PEMBINAAN LEKSIKON SENTIMEN BAGI ULASAN RESTORAN MENGGUNAKAN SET GOLONGAN KATA DAN SKOR SENTIMEN

NIK AFIFAH BINTI NIK ABDUL RAHMAN

UNIVERSITI KEBANGSAAN MALAYSIA

PEMBINAAN LEKSIKON SENTIMEN BAGI ULASAN RESTORAN MENGGUNAKAN SET GOLONGAN KATA DAN SKOR SENTIMEN

NIK AFIFAH BINTI NIK ABDUL RAHMAN

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN DARIPADA SYARAT MEMPEROLEH IJAZAH SARJANA SAINS DATA

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2024

PENGAKUAN

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

04 Februari 2024

NIK AFIFAH BINTI NIK ABDUL RAHMAN
P118061

PENGHARGAAN

Dengan nama Allah Yang Maha Pemurah lagi Maha Mengasihani. Syukur ke hadrat Illahi dengan limpah kurniaNya dapat saya menyiapkan kajian projek akhir ini dengan jayanya. Kejayaan ini adalah dorongan banyak pihak.

Ucapan terima kasih yang tidak terhingga kepada penyelia saya iaitu Prof. Madya Dr. Sabrina binti Tiun atas kesabaran, tunjuk ajar, bantuan dan ruang yang diberikan kepada saya. Banyak ilmu, dan pengalaman yang saya timba sepanjang perjalanan menyiapkan kajian projek akhir ini. Ucapan ribuan terima kasih juga kepada semua pensyarah FTSM yang telah mencurahkan ilmu dan memberikan tunjuk ajar sepanjang pengajian saya di UKM.

Terima kasih juga kepada penaja pengajian saya, Jabatan Perkhidmatan Awam Malaysia di atas pembiayaan pengajian ini dan peluang yang diberikan.

Untuk keluarga tercinta khususnya ayahanda dan bonda tersayang, Haji Nik Abdul Rahman bin Nik Kob dan Hajjah Wan Faridah Mardziah binti Wan Mohamed yang sentiasa mendoakan kejayaan, kelapangan, kemudahan dan kesejahteraan untuk saya. Teristimewa untuk suami tercinta Mohd Firdaus bin Abdul Rahman yang sentiasa menjadi tulang belakang, penyokong kuat dan pendengar setia tanpa mengira masa. Tidak dilupakan ahli keluarga yang lain dan terima kasih atas doa kalian yang sentiasa mendoakan dan memberi sokongan kepada saya tanpa mengira waktu.

Kepada rakan-rakan seperjuangan dalam pengajian bidang sarjana sains data UKM, terima kasih di atas bantuan dalam berkongsi ilmu dan pandangan. Akhir sekali, sekalung terima kasih buat semua yang terlibat sama ada secara langsung mahupun tidak langsung sepanjang pengajian ini dijalankan. Terima kasih dan jasa kalian tidak akan saya lupakan. Semoga Allah memberi rahmat yang tidak terhingga kepada kalian.

ABSTRAK

Kepesatan teknologi internet dan kemajuan dunia digital pada masa kini menjadi dasar kepada perkembangan jaringan perhubungan telefon dan juga teknologi tanpa wayar. Segala maklumat dapat disalurkan dan dikongsi dengan mudah dihujung jari. Dengan perkembangan yang positif ini dan ramai penggunanya maka setiap saat dan ketika banyak data dikumpulkan serta disimpan. Maklum balas, ulasan dan pandangan berkenaan sesuatu perkhidmatan atau produk melalui media sosial seperti Twitter, Facebook dan TikTok menggambarkan sentimen pelanggan terhadap perkhidmatan atau produk tersebut. Data ulasan dalam bentuk teks ini perlu diproses untuk mendapatkan maklumat yang berguna daripadanya. Analisis sentimen merupakan teknik yang amat tepat diaplikasikan untuk mengkaji ulasan-ulasan ini. Sentimen diklasifikasikan kepada polariti positif dan negatif. Terdapat tiga teknik pendekatan yang biasa digunakan dalam analisis sentimen iaitu pendekatan berasaskan leksikon, pendekatan berasaskan pembelajaran mesin dan pendekatan hibrid. Pendekatan berasaskan leksikon menjadi pilihan kerana bersesuaian dengan sifat data kerana ia tidak memerlukan latihan. Antara unsur penting dalam membina model analisis sentimen yang baik adalah pembinaan leksikon sentimen. Terdapat dua teknik dalam pembinaan leksikon sentimen iaitu secara automatik dan secara manual yang menggunakan kepakaran pakar bidang tertentu. Pembinaan leksikon sentimen secara automatik dipilih kerana teknik ini memberikan liputan yang jauh lebih baik berbanding dengan leksikon manual dan ia dapat memproses set data yang besar dengan cepat. Kajian ini mengkaji secara terperinci mengenai set golongan kata (POS) dan skor sentimen yang terbaik untuk membina sentimen leksikon secara automatik bagi set data ulasan restoran. Sebanyak 1,000 ulasan positif dan negatif diuji. Kaedah berasaskan leksikon digunakan untuk menjalankan analisis sentimen. Tiga jenis skor sentimen iaitu nilai skor 1 dan 0, TF-IDF dan PMI digunakan untuk mencari perbandingan antara dua jenis set golongan kata. Matlamat kajian ini adalah untuk membina set leksikon sentimen secara automatik terbaik bagi pendekatan analisis sentimen berasaskan leksikon bagi set data ulasan restoran. Untuk mendapatkan ketepatan yang tinggi set data terlebih dahulu dimurnikan dengan langkah-langkah prapemprosesan seperti menyeragamkan teks kepada huruf kecil, membuang perkataan yang tidak mempunyai maksud, membuang semua tanda baca dan menjadikan semua perkataan kepada kata akar. Kajian perbandingan ini mengambil hasil ketepatan yang tertinggi untuk menentukan set leksikon sentimen yang terbaik. Setelah membandingkan prestasi antara analisis sentimen model yang dibina, analisis sentimen model yang terbaik dalam kajian ini adalah model yang menggunakan set PoS-2 dengan skor sentimen TF-IDF di mana ketepatan diperoleh melebih 77%.

CONSTRUCTION OF A SENTIMENT LEXICON FOR RESTAURANT REVIEWS USING PART OF SPEECH AND SENTIMENT SCORE

ABSTRACT

The rapid pace of internet technology and the advancement of the digital world today is the basis for the development of telephone communication networks and wireless technology. All information can be easily channeled and shared at your fingertips. With this positive development and many users, at any time and when a lot of data is collected and stored. Feedback, reviews and views on a service or product through social media such as Twitter, Facebook and TikTok reflect customer sentiment towards the service or product. This review data in text form need to be processed in order to obtain useful information from it. Sentiment analysis is a very accurate technique used to study these reviews. Sentiments are classified into positive and negative polarities. There are three approach techniques commonly used in sentiment analysis namely lexicon-based approach, machine learning-based approach and hybrid approach. A lexicon-based approach is preferred because it is compatible with the nature of the data because it does not require training. Among the important elements in building a good sentiment analysis model is the construction of a sentiment lexicon. There are two techniques in the construction of sentiment lexicon, namely automatically and manually that use the expertise of experts in a particular field. The construction of a sentiment lexicon is automatically chosen because this technique provides much better coverage compared to a manual lexicon and it can process large datasets quickly. This study examines in detail the best part of speech (POS) and sentiment scores to automatically build a sentiment lexicon for the Restaurant Reviews dataset. A total of 1,000 positive and negative reviews were tested. Lexicon-based methods are used to conduct sentiment analysis. Three types of sentiment scores, namely score values 1 and 0, TF-IDF and PMI, were used to find comparisons between two types of part of speech. The aim of this study was to automatically build the best sentiment lexicon set for a lexicon-based sentiment analysis approach for restaurant review datasets . To obtain high accuracy the data set is first purified by preprocessing steps such as standardizing the text to lowercase, removing words that have no meaning, removing all punctuation marks and turning all words to the root word. This comparative study took the highest accuracy results to determine the best set of sentiment lexicons. After comparing the performance of the sentiment analysis model built, the best sentiment analysis model in this study is a model that uses POS-2 set with TF-IDF sentiment score where accuracy is obtained over 77%.

KANDUNGAN

Halaman

SENARAI JADUAL

SENARAI ILUSTRASI

SENARAI SINGKATAN

BOW	Bag-of-Words
DT	Pepohon Keputusan
KNN	K-Jiran Terdekat
NLP	Natural Language Processing
NPMI	Weighted Normalized Pointwise Mutual Information
PMI	Maklumat Timbal Balik Titik
PoS	Part of Speech
RF	Hutan Rawak
SVM	Mesin Vektor Sokongan
TF-IDF	Term Frequency Inverse Document Frequency
UKM	Universiti Kebangsaan Malaysia
VADER	Valence Aware Dictionary and sEntiment Reasoner

BAB I

PENGENALAN

1.1 PENDAHULUAN

Daya tarikan kepada sesuatu perkhidmatan atau produk boleh berpunca daripada pendapat dan pandangan pelanggan yang pernah menggunakan perkhidmatan atau produk tersebut. Melalui pembelian dalam talian pengguna lebih mudah mendapatkan maklumat perkhidmatan atau produk melalui ulasan yang diberikan dan mengetahui baik buruk. Menurut tinjauan yang dijalankan oleh (‘ L o c a l C o n s u m e r R e v
2 0 2 3 : C u s t o m e r R e v i e 98% pengguna membaca ulasan r ’ n .
perniagaan tempatan dalam talian. Dalam kajian ini fokus utama pengkajian adalah sentimen berkaitan perkhidmatan dan produk restoran. Ulasan daripada pelanggan terdahulu impaknya sama ada dapat menarik lebih ramai pelanggan baharu untuk mencuba perkhidmatan dan produk yang ditawarkan di sesebuah restoran atau sebaliknya.

Sebelum kemajuan dunia internet seseorang akan bertanya kepada individu yang lain berkaitan pengalaman menggunakan sesuatu produk atau perkhidmatan. Dalam kepesatan dan kepantasan dunia digital sekarang segala maklumat adalah dihujung jari begitu jua maklum balas, pandangan dan ulasan ke atas perkhidmatan atau produk boleh diperolehi dalam masa nyata. Pengguna menjadikan media sosial seperti Twitter, Facebook dan TikTok antaranya sebagai medium perantara. Kebanjiran data bertekstual yang dijana daripada media sosial ini boleh dijadikan sumber maklumat yang berharga dan utama untuk mengkaji sentimen pengguna. Namun, antara masalah yang dihadapi dengan penilaian dalam talian adalah ulasan yang mengandungi kesalahan ejaan, sintaks bahasa dan tatabahasa menjadi satu kekangan untuk memahami

serta menganalisa ulasan tersebut Liapakis et al. (2020). Di sini analisis sentimen berperanan untuk merungkai masalah yang dihadapi.

Analisis sentimen merupakan kaedah Pemprosesan Bahasa Tabii (NLP) dalam menganalisis teks digital untuk mengesan nada emosi, pendapat dan sikap pengguna, isu-isu, peristiwa atau topik perbincangan di sebalik kandungan media sosial Nerabie et al. (2021). Analisis sentimen juga menjadikan pengalaman yang lepas sebagai panduan untuk membuat keputusan sekarang yang lebih tepat Nandi & Agrawal (2016). Ia merupakan proses mengumpul dan menganalisa pendapat dan pandangan ke atas perkhidmatan atau produk. Dengan teknik ini, kita dapat mengenal pasti nada emosi secara automatik melalui teks sama ada secara masa nyata dengan tepat walaupun berurusan dengan jumlah data yang berskala besar dan dari sumber yang pelbagai.

Dalam bidang analisis sentimen terdapat tiga teknik pendekatan yang biasa digunakan iaitu pendekatan berasaskan leksikon, pendekatan berasaskan pembelajaran mesin dan gabungan antara pembelajaran mesin dan leksikon iaitu pendekatan hibrid.

Pendekatan berasaskan leksikon adalah pendekatan tanpa selia melibatkan pengiraan orientasi dokumen teks daripada orientasi semantik perkataan atau frasa dalam dokumen. Polariti positif, negatif atau neutral sesuatu teks juga dapat dihuraikan. Pendekatan ini tidak memerlukan jumlah data latihan yang mempunyai anotasi manusia yang mencukupi untuk mendapatkan hasil yang boleh diterima Enjop et al. (2022). Oleh kerana pendekatan ini tidak memerlukan data latihan disebabkan itu ia boleh dikategorikan sebagai teknik pembelajaran tanpa selia.

Walaupun pendekatan ini mempunyai kelemahan iaitu kebergantungan kepada domain di mana sesuatu perkataan boleh juga mempunyai pelbagai takrifan mengikut perspektif dan penggunaan perkataan tersebut. Sebagai contoh penggunaan perkataan "besar" boleh memberi sentimen yang positif atau negatif. Penggunaan perkataan "besar" dalam ayat "buah tembikai itu masak berbanding "badan bersaiz besar" yang sel
Namun dalam kajian ini, pendekatan ini menjadi pilihan kerana bersesuaian dengan sifat data kerana ia tidak memerlukan latihan, oleh itu label ditentukan secara manual

dan makna perkataan boleh diakses dengan pantas (Yıl ma. Manakala pendekatan pembelajaran mesin memerlukan data latihan untuk dipelajari dan digeneralisasikan. Selain itu peraturan dan corak dalam pendekatan berasaskan leksikon adalah lebih telus dan mudah ditafsir.

Pendekatan berasaskan pembelajaran mesin merupakan teknik yang menggunakan algoritma pembelajaran mesin seperti *Naïve Bayes, Support Vector Machine, Logistic Regression, Decision Tree* dan sebagainya untuk mendapatkan keputusan sentimen. Set data dibahagikan kepada data latihan dan pengujian untuk diklasifikasikan. Terdapat dua kaedah yang utama iaitu pembelajaran dengan selia dan tanpa selia. Pembelajaran dengan selia menggunakan set data yang mempunyai label manakala pembelajaran tanpa selia set data yang tidak mempunyai label. Para penyelidik terdahulu telah membuktikan bahawa kaedah pembelajaran dengan selia adalah lebih banyak digunakan dan memberikan keputusan yang lebih tepat berbanding tanpa selia Nandi & Agrawal 2016; Wankhade et al. (2022). Walau bagaimanapun pendekatan ini memerlukan data latihan yang besar untuk mendapatkan hasil yang baik.

