatan merupakan peratusan nisbah ramalan yang betul berbanding keseluruhan ramalan yang telah dijalankan. Nilai ketepatan yang tinggi menunjukkan keputusan ramalan yang baik. Nilai kejituan adalah keupayaan untuk memberikan ramalan yang betul manakala dapatan semula menunjukkan keupayaan untuk mengenal pasti semua ramalan yang betul (Goutte & Gaussier 2005).

Bab ini terdiri daripada penerangan keputusan eksperimen yang akan yang telah dijalankan iaitu terdiri daripada eksperimen pemilihan kaedah pemprosesan teks dan juga eksperimen pemilihan fitur dengan menggunakan algoritma APB. Seterusnya, bab ini diteruskan bahagian kedua yang menyatakan rumusan yang diperoleh daripada hasil eksperimen yang telah dijalankan

4.2 KEPUTUSAN EKSPERIMEN I

Pada seksyen ini akan menerangkan hasil eksperimen berkaitan dengan kaedah penggunaan teknik pemprosesan teks.

4.2.1 Kaedah Prapemprosesan Teks

Pada fasa ini terdapat empat kategori model gabungan pemprosesan teks yang dibandingkan dalam eksperimen ini iaitu Model A yang terdiri gabungan kaedah pemprosesan linguistik dan *stemming*. Seterusnya model B yang terdiri daripada gabungan kaedah pemprosesan linguistik dan *lemmatize*. Kemudian, model C yang merupakan gabungan kaedah pemprosesan linguistik dan juga pembetulan ejaan. Akhir sekali ialah model D yang hanya menggunakan pemprosesan linguistik sahaja. Set data ulasan yang telah melalui model gabungan pemprosesan teks ini seterusnya akan melalui proses tokenisasi untuk menghasilkan senarai fitur. Langkah terakhir ialah senarai fitur yang dihasilkan daripada model-model ini akan diuji dengan algoritma pengelasan iaitu SVM untuk mendapatkan nilai ketepatan pengelasan kerana pada eksperimen pemilihan fitur pengelas SVM telah mencapai keputusan pengelasan yang terbaik berbanding 3 yang lain.

Daripada eksperimen yang dijalankan, didapati bahawa model C memperoleh bacaan kadar ketepatan pengelasan yang terbaik iaitu 99.15% berbanding dengan model A yang mencapai 98.99% , seterusnya model B mencapai 98.14% dan model D dengan 97.4%. Keputusan berdasarkan kejituan juga menunjukkan model C mempunyai kejituan yang terbaik iaitu 99%. Model C juga memberikan keputusan metrik pengukuran dapatan semula yang terbaik iaitu 99.15%. Dengan itu model C memberikan keputusan keseluruhan terbaik berbanding dengan model-model lain.

Jadual 4.1 menunjukkan perbandingan prestasi pemprosesan teks berdasarkan eksperimen yang telah dijalankan.

Jadual 4.1 Perbandingan hasil model prapemprosesan teks

Model	Ketepatan	Kejituan	Dapatan Semula
Model A	98.99%	99.0%	99.0%
Model B	98.14%	98%	98.14%
Model C	**99.15%**	**99.2%**	**99.15%**
Model D	97.4%	95.2%	97.4%

Selain daripada perbandingan hasil model pemprosesan teks, penentuan model terbaik juga ditentukan berdasarkan jumlah skor kedudukan bagi setiap model berdasarkan metrik penilaian ketepatan, kejituan dan dapatan semula. Jadual 4.2 menunjukkan model C berada di kedudukan yang terbaik dengan skor 3 dan diikuti oleh model A dengan skor 6 dan seterusnya model B dengan skor 9 dan akhir sekali model D dengan skor 12. Oleh itu berdasarkan kedudukan skor ini, model A dipilih sebagai model pemprosesan teks yang akan digunakan dalam kajian ini.

Jadual 4.2 Skor kedudukan hasil model prapemprosesan teks

Model	Ketepatan	Kejituan	Dapatan Semula	Jumlah Skor	Skor Akhir
Model A	2	2	2	6	2
Model B	3	3	3	9	3
Model C	**1**	**1**	**1**	**3**	**1**
Model D	4	4	4	12	4

4.3 KEPUTUSAN EKSPERIMEN II

Seksyen ini akan menerangkan keputusan eksperimen pembangunan algoritma APB sebagai algoritma pemilihan fitur.