Manakala pendekatan hibrid merupakan gabungan antara pendekatan berasaskan leksikon dan pendekatan pembelajaran mesin. Gabungan antara dua pendekatan ini telah mengaplikasikan penggunaan kaedah statistik dan kaedah berasaskan pengetahuan bagi pengecaman polariti. Kaedah yang biasa dilaksanakan adalah dimulakan dengan pendekatan berasaskan leksikon diikuti dengan pendekatan pembelajaran mesin. Ali & Hameed (2017) pendekatan hibrid telah meningkatkan prestasi klasifikasi berbanding pendekatan yang menggunakan pembelajaran mesin dan pendekatan leksikon sahaja. Menurut Rajeswari et al. (2020) polariti neutral telah dapat dikenalpasti selain daripada positif dan negatif dengan menggunakan pendekatan hibrid

Penandaan golongan kata (POS) adalah proses untuk mengekstrak lebih banyak maklumat daripada teks asal di mana setiap perkataan akan dilabelkan mengikut kumpulan tatabahasa seperti kata nama (*noun*), kata kerja (*verb*), kata sifat (*adjective*) dan lain-lain. Menurut Nerabie et al. (2021) ketepatan hasil sesuatu analisis sentimen bergantung kepada penanda POS di mana setiap perkataan dilabelkan mengikut kumpulan tatabahasa. Terdapat tiga teknik pendekatan yang biasa digunakan dalam

analisis sentimen iaitu pendekatan berasaskan leksikon, pendekatan berasaskan pembelajaran mesin dan pendekatan hibrid. Pendekatan berasaskan leksikon menjadi pilihan kerana ia bersesuaian dengan sifat data kerana tidak memerlukan latihan.

Antara unsur penting dalam membina model analisis sentimen yang baik adalah pembinaan leksikon sentimen. Pembinaan leksikon sentimen melibatkan membuat senarai perkataan atau frasa dengan polariti sentimen seperti positif, negatif atau neutral. Terdapat dua teknik dalam pembinaan leksikon sentimen iaitu secara automatik dan secara manual. Teknik secara manual menggunakan khidmat kepakaran pakar bidang tertentu untuk klasifikasi nilai sentimen sesuatu perkataan. Manakala teknik secara automatik menggunakan pengaturcaraan komputer untuk mengklasifikasi sentimen perkataan.

Justeru itu dengan menggunakan set data ulasan restoran ini serta bereksperimen dengan pendekatan yang dinyatakan adalah sesuai untuk menghuraikan emosi, pendapat dan sikap pelanggan terhadap produk serta perkhidmatan yang ditawarkan. Dengan itu, ia dapat memberi panduan bukan sahaja kepada pengusaha restoran malahan juga kepada pelanggan.

1.2 PENYATAAN MASALAH

Pembinaan leksikon sentimen secara manual dianggap lebih mahal dan memperuntukkan tenaga dan masa pakar bidang dalam tempoh yang lama berbanding secara automatik Dragut et al. (2012). Manakala pembinaan leksikon sentimen secara automatik memberikan liputan yang jauh lebih baik berbanding dengan leksikon manual dan ia dapat memproses set data yang besar dengan cepat. Pembinaan leksikon sentimen secara automatik menggunakan sama ada kamus atau korpus. Menurut Wankhade et al. (2022) kelemahan yang dihadapi dengan pendekatan berasaskan kamus adalah tidak berupaya mencari istilah ulasan dengan domain berorientasikan kandungan tertentu yang tidak termasuk dalam leksikon. Namun sebaliknya, dengan menggunakan pendekatan berasaskan korpus ia berupaya untuk mengenal pasti istilah ulasan dengan orientasi kandungan tertentu di mana apabila domain adalah berbeza akan memberikan hasil yang lebih baik Feng et al. (2018).

Sesuatu ayat boleh terkandung perkataan yang kabur maksudnya dan mempunyai pelbagai makna atau sentimen Saraswathi et al. (2023). Golongan kata boleh membantu dalam memberikan makna dan skor sentimen yang sesuai kepada perkataan berdasarkan konteks sintaksis dan semantiknya. Pemilihan golongan kata adalah langkah penting kerana ia secara langsung memberi kesan kepada ketepatan dalam analisis sentimen.

Dalam pemarkahan sentimen, ternyata terdapat beberapa isu ketidaktepatan dan tidak konsisten yang perlu ditangani Fang & Zhan (2015). Isu utama berpunca daripada penggunaan leksikon sentimen yang lapuk, yang membawa kepada salah tafsir teks dan pemarkahan yang tidak tepat (Maynard et al. 2016). Walau bagaimanapun, meramalkan markah penilaian ulasan menimbulkan cabaran yang berbeza. Ukuran sentimen adalah bergantung pada konteks ayat dan teks di mana berkemungkinan semakin kerap perkataan itu disebut dalam ulasan, semakin tinggi skor sentimennya.

1.3 PERSOALAN KAJIAN

Bagi memastikan objektif kajian tercapai, berikut adalah persoalan kajian yang perlu dihuraikan.

1. Berdasarkan eksperimen yang dijalankan apakah set golongan kata (POS) dan skor sentimen yang terbaik serta memberikan keputusan yang lebih jitu.

1.4 OBJEKTIF KAJIAN

Objektif kajian ini adalah untuk:

1. Mencadangkan set golongan kata yang sesuai untuk model analisis sentimen untuk set data ulasan restoran.
2. Mencadangkan ukuran skor sentimen yang sesuai untuk model analisis sentimen untuk set data ulasan restoran.
3. Membina leksikon sentimen yang sesuai berdasarkan gabungan output objektif 1 dan objektif 2 bagi model analisis sentimen untuk set data ulasan restoran

1.5 SKOP KAJIAN

Skop kajian ini adalah seperti berikut:

1. Menggunakan set data ulasan restoran yang diperolehi daripada laman sesawang *Kaggle Dataset Repository*.
2. Memfokuskan kepada pemilihan penandaan dua set POS iaitu set POS 1 melibatkan penggunaan kata bantu, kata nama, kata kerja dan kata sifat. Manakala set POS 2 ialah kata sifat.
3. Menggunakan skor 1/0, frekuensi songsang jangka kekerapan dokumen (TF-IDF) dan maklumat timbal balik titik (PMI) sebagai skor sentimen.

1.6 KEPENTINGAN KAJIAN

Antara kepentingan dan jangkaan sumbangan kajian ini adalah:

1. Membantu pengusaha restoran mendapatkan pandangan pelanggan dan memahami perasaan serta emosi pelanggan selepas mendapatkan perkhidmatan dan produk daripada restoran.
2. Pengusaha restoran boleh meningkatkan produk atau perkhidmatan untuk memenuhi keperluan dan harapan pelanggan.
3. Membantu pelanggan membuat keputusan dan pemilihan yang tepat sebelum mengunjungi dan menggunakan perkhidmatan atau produk sesuatu restoran.
4. Pemilihan gabungan set POS yang terbaik dan ukuran skor yang sesuai akan dapat membina sentimen leksikon yang lebih relevan seterusnya meningkatkan kualiti dan ketepatan analisis sentimen.
5. Pembentukan leksikon sentimen yang dihasilkan dapat diaplikasikan ke dalam industri restoran dan dalam bahasa set data yang dikaji.

1.7 BAB ORGANISASI

Kandungan laporan ini dibahagi dan distrukturkan kepada beberapa bahagian seperti berikut.

1. **BAB 1** secara umumnya menerangkan pengenalan dan latar belakang tentang analisis sentimen dalam industri restoran. Seterusnya menghuraikan penyataan masalah yang hendak dikaji. Juga persoalan, objektif, skop dan kepentingan kajian.

2. **BAB 2** menghuraikan beberapa penyelidikan yang telah dijalankan terdahulu yang berkaitan analisis sentimen dengan pendekatan berasaskan leksikon, pendekatan berasaskan pembelajaran mesin dan pendekatan hibrid.

3. **BAB 3** menerangkan kaedah yang digunakan dalam membina sentimen leksikon. Ia merangkumi proses pra-pemprosesan dan seni bina eksperimen yang dicadangkan.

4. **BAB 4** melaporkan hasil kajian ke atas dua set POS.

5. **BAB 5** merangkumi kesimpulan dan cadangan penambahbaikan untuk kajian masa hadapan.

1.8 PENUTUP

Bab 1 ini telah menerangkan secara terperinci pengenalan kepada kajian serta motivasi supaya kajian ini dijalankan. Persoalan, objektif, skop dan kepentingan kajian ini dilakukan turut dijelaskan untuk memberi gambaran keseluruhan laporan. Laporan ini dibahagikan kepada 5 Bab yang teridiri daripada Pengenalan, Kajian Kesusasteraan, Metodologi Kajian, Hasil Kajian dan Rumusan yang disusun masing-masing sebagai Bab I, Bab II, Bab III, Bab IV dan Bab V.

BAB II

KAJIAN KESUSASTERAAN

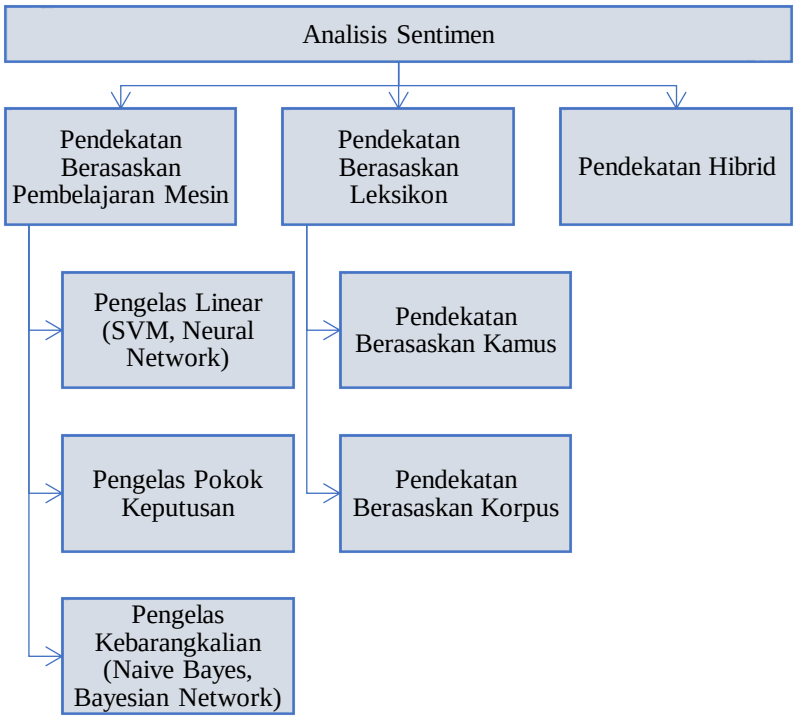
2.1 PENGENALAN

Bab ini membincangkan tentang kajian-kajian terdahulu yang telah dijalankan dalam bidang analisis sentimen. Kajian kesusasteraan dilakukan dengan bertujuan mendapatkan gambaran jelas tentang analisis sentimen, pembinaan leksikon sentimen serta pendekatan-pendekatan yang digunakan dalam analisis sentimen. Ia menfokuskan kepada bidang perkhidmatan dan produk restoran serta makanan. Walau bagaimanapun untuk memberikan pandangan yang lebih meluas dalam analisis sentimen turut dihuraikan kajian-kajian sentimen yang berlainan domain seperti ulasan produk perhotelan, ulasan produk telefon mudah alih dan lain-lain yang boleh juga dijadikan sebagai rujukan. Bukan sekadar ulasan produk malahan jua kajian terhadap sentimen pilihan raya serta sentimen bahasa-bahasa selain Bahasa Inggeris. Kajian-kajian yang telah dijalankan dapat menjadi rujukan dalam perbincangan masalah yang dikaji seterusnya mencadangkan cara penyelesaian yang relevan.

2.2 ANALISIS SENTIMEN

Sepertimana yang dihuraikan pada Bab I, analisis sentimen merupakan kaedah NLP untuk menganalisa teks dan mengklasifikasikan kepada polariti positif, negatif dan neutral. Analisis sentimen dapat menganalisa ekspresi subjektif seperti komen, pandangan, ulasan dan emosi yang pada kelazimannya boleh dilihat pada media sosial seperti Facebook, TikTok dan sebagainya ditukar kepada data kuantitatif. Penggunaan media sosial pada era ini adalah popular diseluruh dunia dan penjanaan data adalah sangat luar biasa. Kuantiti besar data bertekstual yang dijana setiap hari tidak mempunyai nilai melainkan ia diproses untuk mendapatkan maklumat berharga. Dengan pertumbuhan data dalam volum yang besar tersebut memerlukan kewujudan

sistem berautomatik untuk mengklasifikasikan data tersebut berdasarkan aspek yang berbeza Sadia et al. (2018). Menerusi analisis sentimen, ia membolehkan sesebuah organisasi memperoleh pandangan berharga mengenai jenama, produk atau perkhidmatan yang ditawarkan. Analisis sentimen mengkaji matlamat subjektif dalam suatu ekspresi atau ungkapan iaitu pendapat, penilaian, emosi atau sikap terhadap suatu topik, orang atau entiti. Dalam bidang analisis sentimen terdapat tiga teknik pendekatan yang biasa digunakan iaitu pendekatan berasapskan leksikon, pendekatan berasaskan pembelajaran mesin dan pendekatan hibrid. Bahagian ini juga menghuraikan secara umum tiga kaedah yang terdapat dalam analisis sentimen ini. Rajah 2.1 menunjukkan kaedah analisis sentimen.



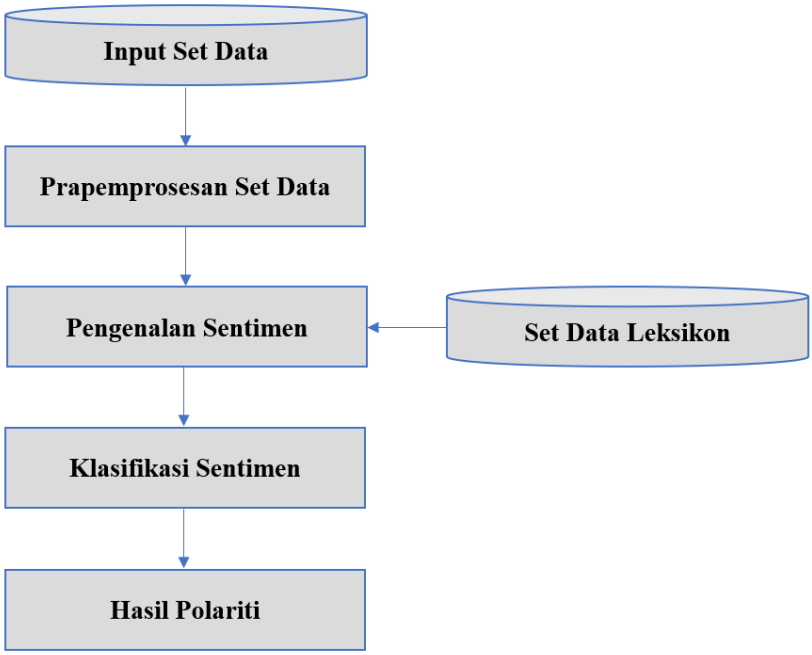
Rajah 2.1 Kaedah Analisis Sentimen

Sumber : Sadia et al. 2018

2.2.1 Pendekatan Berasaskan Leksikon

Pendekatan berasaskan leksikon merupakan pendekatan tanpa seliaan melibatkan pengiraan orientasi dokumen teks daripada orientasi semantik perkataan atau frasa dalam dokumen. Ulasan yang dikaji akan ditokenkan dan diberikan jumlah skor secara berasingan mengikut polariti seperti positif, negatif dan neutral atau berdasarkan

kekuatan polariti dan nilainya dalam julat (+1,-1) d i m a n a + 1 m e w a k i l i manakala -1 ‘ s a n g a W Wankhade et al.t (2022). Rajah 2.2 memaparkan proses umum analisis sentimen berasaskan leksikon.



Rajah 2.2 Proses Umum Pendekatan Berasaskan Leksikon

Sumber : Sadia et al. 2018

Sentimen leksikon merupakan koleksi kosa kata dan frasa yang terkandung p o l a r i t i p o s i t i f , n e g a t i f m a h u p u n n e u t r a b o l e h d i k e l a s k a n s e b a g a i p o l a r i t i p o s i t i selalunya dianggap berpolariti negatif. Manakala polariti neutral adalah perkataan yang tidak membawa pengertian positif mahupun negatif. Sentimen leksikon boleh dianggap sebagai elemen utama untuk menganalisis sentimen Oliveira et al. (2016). Dalam pelombongan pendapat leksikon sentimen adalah sumber penting dalam tugasan analisis sentimen bagi menentukan polariti pendapat. Selain menganalisa sentimen, sentimen leksikon juga amat membantu dalam pelbagai penggunaan seperti perlombongan pendapat dan turut digunakan untuk pemantauan media sosial.

Terdapat dua teknik dalam pembinaan sentimen leksikon iaitu teknik secara manual dan teknik secara automatik Sharmini Alexander (2017). Teknik secara manual menggunakan khidmat kepakaran pakar bidang tertentu untuk klasifikasi nilai sentimen

sesuatu perkataan. Manakala teknik secara automatik menggunakan pengaturcaraan komputer untuk menklasifikasi sentimen perkataan. Teknik secara manual dianggap lebih mahal dan memperuntukkan masa yang lama berbanding secara automatik.

TextBlob, SentiWordNet dan VADER adalah antara alatan leksikon yang umum dalam analisis sentimen berasaskan leksikon yang digunakan untuk meneka nilai sentimen setiap teks. TextBlob, SentiWordNet dan VADER menggunakan bahasa pengaturcaraan Python untuk melaksanakan tugas pemprosesan bahasa tabii. TextBlob boleh digunakan untuk penandaan golongan kata, pengekstrakan frasa nama, tokenisasi dan analisis sentimen Hazarika et al. (2020). Polariti bagi TextBlob adalah di antara [-1,1], -1 menunjukkan sentimen negatif dan +1 menunjukkan sentimen positif. SentiWordNet berasal dari WordNet di mana setiap istilah dikaitkan dengan skor berangka yang menunjukkan maklumat sentimen positif dan negatif. Keputusan yang diperoleh dengan SentiWordNet adalah sama dengan pendekatan menggunakan leksikon manual Ohana & Tierney (2009).VADER juga seperti TextBlob yang merupakan modul NLTK yang menyediakan skor sentimen berdasarkan perkataan yang digunakan. Kelebihan yang ada pada VADER adalah boleh memahami dengan baik sentimen teks yang mengandungi emotikon, slanga, kata hubung, huruf besar, tanda baca dan banyak lagi Singh (2020).

Terdapat dua pendekatan dalam Pendekatan Berasaskan Leksikon iaitu pendekatan berasaskan kamus dan pendekatan berasaskan korpus. Pendekatan berasaskan kamus terkandung set kecil perkataan dasar dan menggunakan kamus sentimen seperti WordNet untuk mengembangkan perkataan melalui sinonim dan antonimnya Liapakis et al. (2020) dengan andaian sinonim mempunyai polariti yang sama dengan kata dasar manakala antonim berlawanan polariti Wankhade et al. (2022). Perkataan baharu yang diperolehi daripada proses pengembangan ini akan dimasukkan ke dalam senarai set dan proses ini akan berulang sehingga kesemua perkataan baharu ditemui daripada ulasan tersebut Aye & Aung (2018).

Manakala pendekatan berasaskan korpus menggunakan data korpus iaitu set rangkai kata sentimen yang telah ditetapkan dan seterusnya boleh dibahagikan kepada pendekatan statistik dan semantik Sadia et al. (2018); Wankhade et al. (2022).

Pendekatan berasaskan korpus digunakan untuk mengatasi isu kata dasar yang mungkin menunjukkan perkataan positif atau tidak bagi memastikan penggunaan perkataan sentimen dengan domain yang berbeza adalah cekap dan berkesan.

Walaupun pendekatan ini mempunyai kelemahan iaitu kebergantungan kepada domain di mana sesuatu perkataan boleh juga mempunyai pelbagai takrifan mengikut perspektif dan penggunaan perkataan tersebut. Namun dalam kajian ini, pendekatan ini menjadi pilihan kerana bersesuaian dengan volum data kerana ia tidak memerlukan latihan sekiranya terdapat penambahan lesikon. Seperti yang kita sedia maklum, teks ulasan adalah teks yang dinamik dan kadar evolusi sangat tinggi berbanding kadar evolusi teks formal (seperti teks berita).

2.3 PEMBINAAN LEKSIKON DALAM DOMAIN RESTORAN DAN MAKANAN

Analisis sentimen dalam industri restoran dan makanan amat bermanfaat untuk menarik pelanggan baharu untuk mencuba perkhidmatan dan produk yang ditawarkan di sesebuah restoran atau sebaliknya. Maklum balas memberangsangkan yang diperolehi juga dapat menaikkan kadar kepopularan sesebuah premis makanan tersebut serta meningkatkan jumlah keuntungan perniagaan.

Terdapat beberapa kajian yang telah dijalankan berkaitan analisis sentimen dalam industri restoran dan makanan. Dalam kajian Liapakis et al. (2020) telah menguji 2,000 ulasan pelanggan dalam Bahasa Greek. Penilaian adalah berdasarkan beberapa fungsi yang telah dikategorikan seperti kualiti (DoQI), perkhidmatan (DoS), harga (DoP), kuantiti (DoQn) dan imej (DoI) syarikat tempatan. Untuk pembinaan leksikon, kaedah beg kata-kata (BOW) telah digunakan sebagai perwakilan pembolehubah berasingan yang mempunyai berat berangka dengan kepentingan yang berbeza-beza. TF-IDF digunakan untuk mengira berat angka setiap perkataan. Analisis data menunjukkan bahawa kualiti makanan, perkhidmatan pelanggan dan imej sesebuah syarikat merupakan faktor yang menjadi keutamaan pelanggan berbanding harga dan kuantiti makanan. Dengan menggunakan kaedah pendekatan berasaskan leksikon ketepatan yang diperolehi adalah amat tinggi iaitu 99.35%. Pemilihan pendekatan berasaskan leksikon adalah bersesuaian dalam konteks kajian ini bagi Liapakis et al. (2020) kerana ia dapat menentukan polariti, kekuatan orientasi dan memberikan hasil

yang lebih baik berbanding pendekatan pembelajaran mesin sekiranya berurusan dengan perkataan penafian.

Kajian oleh Aye & Aung (2017) adalah membina leksikon sentimen bagi industri makanan dan restoran dalam bahasa Myanmar. Pendekatan yang dijalankan juga menggunakan pendekatan berasaskan leksikon di mana data ulasan pelanggan restoran diperolehi daripada Facebook. Bagi membina leksikon sentimen dalam bahasa Myanmar ulasan dikelaskan dalam beberapa bahagian iaitu makanan dan rasa, tempat, harga, kakitangan, perkhidmatan dan penilaian umum. Penyelidik menggunakan kamus asas yang boleh didapati daripada leksikon Myanmar dan mengumpul kata-kata sentimen secara manual berdasarkan pengetahuan dan mempelbagaikan perkataan dengan mencari sinonim dan antonim. Dapatan kajian menunjukkan bahawa peraturan sentimen bebas konteks untuk Bahasa Myanmar yang dihasilkan mampu memberikan ketepatan yang jitu.

Seterusnya melalui kajian pada 2018, Aye & Aung (2018) memfokuskan kepada kata penguat dan perkataan objektif untuk menambah baik klasifikasi sentimen bagi teks yang tidak formal dalam Bahasa Myanmar. Aye & Aung (2018) menemui kata penguat dan perkataan objektif tidak diambil kira semasa proses klasifikasi walaupun ia sebenarnya amat berperanan untuk mengubah arah sesuatu polariti sentimen. Kata penguat berfungsi menguatkan maksud yang terkandung dalam frasa seperti sungguh sedap atau sedap sungguh. Dalam pembinaan leksikon sentimen penyelidik telah membangunkan Senti-Lexicon Myanmar untuk domain makanan dan restoran. Dalam kajian ini juga penyelidik menggunakan kamus asas yang boleh didapati daripada leksikon Myanmar dan mengumpul kata-kata sentimen secara manual berdasarkan pengetahuan dan mempelbagaikan perkataan dengan mencari sinonim dan antonim. Masih menggunakan pendekatan yang sama iaitu pendekatan berasaskan leksikon, keputusan ketepatan yang diperolehi adalah amat memberangsangkan selepas mengambil kira kata penguat dan perkataan objektif dalam eksperimen.

Alqadi et al. (2020) mengkaji kaitan pempengaruh media sosial dalam pemilihan restoran dalam kalangan penduduk di Riyadh, Arab Saudi serta faktor-faktor yang mempengaruhi seseorang dalam pemilihan restoran. Penyelidik menggunakan 799

perkataan leksikon sentimen yang berpolariti positif dan negatif yang dibangunkan oleh Alrumayyan et al. (2018). Hasil analisis statistik soal selidik, ulasan 30 restoran diekstrak dan dianalisa. Dapatan hasil soal selidik mendapati 62.1% responden memilih restoran berdasarkan cadangan pempengaruh Snapchat dan ini adalah selaras dengan hasil dapatan analisis sentimen. Kajian ini menggunakan pendekatan pembelajaran mesin bagi analisis sentimen ulasan Google Map dan pendekatan berasaskan leksikon.

Berdasarkan hasil kajian-kajian yang lepas, boleh disimpulkan bahawa kajian analisis sentimen dalam industri restoran dan makanan telah dijalankan oleh ramai penyelidik. Para penyelidik memilih menggunakan leksikon sentimen yang dibangunkan sendiri berbanding leksikon umum kerana faktor bahasa asing selain Bahasa Inggeris. Leksikon sentimen bagi bahasa asing seperti kajian di atas belum diterokai secara meluas serta ketiadaan pelbagai sumber dan alatan seperti penandaan golongan kata, perbendaharaan kata dan korpus beranotasi.

2.4 PEMBINAAN LEKSIKON DALAM DOMAIN LAIN

Analisis sentimen juga amat penting dalam lain-lain industri serta domain selain perniagaan restoran dan makanan. Kajian yang dilakukan oleh Rodzman et al. (2019) menggunakan pendekatan berasaskan leksikon untuk mengkaji domain khusus lagu, isu politik dan produk dalam Bahasa Melayu. Pendekatan berasaskan leksikon menggunakan model linguistik bagi proses mengklasifikasi sentimen dan teknik ini lebih baik untuk mendapatkan hasil klasifikasi yang optimum dan konsisten walaupun berbeza domain Rodzman et al. (2019). Penemuan dalam kajian ini dapat membuktikan pendekatan berasaskan leksikon berjaya mengklasifikasikan ulasan lebih baik daripada Naïve Bayes dengan hasil nilai purata kejituan 70% berbanding 50%.

Seterusnya dalam kajian sama ada penggunaan aplikasi terjemahan Google memberikan kesan ke atas analisis sentimen leksikon dalam Bahasa Melayu, Enjop et al. (2022) mencari perbezaan antara set data manual yang didapati daripada terjemahan secara manual dan set data automatik, ulasan dalam Bahasa Inggeris diterjemahkan kepada Bahasa Melayu menggunakan aplikasi terjemahan Google. Didapati analisis sentimen menggunakan set data manual masih dapat mengekalkan konteks sentimen dan set data automatik telah mengurangkan sentimen berkaitan gangguan spektrum

autism. Leksikon sentimen VADER diaplikasikan untuk mengklasifikasikan polariti sentimen positif, negatif dan neutral.

Menurut Yerpude et al. (2019) analisis sentimen berupaya mengasingkan ulasan produk dan berasaskan ciri produk tersebut untuk mengetahui keburukan atau kebaikan sesuatu produk walaupun sebagai contoh kamera telefon mudah alih mendapat ulasan yang positif namun ruang penyimpanan menerima maklum balas yang negatif. Kajian ini dilaksanakan dalam dua bahagian iaitu bahagian yang pertama analisis sentimen secara menyeluruh ke atas ulasan produk dan analisis sentimen berasaskan ciri produk. Bagi pembangunan leksikon sentimen, penyelidik menggunakan leksikon umum SentiWordNet untuk mendapatkan skor positif dan negatif bagi sesuatu perkataan.

Feng et al. (2018) membina leksikon sentimen pendekatan automatik iaitu khusus domain daripada ulasan pembelian atas talian kerana leksikon sentimen yang tersedia sangat general dan tidak sesuai. Mereka menggunakan kaedah timbal balik titik (PMI) untuk membina matriks sentimen primitif, frekuensi songsang jangka kekerapan dokumen (TF-IDF) untuk menyaring ciri-ciri tujuh jenis produk serta mengukur hubung kait antara perkataan sentimen A dan ciri produk B dengan maklumat timbal balik titik yang ditambah baik (EPMI). Pada setiap dimensi sentimen, setiap perkataan sentimen boleh mengambil nilai 0 atau 1, di mana 1 menunjukkan bahawa perkataan itu tergolong dalam kategori tertentu manakala 0 menunjukkan bahawa ia tidak tergolong dalam kategori tersebut. Leksikon sentimen yang dibina dinilai dengan membina tugas klasifikasi sentimen menggunakan beberapa ulasan produk yang ditulis dalam bahasa Cina dan Inggeris. Seterusnya membandingkan keputusan dengan dua model daripada pembelajaran mesin iaitu Naive Bayes dan SVM.

Nugroho (2021) mencadangkan kaedah VADER untuk pembangunan leksikon sentimen dan membuat analisis sentimen daripada data Twiter serta meramal keputusan pilihan raya presiden Amerika Syarikat bagi tahun 2020. Ulasan daripada Twitter boleh juga terkandung ulasan daripada bukan pengundi yang berdaftar dan untuk memastikan ulasan yang diperolehi adalah daripada penduduk tempatan pemetaan kawasan dibuat terlebih dahulu. Dapatan hasil ramalan adalah hampir tepat dengan keputusan pilihan raya iaitu kemenangan kepada Parti Demokratik dengan undian sebenar sebanyak 22

undian berbanding ramalan 24 undian. Keputusan ini menunjukkan kaedah VADER terbukti boleh djadikan model ramalan hampir tepat untuk pilihan raya bagi Amerika Syarikat.

Dalam kajian Karamollaoglu et al. (2018) dengan pendekatan berasaskan leksikon dan menggunakan kamus SentiWordNet untuk menentukan nilai sentimental perkataan berkaitan dalam Bahasa Turki. Perkataan yang mengandungi kata nafi dikelaskan sebagai negatif dan dikurangkan dengan menggunakan Zemberek API. Kadar kejayaan yang diperolehi dalam kajian ini adalah 80% ketepatan dan pemarkahan-f.

Nguyen et al. (2018) melabelkan ulasan yang berskala satu sehingga tiga sebagai "negatif" dan empat atau lima sebagai "positif". Seterusnya selepas mendapat skor TF-IDF, diuji dengan tiga kaedah pembelajaran mesin iaitu SVM, Regresi Logistik dan Peningkatan Kecerunan serta tiga leksikon sentimen umum iaitu Pattern, VADER dan SentiWordNet. Ketiga-tiga kaedah pembelajaran mesin mengatasi prestasi model pendekatan berasaskan leksikon.

Pamungkas & Putri (2017) menjalankan analisis sentimen berasaskan leksikon pada set data ulasan berbahasa Indonesia yang diperolehi daripada Google PlayStore dan Apple AppStore dengan menggunakan leksikon sentimen umum SentiWordNet. Antara langkah yang dijalankan adalah menterjemah data daripada Bahasa Indonesia ke dalam Bahasa Inggeris supaya boleh dianalisis menggunakan SentiWordNet. Proses terjemahan menggunakan alat yang dipanggil Penterjemah Bing yang dibangunkan oleh Microsoft. Seterusnya, pemilihan golongan kata dijalankan dengan menggunakan perpustakaan daripada Stanford POS Tagger. Untuk pengiraan skor sentimen, SentiWordNet digunakan sebagai leksikon untuk menetapkan skor sentimen bagi setiap perkataan dalam set data. Output yang diperolehi daripada proses ini ialah skor sentimen bagi setiap data pendapat.

Sumber digital dalam Bahasa Melayu adalah terbatas menurut Sharmini Alexander (2017), sehubungan itu pembangunan leksikon sentimen dalam Bahasa Melayu adalah amat relevan. Penyelidik memadankan WordNet Bahasa dengan

WordNet Bahasa Inggeris bagi mendapatkan ofset set perkataan awal. Antara langkah pembangunan leksikon yang dijalankan adalah pemilihan set perkataan awal, pemadanan set perkataan dengan WordNet Bahasa, penjanaan kata sinonim dan antonim di WordNet Bahasa Inggeris dan penterjemahan ke Bahasa Melayu dalam WordNet Bahasa. Keputusan pengujian mendapati algoritma yang dicadangkan dalam penjanaan leksikon sentimen dalam Bahasa Melayu berasaskan leksikon sentimen umum WordNet adalah terbukti berkesan dengan peratusan persetujuan sebanyak 86.58%.

Manakala dalam kajian Amira Sumitro et al. (2021) berkaitan analisis sentimen terhadap vaksin Covid-19 di Indonesia menggunakan leksikon sentimen VADER. Pendapat berkaitan vaksin dikelaskan kepada lima sentimen polariti iaitu sangat positif, positif, negatif, kurang negatif dan neutral. Skor polariti ditetapkan seperti berikut, skor kurang dari -0.5 adalah sentimen negatif, skor lebih dari -0.5 dan kurang dari 0 adalah sentimen kurang negatif, skor 0 adalah sentimen neutral, skor lebih dari 0 dan kurang dari 0.5 adalah sentimen sangat positif, skor lebih besar dari 0.5 adalah sentimen positif. Dapatan kajian mendapati sentimen masyarakat terhadap vaksin Covid-19 pada Twitter lebih cenderung kepada sentimen neutral dengan peratusan ketepatan sebanyak 44.36%.

Berdasarkan contoh-contoh kajian yang telah dihuraikan di atas, sudah terbukti terdapat banyak kajian analisis sentimen yang dijalankan berkaitan pelbagai industri menggunakan antara tiga pendekatan yang dirasakan bersesuaian dengan set data serta skop pengkajian. Dalam pemilihan leksikon sentimen kebanyakan penyelidik membangunkan leksikon sentimen yang dibangunkan sendiri dan terdapat juga yang menggunakan leksikon sentimen umum seperti VADER, WordNet dan SentiWordNet. Kebanyakan kajian dalam analisis sentimen ulasan produk yang telah dijalankan adalah dalam kelas binari iaitu diklasifikasikan kepada positif dan negatif berbanding pengelasan yang pelbagai. Jadual 2.1 menunjukkan ringkasan kajian-kajian yang telah dijalankan.

Jadual 2.1 Ringkasan Pembinaan Leksikon

No	Tajuk	Penerangan	Algoritma/Kaedah Pembinaan Leksikon	Sumber Data

1	A Sentiment Lexicon-based Analysis for Food and Beverage Industry Reviews. The Greek Language Paradigm
	
	Liapakis et al. (2020)
		Kajian ini menguji 2,000 ulasan pelanggan dalam Bahasa Greek. Penilaian adalah berdasarkan beberapa fungsi yang telah dikategorikan seperti kualiti (DoQI), perkhidmatan (DoS), harga (DoP), kuantiti (DoQn) dan imej (DoI) syarikat tempatan.	- Pendekatan berasaskan leksikon
			  - Kaedah berasaskan kamus
			  - Kaedah berasaskan korpus
			- BOW
			- TF-IDF
			- PMI
				8,950 ulasan pelanggan diekstrak daripada 690 syarikat yang dipilih secara rawak daripada 6,795 syarikat di Greece

2	Sentiment Analysis For Reviews of Restaurants in Myanmar Text
	
	Aye & Aung (2017)
		Kajian ini membina leksikon sentimen Myanmar untuk domain makanan dan restoran serta menganalisis ulasan teks pelanggan.	- Pendekatan berasaskan leksikon
			  - Kaedah berasaskan kamus
			  - Kaedah berasaskan korpus
				500 ulasan restoran daripada halaman Facebook dikumpulkan secara manual

3	Enhanced Sentiment Classification for Informal Myanmar Text of Restaurant Reviews Aye & Aung (2018)	Kajian ini bertujuan untuk mengatasi masalah khusus bahasa iaitu Bahasa Myanmar dalam meningkatkan klasifikasi sentimen untuk teks tidak formal. Memfokuskan kepada perkataan penguat dan objektif untuk meningkatkan klasifikasi sentimen untuk domain makanan dan restoran.	- Pendekatan berasaskan leksikon - Senti-Lexicon	1000 ulasan restoran daripada halaman Facebook dikumpulkan secara manual

b e r s a m b

...sambungan

4	How Social Media Influencers Affect Consumers' Selection: Statistical and Sentiment Analysis Alqadi et al. (2020)	Fokus dalam kajian ini adalah mengkaji kesan pempengaruh media sosial Snapchat ke atas pemilihan restoran di Riyadh dan mengenal pasti faktor yang dianggap paling penting oleh orang ramai dalam pemilihan restoran.	- Pendekatan berasaskan leksikon - SVM - Analisis statistik	- Soal selidik tentang pengaruh Snapchat dalam pemilihan restoran. - Ulasan orang ramai tentang restoran daripada Peta Google
5	Experiment with Lexicon Based Techniques on Domain-Specific Malay Document Sentiment Analysis Rodzman et al. (2019)	Kertas kerja ini bertujuan untuk menjalankan Analisis Sentimen ke atas dokumen bahasa Melayu dan mencadangkan teknik berasaskan Leksikon bagi set data domain khusus Lagu, Politik dan Produk. Seterusnya mencari pengelasan yang terbaik pada Dokumen Bahasa Melayu Khusus Domain.	- Pendekatan berasaskan leksikon - Naïve bayes	Jawapan yang diberikan oleh pelajar-pelajar.
6	Does Google Translate Affect Lexico,n-based Sentiment Analysis of Malay Social Media Text? Enjop et al. (2022)	Matlamat kajian ini adalah untuk menilai prestasi Terjemahan Mesin (MT) dalam Terjemahan Google terhadap analisis sentimen teks media sosial Melayu di laman Facebook seorang penjaga penghidap autisme.	VADER	3,525 komen Facebook dalam bahasa Melayu telah dikumpulkan dari Mei hingga Oktober 2020

bersamb

...sambungan

7	Sentiment Analysis on Product Features Based on Lexicon Approach Using Natural Language Processing Yerpude et al. (2019)	Kajian ini menjalankan analisis sentimen terhadap ulasan produk elektronik dan mengenai ciri produk yang terdapat dalam ulasan produk iaitu telefon mudah alih.	SentiWordNet	Set data ulasan pelanggan telefon mudah alih
8	US Presidential Election 2020 Prediction Based on Twitter Data Using Lexicon-Based Sentiment Analysis Nugroho (2021)	Penyelidik menjalankan kajian ke atas ulasan Twitter untuk meramal keputusan pilihan raya Presiden Amerika Syarikat bagi tahun 2020.	- Pendekatan berasaskan leksikon - VADER	Twitter
9	Automatic Approach of Sentiment Lexicon Generation for Mobile Shopping Reviews Feng et al. (2018)	Penyelidik mencadangkan pendekatan automatik untuk membina leksikon sentimen khusus domain dengan mempertimbangkan hubungan antara perkataan sentimen dan ciri produk dalam ulasan beli-belah mudah alih.	- PoS - BOW - PMI - TF-IDF - EPMI - Naive Bayes - SVM	- Menggunakan dua set data ulasan membeli belah iaitu dalam Bahasa Cina dan Bahasa Inggeris bagi tujuh jenis produk - Set data berbahasa Cina diperolehi daripada Dr. Tan(http://download.csdn.net/download/lssc4205/9903298) dan aplikasi membeli belah JD (https://www.jd.com/). - Set data berbahasa Inggeris diperolehi daripada Amazon

b e r s a m b

...sambungan

10	Sentiment Analysis on Turkish Social Media Shares through Lexicon Based Approach Karamollaoglu et al. (2018)	Matlamat kajian ini adalah untuk mengkaji analisis sentimen dalam Bahasa Turki dengan menggunakan kamus SentiWordNet untuk menentukan sentimen perkataan dalam Bahasa Turki. Perkataan yang mengandungi kata nafi dikelaskan sebagai negatif dan dikurangkan dengan menggunakan Zemberek API	- Pendekatan berasaskan leksikon - SentiWordNet - Ketepatan, Precision, Recall, Certainty dan pemarkahan-F	Mengandungi 300 ulasan positif dan 300 ulasan negatif yang diperolehi daripada Twitter
11	Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches Nguyen et al. (2018)	Kajian ini mencari perbandingan antara pendekatan berasaskan pembelajaran mesin dan pendekatan berasaskan leksikon ke atas ulasan produk. Enam kaedah klasifikasi sentimen yang berbeza digunakan iaitu tiga pendekatan pembelajaran mesin yang diselia dan tiga teknik berasaskan leksikon.	- Regresi Logistik - SVM - Gradient Boosting - VADER - Pattern - SentiWordNet	Ulasan pelanggan daripada Amazon
12	Sentiment Analysis of Online Customer Reviews for Hotel Industry: An Appraisal of Hybrid Approach Raheem & Lai (2020)	Fokus utama kajian ini adalah menganalisis sentimen ulasan hotel dengan menggunakan kaedah hibrid. Kajian ini memberi tumpuan kepada ulasan dalam talian yang dikongsi oleh pelanggan hotel di mana ia menggambarkan pengalaman dan kepuasan melalui sentimen yang bersifat positif, negatif atau neutral.	Pendekatan hibrid	Ulasan pelanggan hotel daripada laman web hotel

bersambung...

...sambungan

13	An Experimental Study of Lexicon-based Sentiment Analysis on Bahasa Indonesia Pamungkas & Putri (2017)	Kajian ini menjalankan analisis sentimen berasaskan leksikon dengan menggunakan leksikon sentimen SentiWordNet. Teks ulasan terlebih dahulu diterjemah ke dalam Bahasa Indonesia supaya boleh dianalisa. Stanford POS Tagger dipilih sebagai penandaan golongan kata. SentiWordNet digunakan sebagai skor sentimen untuk setiap perkataan.	- Pendekatan berasaskan leksikon - SentiWordNet	553 data ulasan pengguna daripada PlayStore dan AppStore
14	Generating a Malay Sentiment Lexicon Based on Wordnet Sharmini Alexander (2017)	Penyelidik menjalankan kajian berkaitan penjanaan leksikon sentimen dalam Bahasa Melayu dengan menggunakan leksikon sentimen umum WordNet Bahasa. Keputusan pengujian mendapati algoritma yang dicadangkan dalam penjanaan leksikon sentimen dalam Bahasa Melayu berasaskan leksikon sentimen umum WordNet adalah terbukti berkesan dengan peratusan persetujuan sebanyak 86.58%	- Pendekatan berasaskan leksikon - WordNet	
15	Analisis Sentimen Terhadap Vaksin Covid-19 di Indonesia pada Twitter Menggunakan Metode Lexicon Based Amira Sumitro et al. (2021)	Kajian ini adalah berkaitan analisis sentimen masyarakat di Indonesia terhadap vaksin Covid-19. Ulasan dalam Bahasa Indonesia terlebih dahulu diterjemahkan kepada Bahasa Inggeris dengan aplikasi Bing Translator supaya dapat diproses dengan menggunakan alatan VADER. Ulasan dibahagikan kepada lima polariti iaitu positif, sangat positif, neutral, negatif dan kurang negatif. Dapatan kajian mendapati sentimen masyarakat terhadap vaksin Covid-19 pada Twitter lebih cenderung kepada sentimen neutral dengan peratusan ketepatan sebanyak 44.36%.	- Pendekatan berasaskan leksikon - VADER	-Twitter

2.5 LATAR BELAKANG TEKNIKAL

Bahagian ini akan menerangkan tentang latar belakang teknikal seperti pengumpulan data, golongan kata dan penentuan skor sentimen.

2.5.1 Pengumpulan Data

Analisis sentimen memerlukan set data untuk menjalankan sesuatu eksperimen seperti latihan, pengujian dan menilai sesuatu model analisis sentimen yang ingin dibangunkan. Terdapat pelbagai sumber yang tersedia di alam maya yang boleh digunakan dan terbukti sebagai sumber yang boleh dipercayai dan berharga. Dalam kajian ini set data ulasan restoran diperolehi daripada *Kaggle Dataset Repository*.

Kaggle merupakan repositori set data yang merangkumi pelbagai topik. Ia juga sebagai platform komuniti dalam talian untuk saintis data, penyelidik dan sesiapa jua yang berminat meneroka bidang pembelajaran mesin (‘ Wh a t i s K a g g l e ? | n.d.). Pengguna boleh memuat naik, berkongsi dan meneroka set data yang boleh digunakan sebagai bahan penyelidikan atau analisis. Kaggle menyediakan platform untuk pertandingan di mana ia menjadikan Kaggle sebagai sumber popular. Dalam pertandingan sains data ini penyertaan syarikat atau organisasi sebagai penaja dengan memberikan ganjaran yang lumayan telah menarik lebih ramai saintis data sama ada yang Baharu mahupun yang berpengalaman untuk turut serta dan bersaing antara satu sama lain.

2.5.2 Penandaan Golongan Kata (POS)

Penandaan golongan kata atau POS adalah proses untuk mengekstrak lebih banyak maklumat daripada teks asal di mana setiap perkataan akan dilabelkan mengikut kumpulan tatabahasa seperti kata nama (noun), kata kerja (verb), kata sifat (adjective), kata keterangan (adverb) dan lain-lain. Perkataan yang mengandungi sentimen ialah perkataan yang menerangkan huraian pengguna tentang produk. Terdapat kajian yang menyatakan kata sifat dan kata keterangan mengandungi kata-kata sentimen Benamara et al. (2007). Selain kata sifat dan kata nama, kata kerja dan kata keterangan juga mengandungi kata-kata sentimen dan memberi impak yang besar terhadap sentimen Rajeswari et al. (2020).

2.5.3 Penentuan Skor Sentimen

Penentuan skor sentimen merupakan proses mendapatkan maklumat yang bernilai dan akan memberi kesan kepada prestasi sesuatu eksperimen. Pemilihan set skor penting untuk mengubah maklumat kualitatif iaitu sentimen dalam teks kepada data kuantitatif. Terdapat banyak teknik untuk menentukan skor sentimen antaranya ialah TF-IDF, Word2vec, Doc2vec, PMI, BOW dan sebagainya.

a. TF-IDF

TF-IDF atau frekuensi songsang jangka kekerapan dokumen merupakan kaedah statistik untuk menilai kepentingan sesuatu perkataan berbanding teks tertentu dalam keseluruhan korpus Bijoyan & Sarit (2018). Kekerapan meningkat apabila sesuatu perkataan telah berulang beberapa kali. Perkataan dengan skor yang lebih tinggi dianggap lebih penting. TF-IDF dikira berdasarkan formula 2.1 seperti berikut:

…(2.1)

TF(*t, d*), adalah kekerapan istilah di mana merujuk kepada nisbah kekerapan sebutan (t) muncul dalam dokumen (d). Ia mengukur kekerapan istilah itu berlaku dalam dokumen. IDF(*t*), adalah kekerapan dokumen songsang yang mengukur kepentingan istilah. Ia dikira berdasarkan pengiraan 2.2 seperti berikut:

…(2.2)

N, merujuk kepada jumlah bilangan dokumen di dalam korpus manakala (n_t) merupakan bilangan dokumen yang mengandungi istilah (t). Penyebut "+1" digunakan untuk mengelakkan pembahagian dengan sifar.

b. PMI

PMI atau maklumat timbal balik titik berperanan untuk menentukan kekuatan jumlah maklumat yang dikongsi antara dua istilah dalam korpus. Ia menerangkan kebarangkalian dua istilah berkait antara

mempunyai maksud yang khusus apabila dua perkataan ini digunakan bersama namun sekiranya digunakan secara berasingan akan membawa maksud dan konteks yang berbeza. PMI dikira berdasarkan formula 2.3 seperti berikut:

…(2.3)

P(perkataan$_1$, perkataan$_2$), adalah kebarangkalian perkataan$_1$ dan perkataan$_2$ dalam teks yang sama manakala *P(perkataan$_1$)* dan *P(perkataan$_2$)* mewakili kebarangkalian kejadian perkataan$_1$ dan perkataan$_2$ dalam korpus. Sekiranya jawapan yang diterima bersamaan 0, ia bermaksud dua perkataan tersebut tiada perkaitan antara satu sama lain dan tiada maksud khusus. Semakin tinggi jawapan yang diterima maka semakin tinggi perkaitan antara dua perkataan dan sebaliknya.

2.6 PENUTUP

Bab ini menerangkan Kajian Kesusateraan daripada penyelidikan yang pernah dijalankan. Di dalam analisis sentimen terdapat tiga pendekatan yang biasa digunakan oleh para penyelidik iaitu pendekatan berasaskan leksikon, pendekatan berasaskan pembelajaran mesin dan pendekatan hibrid. Dalam kajian pembinaan lesikon dalam industri makanan dan restoran, para penyelidik menggunakan leksikon sentimen yang dibangunkan sendiri berbanding leksikon sentimen yang telah tersedia. Dalam kajian pembinaan leksikon dalam domain lain pula selain membangunkan lesksikon sentimen terdapat juga penyelidik menggunakan leksikon sentimen yang umum seperti VADER, SentiWordNet dan WordNet.

BAB III

METODOLOGI KAJIAN

3.1 PENGENALAN

Bahagian ini akan menerangkan kaedah dan rangka kerja penyelidikan yang akan dilaksanakan bagi mencapai matlamat kajian. Teknik analisis sentimen secara umumnya yang selalu digunakan adalah proses yang merangkumi pengumpulan data, pembersihan data, pemilihan ciri dan menganalisa sentimen.

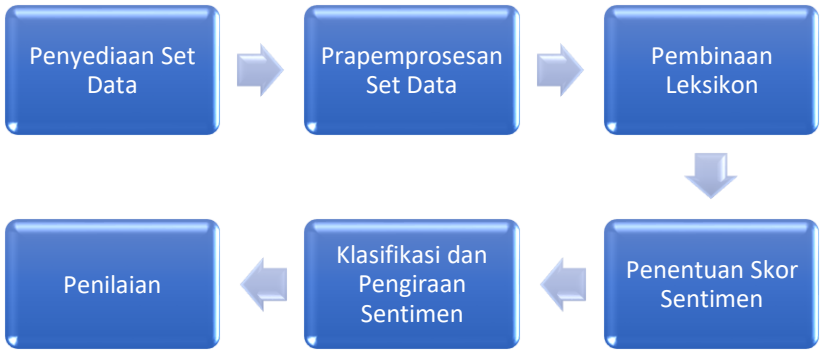
3.2 ALATAN EKSPERIMEN

Alatan dan aplikasi yang digunakan dalam kajian ini adalah Google Colab yang menggunakan bahasa pengaturcaraan Python. Google Colab merupakan aplikasi dalam talian yang disediakan secara percuma kepada pelajar, saintis data atau penyelidik pembelajaran mesin. Ia membolehkan pengguna menulis dan melaksanakan Python dengan menggunakan pelayar sedia ada. Selain itu, ia tidak memerlukan sebarang konfigurasi dan mudah untuk perkongsian file.

Manakala bahasa pengaturcaraan Python merupakan salah satu bahasa pengaturcaraan aras tinggi, sumber terbuka dan berorientasikan objek yang dapat beroperasi dalam hampir semua platform. Rangka kerja pembelajaran mesin atau *Library* seperti SciKitLearn, Pandas, Matplotlib dan sebagainya telah menjadikan tugas analisis sentimen lebih cekap dan pantas yang amat diperlukan untuk menjalankan kajian ini.

3.3 METODOLOGI KAJIAN

Metodologi kajian ini terdiri daripada enam (6) fasa. Fasa pertama ialah penyediaan set data, fasa kedua proses prapemprosesan set data, seterusnya pembinaan leksikon yang melibatkan proses penandaan POS dilaksanakan pada fasa ketiga, pada fasa keempat penentuan skor sentimen, manakala fasa seterusnya adalah klasifikasi sentimen serta pengiraan sentimen dan akhirnya penilaian dibuat pada fasa terakhir. Aliran kerja bagi metodologi kajian ini diringkaskan seperti Rajah 3.2 berikut:



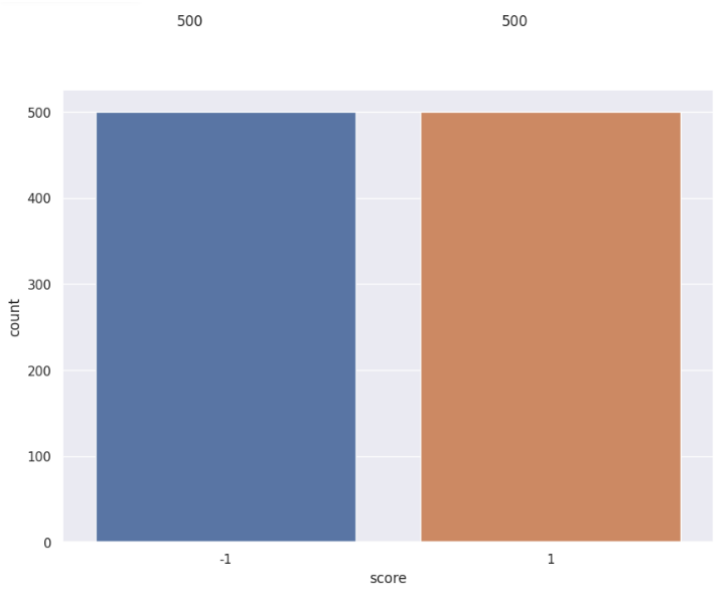
Rajah 3.1 Aliran Kerja Metodologi Kajian

3.4 FASA 1 : PENYEDIAAN SET DATA

Analisis sentimen memerlukan set data untuk menjalankan sesuatu eksperimen seperti latihan, pengujian dan menilai sesuatu model analisis sentimen yang ingin dibangunkan. Set data ulasan restoran diperoleh dari *Kaggle Dataset Respository*, https://www.kaggle.com/datasets/anuragmishra2311/restaurant-reviews dan mengandungi 1,000 rekod dalam format CSV. Set data ini terdiri daripada ulasan restoran berserta label 0 yang menunjukkan bahawa pelanggan tidak menyukai perkhidmatan Restoran atau 1 menunjukkan bahawa pelanggan menyukai perkhidmatan yang disediakan oleh Restoran. Kesemua set data ini digunakan sepenuhnya bagi melaksanakan kajian ini. Dalam set data ini, label menyukai dan tidak menyukai iaitu atribut 'Liked' adalah seimbang iaitu 500 ulasan positif dan 500 negatif. Jadual 3.1 memaparkan senarai atribut, keterangan dan jenis set data. Manakala Rajah 3.3 menunjukkan plot bar bagi set data ini.

Jadual 3.1 Senarai Atribut dengan Keterangan & Jenis Set Data Ulasan Restoran

Atribut	**Keterangan**	**Jenis**
Review	Ulasan pelanggan	Nominal
Liked	0 - menunjukkan bahawa pelanggan tidak menyukai perkhidmatan Restoran 1- menunjukkan bahawa pelanggan menyukai perkhidmatan yang disediakan oleh Restoran.	Kategori



Rajah 3.2 Graf Plot Set Data Ulasan Restoran

Sebelum melakukan sebarang analisis, data perlu diekstrak. Set data ulasan restoran yang tersedia dalam format CSV dibaca dengan menggunakan skrip Python melalui aplikasi dalam talian Google Colab. Kod aturcara Python bermula dengan memanggil *library* yang berkaitan seperti NLTK dan Pandas. Fail CSV akan dibaca menggunakan kod *library* pandas. Arahan Pandas akan memasukkan data untuk dibaca dan ditulis data ke dalam atau dari sesebuah file. Rajah 3.4 menunjukkan pengekodan untuk membaca set data.

```
import pandas as pd
import csv

# Load your CSV file into a DataFrame
df = pd.read_csv('restaurant_reviews_ori.csv')
df
```

Rajah 3.3 Import Library dan Baca Fail CSV

3.5 FASA 2 : PRAPEMPROSESAN SET DATA

Teks ulasan yang diperolehi selalunya mengandungi banyak data hingar dan data yang tidak diperlukan. Untuk mendapatkan hasil ketepatan yang tinggi set data terlebih dahulu hendaklah dimurnikan dengan langkah-langkah prapemprosesan. Dengan prapemprosesan, langkah klasifikasi dapat dipercepatkan dan membantu dalam menganalisis sentimen secara masa nyata. Prapemprosesan set data dilaksanakan bagi memastikan data tiada nilai yang tidak lengkap, diolah dan dibersihkan untuk meningkatkan kualiti data. Dalam konteks analisis sentimen, data hendaklah dibersihkan dan ditransformasikan kepada perkara berikut:

1. Menyeragamkan teks kepada huruf kecil;
2. Tokenisasi teks iaitu membahagikan teks kepada satu perkataan di mana teks diubah kepada urutan ayat dan kemudian ayat tersebut diubah kepada turutan perkataan;
3. Membuang semua tanda baca seperti tanda !@#%;
4. Membuang perkataan yang tidak mempunyai maksud dan tidak berguna iaitu kata henti seperti 'a', 'the', 'is' dan 'are', dan
5. Menjadikan semua perkataan kepada kata akar seperti '*likes*' menjadi '*like*'.

Prapemprosesan set data sangat penting kerana tanpa membersih, mengubah dan menyusun data sebelum melakukan analisis, hasil kajian tidak akan diperolehi seperti mana yang diharapkan dan ketepatannya boleh diragui. Rajah 3.5 menunjukkan skrip bagi prapemprosesan yang telah dilaksanakan seperti berikut:

```
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize, RegexpTokenizer
from nltk import pos_tag
import spacy
nlp = spacy.load('en_core_web_sm')

nltk.download('punkt')
nltk.download('stopwords')
nltk.download('averaged_perceptron_tagger')

# Define a function to preprocess text
def preprocess_text(text):
    # Remove numbers and convert to lowercase
    text = ''.join([char for char in text if not char.isdigit()])
    text = text.lower()

    # Tokenize the text
    tokenizer = RegexpTokenizer(r'\w+')
    tokens = tokenizer.tokenize(text)

    # Remove stopwords
    stop_words = set(stopwords.words('english'))
    tokens = [token for token in tokens if token not in stop_words]
```

Rajah 3.4 Pembersihan Data

3.6 FASA 3 : PEMBINAAN LEKSIKON

Terdapat dua aktiviti utama dalam proses pembinaan leksikon sentimen secara automatik iaitu:

1. Pemilihan set golongan kata; dan
2. Pembahagian ulasan positif dan negatif

3.6.1 Pemilihan Set Golongan Kata

Pemilihan set golongan kata dilakukan dengan mengenal pasti set golongan kata atau Part of Speech (POS) dalam Bahasa Inggeris untuk mendedahkan pengertian sesuatu perkataan dan perkataan seterusnya. Leksikon POS dibangunkan dengan menggunakan librari Spacy. Terdapat lapan POS yang terdiri daripada kata nama, kata kerja, kata ganti nama, kata keterangan, kata sendi, preposisi, *participle* dan artikel (Jurafsky & Martin

2023). Dalam kajian ini golongan kata terbahagi kepada dua set iaitu Set POS-1 yang terdiri daripada kata nama, kata kerja, kata sifat dan kata keterangan. Manakala Set POS-2 terdiri daripada kata sifat. Dua set POS ini dibangunkan untuk perbandingan seperti Rajah 3.6 dan Rajah 3.7. Pemilihan set POS ini dibuat bertepatan dengan kajian Benamara et al. (2007) dan Rajeswari et al. (2020)yang mengatakan kata sifat, kata nama, kata kerja dan kata keterangan mengandungi kata-kata sentimen dan memberi impak yang besar terhadap sentimen.

```
# Extract adverb, nouns, verb and adjectives
filtered_tokens = [word for word, pos in tagged_tokens if pos in ['NN', 'NNS', 'JJ', 'JJR', 'JJS','RB','RBR','RBS','VB']]
```

Rajah 3.5 Set POS-1

```
# Extract adjectives
filtered_tokens = [word for word, pos in tagged_tokens if pos in ['JJ', 'JJR', 'JJS']]
```

Rajah 3.6 Set POS-2

Dalam kajian ini juga Penn Treebank digunakan sebagai set tag POS dan ia merupakan penandaan POS dalam Bahasa Inggeris yang sering digunakan dalam kebanyakan kajian. Penn Treebank terdiri daripada kira-kira tujuh juta perkataan teks bertag dan mempunyai 48 set tag POS (Taylor et al. 2003). Jadual 3.2 menunjukkan senarai set tag POS bagi Penn Treebank. Manakala Jadual 3.3 menyenaraikan 10 contoh POS yang terdapat dalam kandungan teks.

Jadual 3.2 Set Tag POS Penn Treebank

Tag	**Keterangan**	**Tag**	**Keterangan**
CC	Coordinating conjunction	TO	Infinitival *to*
CD	Cardinal number	UH	Interjection
DT	Determiner	VB	Verb, base form
EX	Existential there	VBD	Verb, past tense
FW	Foreign word	VBG	Verb, gerund/present tense
IN	Preposition	VBN	Verb, past particle
JJ	Adjective	VBP	Verb, non-3rd ps. sg. present
JJR	Adjective, comparative	VBZ	Verb, 3rd ps. sg. present
JJS	Adjective, superlative	WDT	Wh-determiner
LS	List item marker	WP	Wh-pronoun
MD	Modal	WP$	Wh-possess
NN	Noun, singular or mass	WRB	Wh-adverb

b e r s a m b

…s a m b u n g a n

NNS	Noun, plural	#	Pound sign
NNP	Proper noun, singular	$	Dollar sign
NNPS	Proper noun, plural	.	Sentence-final punctuation
PDT	Predeterminer	,	Comma
POS	Possessive ending	:	Colon, semi-colon
PRP	Personal pronoun	(	Left bracket character
PP$	Possessive pronoun	)	Right bracket character
RB	Adverb	″	Straight double quote
RBR	Adverb, comparative	‘	Left open single quote
RBS	Adverb, superlative	“	Left open double quote
RP	Particle	,	Right close single quote
SYM	Symbol	”	Right close double quote

Jadual 3.3 Contoh POS dalam Kandungan Teks

Perkataan	**Tag**
wow	UH
place	NN
crust	NN
good	JJ
tasty	JJ
texture	NN
nasty	JJ
late	NNP
bank	NNP
holiday	NNP

3.6.2 Pembahagian Ulasan Positif dan Negatif

Aktiviti yang kedua dalam pembinaan leksikon sentimen adalah pembahagian ulasan positif dan negatif. Nilai pemberat -1 diberikan untuk ulasan negatif dan +1 untuk ulasan positif. Ulasan positif dan negatif berdasarkan atribut *Liked* iaitu 0 sebagai ulasan negatif dan *Liked* bernilai 1 sebagai ulasan positif. Seterusnya set golongan kata dibahagikan kepada dua set iaitu Set POS-1 yang terdiri daripada kata nama, kata kerja, kata sifat dan kata keterangan serta Set POS-2 terdiri daripada kata sifat.

3.7 FASA 4 : PENENTUAN SKOR SENTIMEN

Penentuan skor sentimen perlu dilaksanakan untuk memberikan nilai skor sentimen untuk leksikon sentimen. Dalam kajian ini terdapat tiga teknik yang digunakan iaitu skor 1 dan 0, TF-IDF dan PMI telah digunakan dalam set data ulasan restoran.

3.7.1 Skor 1 dan 0

Nilai skor yang diberikan ini adalah nilai yang mudah diberikan untuk pengiraan sesuatu skor sentimen sama ada membawa nilai positif mahupun negatif. Sekiranya jumlah skor adalah lebih besar daripada nilai 0 maka polariti sentimen adalah positif manakala sekiranya kurang daripada 0 polariti sentimen adalah negatif.

3.7.2 TF-IDF

TF-IDF ialah teknik untuk mengira berat angka atau skor setiap perkataan. Semakin besar nilai TF-IDF dalam perkataan artikel semakin penting perkataan tersebut dalam dokumen. TF-IDF digunakan dalam klasifikasi teks dan pemodelan topik. TF-IDF juga boleh digunakan dalam banyak aplikasi pemprosesan bahasa semula jadi. Sebagai contoh, aplikasi carian internet menggunakan TF-IDF untuk menilai perkaitan dokumen untuk pertanyaan. Penggunaan TF-IDF sebagai skor sentimen adalah wajar kerana ia mengambil kira kekerapan istilah atau frasa penting dalam teks tertentu untuk menentukan keseluruhan sentimen. Dengan mengira skor TF-IDF untuk setiap perkataan atau frasa, pemberat yang lebih tinggi dapat ditetapkan kepada frasa di mana ia lebih bermaklumat dan menunjukkan sentimen. Addiga & Bagui (2022) menunjukkan pengiraan skor TF-IDF dalam sesuatu korpus membolehkan perwakilan yang lebih tepat bagi keseluruhan polariti kerana perkataan diberikan pemberat berdasarkan keunikan dan bukannya sekadar kekerapan ia muncul dalam korpus.

Kaedah ini adalah gabungan antara TF dan IDF. Berat angka atau skor tersebut adalah hasil daripada dua faktor, yang pertama adalah pengiraan kekerapan istilah (*tf*) yang dinormalkan sebagai contoh bilangan kali perkataan muncul dalam dokumen, dibahagikan dengan jumlah bilangan perkataan dalam dokumen itu. Istilah kedua ialah kekerapan dokumen songsang (*idf*), yang dikira sebagai logaritma bilangan dokumen dalam koleksi atau korpus dibahagikan dengan bilangan dokumen di mana perkataan

tertentu muncul. Perkataan boleh disusun mengikut turutan dari terbesar ke terkecil. TF-IDF bagi istilah *t* dalam dokumen *d* dikira berdasarkan formula 3.1 seperti berikut:

4 &ط) $ &ðĎΩ سي 4 &ðĎΩ z ì ï ...(3.1)

TF(*t*,*d*) = bilangan kejadian *t* dalam *d*

N = jumlah bilangan dokumen di dalam korpus

(n_t) = bilangan dokumen yang mengandungi istilah (*t*)

Berikut merupakan contoh daripada eksperimen mengunakan teks yang telah dilaksanakan prapemprosesan dan penandaan PoS yang terpilih. Pertimbangkan lima dokumen iaitu D1, D2, D3, D4 dan D5 yang terdiri daripada ayat seperti Jadual 3.4 berikut:

Jadual 3.4 Contoh Dokumen Teks

Dokumen	**Teks Sebelum Prapemprosesan**	**Teks Selepas Prapemprosesan**
D1	Crust is not good	crust good
D2	Not tasty and the texture was just nasty.	tasty texture nasty
D3	The fries were great too	fries great
D4	A great touch	great touch
D5	Now I am getting angry and I want my damn pho	angry damn pho

Untuk langkah pertama, teks seperti yang disenaraikan seperti jadual di atas akan disegmenkan untuk mendapatkan pengiraan yang terdapat dalam frasa yang sepadan dan sebaliknya. Jadual 3.5 menunjukkan hasil segmentasi teks.

Jadual 3.5 Segmentasi Teks

Perkataan	**D1**	**D2**	**D3**	**D4**	**D5**
crust	1	0	0	0	0
good	1	0	0	0	0
tasty	0	1	0	0	0
texture	0	1	0	0	0

bersamb

…s a m b u n g a n					
nasty	0	1	0	0	0
fries	0	0	1	0	0
great	0	0	1	1	0
touch	0	0	0	1	0
angry	1	0	0	0	0
damn	1	0	0	0	0
pho	1	0	0	0	0

Langkah kedua melibatkan pengiraan IDF di mana setiap perkataan yang sepadan dengan lima dokumen yang ditentukan akan dikira. TF dan IDF bagi setiap perkataan boleh ditentukan berdasarkan formula 3.1. Keputusan yang diperolehi seperti Jadual 3.6 berikut:

Jadual 3.6 Pengiraan TF-IDF

Perkataan	**TF**					**IDF**	**TF*IDF**				
	D1	**D2**	**D3**	**D4**	**D5**		**D1**	**D2**	**D3**	**D4**	**D5**
crust	1/2	0	0	0	0	log(5/1) = 0.699	0.35	0	0	0	0
good	1/2	0	0	0	0	log(5/1) = 0.699	0.35	0	0	0	0
tasty	0	1/3	0	0	0	log(5/1) = 0.699	0	0.233	0	0	0
texture	0	1/3	0	0	0	log(5/1) = 0.699	0	0.233	0	0	0
nasty	0	1/3	0	0	0	log(5/1) = 0.699	0	0.233	0	0	0
fries	0	0	1/2	0	0	log(5/1) = 0.699	0	0	0.35	0	0
great	0	0	1/2	1/2	0	log(5/2) = 0.398	0	0	0.199	0.199	0
touch	0	0	0	1/2	0	log(5/1) = 0.699	0	0	0	0.35	0
angry	1/3	0	0	0	0	log(5/1) = 0.699	0.233	0	0	0	0
damn	1/3	0	0	0	0	log(5/1) = 0.699	0.233	0	0	0	0
pho	1/3	0	0	0	0	log(5/1) = 0.699	0.233	0	0	0	0

3.7.3 PMI

Maklumat timbal balik titik atau PMI merupakan ukuran statistik untuk mengukur kekuatan perkaitan antara dua perkataan dalam kandungan teks. PMI menggunakan pengiraan tertentu untuk menghasilkan dan mengklasifikasikan ulasan positif atau negatif. Dengan pengiraan perkataan yang dikaitkan dengan sentimen positif atau negatif dalam konteks ciri yang terdapat dalam ulasan, cerapan berharga tentang sentimen pengguna terhadap ciri tersebut boleh diperolehi. Dalam percubaan perbandingan leksikon oleh (Fredriksen et al. 2018) leksikon yang dicipta

menggunakan PMI sebagai skor mengalahkan leksikon beranotasi secara manual iaitu AFINN dan Sentiment140. Pemilihan PMI sebagai skor sentimen membuktikan bahawa mencipta leksikon sentimen secara automatik ialah kaedah penciptaan leksikon sentimen yang sangat berdaya maju. PMI dikira berdasarkan formula 3.2 seperti berikut:

$$\text{PMI}(w,\text{sentimen}) = \log_2 \frac{p(w,\text{sentimen})}{p(w)\,p(\text{sentimen})} \quad \ldots(3.2)$$

Berikut merupakan contoh teks yang terdapat dalam set data kajian ini. Teks ini juga diambil setelah dilaksanakan prapemprosesan dan penandaan POS yang terpilih. Pertimbangkan ayat seperti Jadual 3.7 di bawah untuk pengiraan PMI:

Jadual 3.7 Contoh Teks

Dokumen	**Teks Sebelum Prapemprosesan**	**Teks Selepas Prapemprosesan**
D	Took an hour to get our food only 4 tables in restaurant my food was Luke warm, Our sever was running around like he was totally overwhelmed.	hour get food tables food luke warm sever around totally overwhelmed

Berdasarkan teks yang terdapat dalam dokumen di atas mulakan langkah yang pertama iaitu mengasingkan teks kepada setiap perkataan seperti berikut:

“hour” “get” “food” “tables” “food” “luke” “ w a r m ” “ s e v e r ” “ a r o
“ o v e r w h e l m e d ”

Seterusnya bagi langkah yang kedua adalah mengira bilangan kekerapan perkataan muncul dalam teks. Jadual 3.8 di bawah menunjukkan bilangan kekerapan setiap perkataan.

Jadual 3.8 Kekerapan Perkataan Muncul

Perkataan	**Kekerapan**
hour	1
get	1
food	2
tables	1
luke	1

b e r s a m b

…s a m b u n g a n	
warm	1
sever	1
around	1
totally	1
overwhelmed	1

Untuk langkah yang ketiga adalah membuat matriks kejadian bersama bagi mengetahui bilangan dua perkataan berlaku serentak. Matrik yang telah dibina seperti Rajah 3.8 berikut:

	hour	get	food	tables	luke	warm	sever	around	totally	overwhelmed
hour	0	1	0	0	0	0	0	0	0	0
get	0	0	1	0	0	0	0	0	0	0
food	0	0	0	1	1	0	0	0	0	0
tables	0	0	1	0	0	0	0	0	0	0
luke	0	0	0	0	0	1	0	0	0	0
warm	0	0	0	0	0	0	1	0	0	0
sever	0	0	0	0	0	0	0	1	0	0
around	0	0	0	0	0	0	0	0	1	0
totally	0	0	0	0	0	0	0	0	0	1
overwhelmed	0	0	0	0	0	0	0	0	0	0

Rajah 3.7 Matriks Kejadian Bersama

Langkah yang terakhir adalah pengiraan PMI dengan menggunakan formula seperti persamaan 3.2. Keputusan yang diperolehi seperti Jadual 3.9 berikut:

Jadual 3.9 Pengiraan PMI

Perkataan 1	**Perkataan 2**	ï ĩ ˜
hour	get	3.4594
tables	food	3.4594
food	luke	3.4594
luke	warm	3.4594
warm	sever	3.4594
sever	around	3.4594
around	totally	3.4594
totally	overwhelmed	3.4594
get	food	2.4594
food	tables	2.4594

3.8 FASA 5 : KLASIFIKASI DAN PENGIRAAN SENTIMEN

Mengklasifikasi teks ulasan restoran mengikut polariti sentimen bertujuan mengelaskan polariti dokumen teks, ayat dan ciri kepada kategori positif dan negatif. Terdapat tiga aktiviti yang dilaksanakan mengikut kepada tiga jenis skor sentimen yang dijalankan.

3.8.1 Skor 1/0

Dalam analisis sentimen setiap teks, polariti setiap teks dikira mengikut kategori negatif atau positif. Pengiraan sentimen adalah pengumpulan jumlah entiti yang membawa sentimen di dalam teks. Berdasarkan POS-1 dan POS-2 yang dihasilkan, pemarkahan akan diberikan kepada setiap teks berdasarkan nilai pemberat yang ada dalam leksikon POS. Pada mulanya skor akan ditetapkan kepada nilai 0. Setiap perkataan positif dan negatif yang diambil daripada leksikon sentimen akan ditambah setiap kali perkataan muncul di dalam teks. Seterusnya, daripada jumlah skor yang dikumpulkan sama ada lebih besar daripada 0 atau kurang daripada 0. Daripada jumlah yang diperolehi pengelasan sentimen dilabelkan kepada kelas Negatif sekiranya jumlah skor yang diperolehi kurang daripada 0 dan kelas Positif sekiranya jumlah skor yang diperolehi lebih besar daripada 0. Formula bagi pengiraan skor ini adalah seperti persamaan 3.3 di bawah dan penjelasan ini diringkaskan seperti pseudocode pada Jadual 3.10 di bawah:

[illegible] …(3.3)

Jadual 3.10 Pseudocode Pengelasan Sentimen

Pseudocode

```
FOR EACH review in reviews
    Set score to 0
    FOR EACH word IN review['Cle
        IF word is in lexicon THEN
            Add score + lexicon[word] to score

    Set the value of review['Score'] to the value of score
    IF score is greater than 0 THEN
        set review['MySentiment'] to 'POSITIVE'
    ELSE
        set review['MySentiment'] to 'NEGATIVE'
```

Polariti sentimen dihasilkan dengan membezakan perkataan menjadi dua kelas iaitu kelas Negatif dan Positif. Jadual 3.11 hingga Jadual 3.12 pula menunjukkan contoh lima ulasan Positif dan Negatif yang teratas.

Jadual 3.11 Ulasan Positif Bagi Skor Sentimen 1/0

Ulasan Asal	**Ulasan Selepas Pemprosesan**	**Skor**
I ordered the Voodoo pasta and it was the first time I'd had really excellent pasta since going gluten free several years ago	ordered voodoo pasta first time i'd really excellent pasta since going gluten free several years ago	13
To summarize... the food was incredible, nay, transcendant... but nothing brings me joy quite like the memory of the pneumatic condiment dispenser	summarize food incredible nay transcendant nothing brings joy quite like memory pneumatic condiment dispenser	13
They have a good selection of food including a massive meatloaf sandwich, a crispy chicken wrap, a delish tuna melt and some tasty burgers	good selection food including massive meatloaf sandwich crispy chicken wrap delish tuna melt tasty burgers	14
If you love authentic Mexican food and want a whole bunch of interesting, yet delicious meats to choose from, you need to try this place	love authentic mexican food want whole bunch interesting yet delicious meats choose need try place	14
Only Pros : Large seating area/ Nice bar area/ Great simple drink menu/ The BEST brick oven pizza with homemade dough!	pros large seating area nice bar area great simple drink menu best brick oven pizza homemade dough	16

Jadual 3.12 Ulasan Negatif Bagi Skor Sentimen 1/0

Ulasan Asal	**Ulasan Selepas Pemprosesan**	**Skor**
Lobster Bisque, Bussell Sprouts, Risotto, Filet ALL needed salt and pepper..and of course there is none at the tables	lobster bisque bussell sprouts risotto filet needed salt pepper and course none tables	-5
I left with a stomach ache and felt sick the rest of the day	left stomach ache felt sick rest day	-5
Worse of all, he humiliated his worker right in front of me..Bunch of horrible name callings	worse humiliated worker right front me bunch horrible name callings	-5
I work in the hospitality industry in Paradise Valley and have refrained from recommending Cibo any longer	work hospitality industry paradise valley refrained recommending cibo longer	-5
I guess I should have known that this place would suck, because it is inside of the Excalibur, but I didn't use my common sense	guess known place would suck inside excalibur use common sense	-4

3.8.2 TF-IDF

Bagi klasifikasi dan pengiraan skor sentimen pada mulanya akan dikira terlebih dahulu skor sentimen bagi ulasan. Seterusnya string ulasan dipisahkan ke dalam senarai perkataan dan tetapkan senarai yang terhasil kepada pembolehubah yang dipanggil 'words'. Skor akan ditetapkan kepada 0. Dalam setiap proses gelung perkataan dalam senarai 'words', sekiranya perkataan ter sentimen akan bertambah. Penerangan ini diringkaskan seperti yang dipaparkan menggunakan pseudocode seperti Jadual 3.13. Seterusnya adalah mengklasifikasikan perkataan kepada binari klasifikasi iaitu postif dan negatif. Dalam klasifikasi ini ramalan positif dan negatif ditetapkan dengan nilai ambang kepada 0.

Jadual 3.13 Pseudocode Pengiraan Sentimen TF-IDF

Pseudocode
BEGIN calculate_sentiment_score(review) Split the review string into a list of words and assign the resulting list to a variable called 'words' Set score to 0 FOR EACH word IN words IF word is in sentiment_lexicon THEN Add score + sentiment_lexicon[word] to score RETURN score

3.8.3 PMI

Seterusnya bagi skor sentimen PMI, proses klasifikasi adalah dengan menggunakan skor PMI yang telah ditentukan terlebih dahulu dalam Fasa 4. Berdasarkan skor PMI yang dikira, skor sentimen kepada perkataan boleh ditetapkan. Skor PMI adalah positif untuk perkaitan dengan sentimen positif dan negatif untuk perkaitan dengan sentimen negatif. Bagi mengklasifikasikan skor fungsi bernama classify_text yang mengambil input teks ditetapkan. Pengiraan skor positif dan negatif berdasarkan beberapa leksikon yang dipratentukan iaitu leksikon positif dan leksikon negatif. Sekiranya skor positif lebih besar daripada skor negatif maka ia dikelaskan kepada positif dan sebaliknya adalah diklasifikasikan kepada negatif. Penerangan ini diringkaskan seperti Jadual 3.14 seperti berikut:

Jadual 3.14 Pseudocode Klasifikasi PMI

Pseudocode

```
FUNCTION classifyText(text):
    Convert text to lowercase and split into words
    words = splitAndLowercase(text)

    Calculate positive and negative scores
    positive_score = calculateScore(words, positiveLexicon, pmiScores)
    negative_score = calculateScore(words, negativeLexicon, pmiScores)

    Determine the overall sentiment
    IF positive_score > negative_score THEN
        return 'positive'
    ELSE
        return 'negative'
```

3.9 PENILAIAN

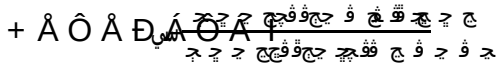
Fasa terakhir dalam kajian ini adalah membuat penilaian ke atas keputusan yang diperolehi. Penilaian ini akan membandingkan antara model yang dibangunkan iaitu Set POS-1 dan Set POS-2 yang menggunakan skor penentuan 1/0, Set POS-1 dengan skor sentimen TF-IDF, Set POS-2 dengan skor sentimen TF-IDF, Set POS-1 dengan skor sentimen PMI dan Set POS-2 dengan skor sentimen PMI. Eksperimen ini dijalankan dengan menggunakan parameter penilaian ketepatan. Jadual 3.15 menunjukkan perbandingan model yang terlibat.

Jadual 3.15 Perbandingan Model

	Kata Sifat	**Kata Nama**	**Kata Kerja**	**Kata Keterangan**
Set POS-1 + Skor 1/0	/	/	/	/
Set POS-2 + Skor 1/0	/			
Set POS-1 + TF-IDF	/	/	/	/
Set POS-2 + TF-IDF	/			
Set POS-1 + PMI	/	/	/	/
Set POS-2 + PMI	/			

Ketepatan ditakrifkan sebagai metrik untuk menilai model klasifikasi. Kaedah pengiraannya adalah dengan membahagikan jumlah ramalan yang betul dengan jumlah keseluruhan ramalan. Formula lengkap pengiraan adalah seperti persamaan 3.4 berikut:

…(3.4)



3.10 PENUTUP

Bab ini telah menerangkan metodologi kajian atau kaedah-kaedah yang dijalankan untuk mencapai matlamat kajian. Bab seterusnya akan membincangkan hasil daripada metodologi yang dilaksanakan.

BAB IV

HASIL KAJIAN

4.1 PENGENALAN

Bab ini membincangkan hasil kajian berdasarkan metodologi sepertimana yang dijelaskan pada Bab 3. Beberapa eksperimen telah dijalankan untuk mendapatkan perbandingan dan mengenal pasti set golongan kata terbaik dengan menggunakan pendekatan berasaskan leksikon dan penandaan golongan kata dalam meramal sama ada ulasan terhadap restoran itu positif atau negatif. Penetapan eksperimen akan dijelaskan terlebih dahulu diikuti dengan langkah-langkah eksperimen yang telah dilaksanakan.

4.2 PENETAPAN EKSPERIMEN

Untuk melaksanakan eksperimen kajian dengan jayanya alatan atau perisian yang cekap amat membantu. Alatan yang digunakan dapat menganalisis perbualan teks dan menilai nada, niat dan emosi di sebalik setiap teks yang dikaji. Sepertimana yang dijelaskan pada bab sebelumnya, aplikasi yang digunakan dalam kajian ini adalah Google Colab dengan bahasa pengaturcaraan Python. Perpustakaan atau *Library* python mengandungi koleksi kod yang berfungsi untuk memudahkan dan mempercepatkan kerja pengekodan. Jadual 4.1 akan menyenaraikan pakej yang digunakan seperti berikut:

Jadual 4.1 Pakej Perpustakaan Google Colab

Pakej Perpustakaan	Fungsi
PANDAS	Untuk membaca dan memproses file set data dari pelbagai format seperti .txt dan .csv,
MATPLOTLIB	Sebagai visualisasi data dan digunakan untuk memplot graf, histogram dan carta bar yang intuitif.
SPACY	Memberi gambaran tentang struktur tatabahasa teks

b e r s a mb

…s a mb u n g a n

SKLEARN	Perpustakaan yang menyokong kebanyakan algoritma pembelajaran mesin samada diselia atau tidak diselia.
NLTK	Untuk pemprosesan data bahasa manusia dan digunakan untuk klasifikasi, tokenisasi, stemming dan sebagainya.
SPACY	Memberi gambaran tentang struktur tatabahasa teks
NUMPY	Cekap mengendalikan pelbagai jenis operasi matematik terutama pengiraan arrays dan matriks.

4.2.1 Pengekstrakan Data

Sepertimana yang dijelaskan pada bab yang sebelumnya, kerangka data yang diperolehi dimuatkan ke dalam perpustakaan python untuk diproses. Sebanyak 1,000 teks ulasan mengandungi 500 ulasan yang dilabelkan sebagai ulasan yang mendapat maklum balas positif dan 500 ulasan dilabelkan sebagai maklum balas negatif. Label 1 menandakan maklum balas positif dan label 0 menandakan maklum balas negatif seperti Rajah 4.1 berikut:

A	B	C
id	review	liked
1	Wow... Loved this place.	1
2	Crust is not good.	0
3	Not tasty and the texture was just nasty.	0
4	Stopped by during the late May bank holiday off Rick Steve recommendation and loved it.	1
5	The selection on the menu was great and so were the prices.	1
6	Now I am getting angry and I want my damn pho.	0
7	Honeslty it didn't taste THAT fresh.)	0
8	The potatoes were like rubber and you could tell they had been made up ahead of time being kept under a warmer.	0
9	The fries were great too.	1
10	A great touch.	1
11	Service was very prompt.	1
12	Would not go back.	0
13	The cashier had no care what so ever on what I had to say it still ended up being wayyy overpriced.	0
14	I tried the Cape Cod ravoli, chicken, with cranberry...mmmm!	1
15	I was disgusted because I was pretty sure that was human hair.	0
16	I was shocked because no signs indicate cash only.	0
17	Highly recommended.	1
18	Waitress was a little slow in service.	0
19	This place is not worth your time, let alone Vegas.	0
20	did not like at all.	0
21	The Burrittos Blah!	0
22	The food, amazing.	1
23	Service is also cute.	1
24	I could care less... The interior is just beautiful.	1

Rajah 4.1 Kerangka Data Format CSV

Seterusnya adalah penyediaan data untuk dianalisa iaitu melibatkan prapemprosesan data. Data hendaklah dibersihkan dengan menyeragamkan teks kepada huruf kecil, membuang perkataan yang tidak mempunyai maksud, membahagikan teks

kepada satu perkataan, membuang semua tanda baca, membuang perkataan yang tidak mempunyai maksud dan menjadikan semua perkataan kepada kata akar. Rajah 4.2 menunjukkan contoh hasil sebelum dan selepas pelaksanaan prapemprosesan.

	review	clean_review
0	Wow... Loved this place.	wow place
1	Crust is not good.	crust good
2	Not tasty and the texture was just nasty.	tasty texture nasty
3	Stopped by during the late May bank holiday of...	late bank holiday rick steve recommendation
4	The selection on the menu was great and so wer...	selection great prices
5	Now I am getting angry and I want my damn pho.	angry damn pho
6	Honeslty it didn't taste THAT fresh.)	honeslty taste fresh
7	The potatoes were like rubber and you could te...	potatoes rubber tell ahead time warmer
8	The fries were great too.	fries great
9	A great touch.	great touch

Rajah 4.2 Sebelum dan Selepas Pembersihan Data

4.3 LEKSIKON SENTIMEN

Hasil pembinaan leksikon sentimen akan digunakan untuk menganalisa sentimen dan pengiraan skor. Dalam kajian ini leksikon sentimen dalam format csv dibangunkan menjadi enam set leksikon sentimen yang terdiri daripada mylexiconPOS1.csv, mylexiconPOS2.csv, positive_lexicon_POS1.csv, positive_lexicon_POS2.csv, negative_lexicon_POS1.csv dan negative_lexicon_POS2.csv. Jadual 4.2 menunjukkan bilangan perkataan yang diperolehi bagi setiap set.

Jadual 4.2 Bilangan Perkataan Leksikon Sentimen

Set	Bilangan Perkataan
mylexiconPOS1.csv	4,398
mylexiconPOS2.csv	1,144
positive_lexicon_POS1.csv	2,262
positive_lexicon_POS2.csv	672
negative_lexicon_POS1.csv	2,136
negative_lexicon_POS2.csv	472

Merujuk kepada Jadual 4.2 di atas, mylexiconPOS1.csv mengandungi kandungan perkataan dan bilangan yang sama setelah digabungkan dengan positive_lexicon_POS1.csv dan negative_lexicon_POS1.csv. Manakala mylexiconPOS2.csv juga mengandungi kandungan perkataan dan bilangan yang sama setelah digabungkan antara positive_lexicon_POS2.csv dan negative_lexicon_POS2.csv. Pembinaan leksikon bagi positive_lexicon_POS1.csv, positive_lexicon_POS2.csv, negative_lexicon_POS1.csv dan negative_lexicon_POS2.csv adalah untuk tentukan leksikon positif dan negatif dengan menggunakan pengukuran skor sentimen PMI. Walau bagaimanapun, nilai pemberat yang sama masih dikekalkan bagi keenam-enam set ini iaitu -1 diberikan untuk ulasan negatif dan +1 untuk ulasan positif. Jadual 4.3 menunjukkan contoh senarai leksikon sentimen yang dihasilkan bagi set POS-1 positif leksikon.

Jadual 4.3 Leksikon Sentimen Set POS-1 Bagi Positif Leksikon

Leksikon Sentimen				
absolute	customer	hawaiian	omg	skimp
absolutely	customize	healthy	onion	slaw
absolutley	cute	heard	opinion	slices
accident	daily	heart	opportunity	small
accommodations	damn	heat	options	smooth
accomodate	dark	hella	order	smoothies
accordingly	date	helpful	ordered	soi
actually	dates	hereas	orders	solid
affordable	daughter	high	original	someone
afternoon	day	highlights	outdoor	somethat
ago	dead	highly	outrageously	something
airport	deal	hip	outstanding	son
almost	decent	hiro	oven	songs
also	decide	hits	overall	soon
always	decision	hole	overwhelmed	soooo
amazing	decor	holiday	owner	sooooo
ambience	def	home	owners	soundtrack
amount	definately	homemade	pace	soup
ample	degree	honest	pack	soups
andddd	delicate	honestly	pancake	sour
anyone	delicioso	hope	pancakes	space
anything	delicious	hot	panna	special

Jadual 4.4 terdiri daripada contoh senarai leksikon sentimen yang dihasilkan untuk set POS-1 bagi negatif leksikon.

Jadual 4.4 Leksikon Sentimen Set POS-1 Bagi Negatif Leksikon

Leksikon Sentimen			
lukewarm	tell	extremely	forgetting
right	ahead	really	note
ok	time	many	ventilation
crowd	warmer	restaurants	upgrading
finally	go	love	letdown
stomach	back	dine	camelback
bill	cashier	weekend	flower
worse	care	helpful	shop
bother	ever	friendly	cartel
job	still	rarely	cheese
doubt	wayyy	husband	boiled
authentic	pretty	ate	legs
special	sure	lunch	gas
selection	human	red	station
trip	hair	curry	sign
find	shocked	bamboo	worker
wife	signs	shoots	front
dinner	cash	menu	bunch
anytime	waitress	always	name
friend	little	quality	callings
bring	slow	served	tragedy
something	service	servers	airline
nice	place	pace	noca
rare	worth	thumbs	gyro
fail	let	attention	lettuce
needless	alone	forty	thoroughly
mid	vegas	crostini	anyway
room	burrittos	hit	greasy
check	blah	bakery	unhealthy
bathroom	never	leftover	similarly
fries	salad	everything	delivery
places	hour	today	man
immediately	get	cost	apology
next	food	spinach	late

Manakala Jadual 4.5 terdiri daripada contoh senarai leksikon sentimen yang dihasilkan untuk set POS-2 bagi positif leksikon.

Jadual 4.5 Leksikon Sentimen Set POS-2 Bagi Positif Leksikon

Leksikon Sentimen					
affordable	dish	hawaiian	middle	prime	subway
amazing	double	healthy	military	professional	super
ample	downtown	helpful	miss	profiterole	sure
atmosphere	drive	high	mixed	public	sushi
attentive	eclectic	homemade	modern	pumpkin	sweet
authentic	empty	hot	mom	puree	table
awesome	enjoyable	hottest	mouth	quick	tasty
bachi	enjoyed	huge	much	quite	ten
bacon	enough	hungry	multiple	rare	thick
bakery	enthusiastic	ice	nargile	real	thin
beateous	ethic	iced	new	realized	thought
beautiful	excellent	imaginative	next	reasonable	tiny
believe	exceptional	impeccable	nice	recent	tiramisu
best	extensive	impressed	nicest	recommend	told
better	extra	incredible	nobu	recommended	top
big	extraordinary	indian	north	red	touch
black	fabulous	inexpensive	nt	regular	tribute
bloddy	familiar	informative	occasional	relaxed	tried
blue	fantastic	interesting	oh	restaurant	typical
bread	fast	interior	ohhh	review	unbelievable
bruschetta	favorite	italian	old	rich	unique
buldogis	feel	jalapeno	ordered	rick	unreal
ca	fine	jamaican	original	right	update
cafe	finish	japanese	outdoor	round	usual
cheese	first	large	outstanding	salad	vegas
chicken	flavorful	larger	overall	salmon	vegetarian
chinese	flavourful	last	overwhelmed	san	veggitarian
classic	francisco	late	past	satisfied	visited
classy	free	least	patio	satisfying	wagyu
clean	fresh	legit	pay	scottsdale	waitress
come	friendly	light	peas	seafood	want
comfortable	fry	lighter	pecan	seal	warm
cooked	fs	like	perfect	seasonal	weekly
cool	full	little	personable	second	welcome

Jadual 4.6 pula terdiri daripada contoh senarai leksikon sentimen yang dihasilkan untuk set POS-2 bagi negatif leksikon.

Jadual 4.6 Leksikon Sentimen Set POS-2 Bagi Negatif Leksikon

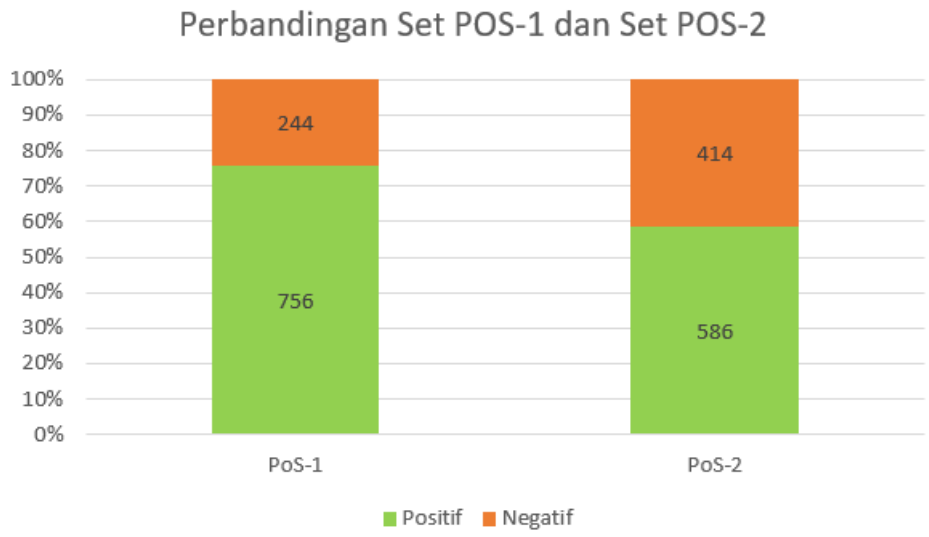
Leksikon Sentimen				
actual	cold	forgetting	low	recommended
allergy	coming	forth	lukewarm	red
amazing	common	forty	lunch	refill
angry	complete	fresh	main	refried
apart	consistent	fried	mall	reviewer
arepas	convenient	front	many	ridiculous
ask	cow	frozen	meal	risotto
ate	crazy	frustrated	mediocre	roasted
atrocious	crumby	full	mellow	rude
authentic	crusty	girlfriend	milkshake	ryan
average	decent	give	much	sad
avoid	despicable	gold	multiple	salad
awful	dessert	good	nachos	salt
bachi	different	gordon	nasty	sandwich
bad	dirty	greasy	neat	sat
bakery	disagree	greatest	needless	scallop
bamboo	disappointed	green	negligent	served
bay	disappointing	gross	new	several
beat	disapppointment	hard	next	shocked
best	disbelief	helpful	nice	short
better	dish	high	non	shower
big	disrespected	hilarious	not	shrimp
bigger	double	honest	nude	sick
biggest	douchey	horrible	ok	similar
black	driest	hostess	old	single
blame	dry	hot	older	slow
blandest	due	huevos	ordered	small
bloodiest	eat	huge	outside	smaller
boba	edible	human	overall	somehow
boiled	eew	ice	overwhelmed	sound
boring	empty	impressed	pale	special
bring	english	impressive	particular	spicy
brownish	enough	inch	pepper	spotty
busy	entire	inconsiderate	petrified	stale

4.4 EKSPERIMEN 1 : SET POS-1 DAN SET POS-2 DENGAN SKOR SENTIMEN 1/0

Eksperimen 1 adalah melibatkan proses mendapatkan ketepatan sentimen menggunakan skor sentimen berdasarkan label 1 dan 0. Eksperimen ini dibahagikan kepada dua bahagian iaitu Set POS-1 dan Set POS-2. Sepertimana yang dijelaskan dalam Bab 3, Set POS-1 terdiri daripada kata nama, kata kerja, kata sifat dan kata keterangan. Manakala Set POS-2 terdiri daripada kata sifat. Pemilihan Penn Treebank digunakan sebagai set tag POS. Bagi pengiraan skor polariti garis dasar ditetapkan iaitu polariti skor lebih daripada 0 merupakan positif manakala sebaliknya merupakan negatif sentimen. Garis dasar ini akan digunakan sama bagi kesemua eksperimen yang dijalankan. Jadual 4.7 memaparkan hasil keputusan yang diperolehi bagi kedua-dua set.

Jadual 4.7 Bilangan Perkataan Sentimen Leksikon

	POS-1 (Kata Nama, Kata Kerja, Kata Keterangan dan Kata Sifat)	**POS-2 (Kata Sifat)**
Positif	756 (75.6 %)	586 (58.6%)
Negatif	244 (24.4 %)	414 (41.4 %)



Rajah 4.3 Graf Perbandingan Set POS-1 dan Set POS-2

Merujuk kepada jadual 4.7 dan Rajah 4.3, set POS-1 menghasilkan 75.6% sentimen positif berbanding 58.6% bagi set POS-2. Manakala bagi sentimen negatif, set POS-1 menghasilkan 24.4% berbanding set POS-2, 41.4 %. Kesimpulannya bagi kedua-dua set, sentimen positif adalah lebih mendominasi berbanding sentimen negatif.

Walau bagaimanapun merujuk kepada Jadual 4.8 berikutnya menunjukkan bahawa set POS-2 mempunyai bilangan ketepatan yang lebih tinggi iaitu 75.80% berbanding set POS-1 yang mempunyai ketepatan sebanyak 74.20%. Keputusan ini menunjukkan set POS-2 merupakan set POS yang terbaik jika dibandingkan antara kedua-dua set ini.

Jadual 4.8 Perbandingan Antara Set POS-1 dan Set POS-2 Mengikut Ketepatan

	POS-1 (Kata Nama, Kata Kerja, Kata Keterangan dan Kata Sifat)	**POS-2 (Kata Sifat)**
Bilangan Ramalan Yang Tepat	742	758
Ketepatan	0.742	0.758
Peratus Ketepatan	74.20 %	75.80 %

4.5 EKSPERIMEN 2 : SET POS-1 DAN SET POS-2 DENGAN SKOR SENTIMEN TF-IDF

Eksperimen 2 melibatkan proses mendapatkan ketepatan sentimen dengan melibatkan penggunaan skor sentimen TF-IDF. TF-IDF ialah kaedah untuk mengira berat atau skor setiap perkataan. Pengiraan TF-IDF dijalankan setelah proses prapemprosesan dan penandaan golongan kata dilaksanakan. Dengan pengiraan TF-IDF telah menghasilkan 1,000 barisan × 1,574 lajur bagi set POS-1 manakala bagi set POS-2 ialah 1,000 barisan × 470 lajur. Ia bermaksud bagi set POS-1 mengandungi 1,574 istilah manakala set POS-2 470 istilah. Rajah 4.4 dan Rajah 4.5 di bawah memaparkan contoh keputusan TF-IDF yang diperolehi.

O	P	Q	R	S	T	U	V	W	X
ahead	airline	airport	albondigas	allergy	almost	alone	also	always	amazing
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0.448805	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0.537807	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0.867871
0	0	0	0	0	0	0	0.798155	0	0
0	0	0	0	0	0	0	0	0	0

Rajah 4.4 TF-IDF Set POS-1

HJ	HK	HL	HM	HN	HO	HP
late	least	legit	light	lighter	like	little
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0.686379	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0.533489
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

Rajah 4.5 TF-IDF Set POS-2

Jadual 4.9 Perbandingan Antara Set POS-1+TF-IDF dan Set POS-2+TF-IDF Mengikut Ketepatan

	POS-1 + TF-IDF (Kata Nama, Kata Kerja, Kata Keterangan dan Kata Sifat)	**POS-2 + TF-IDF (Kata Sifat)**
Ketepatan	0.754	0.773
Peratus Ketepatan	75.40 %	77.30 %

Jadual 4.9 di atas memaparkan keputusan ketepatan atau kejituan apabila set POS-1 dan POS-2 dengan menetapkan skor sentimen menggunakan skor TF-IDF. Keputusan menunjukkan bahawa set POS-2 dengan ketepatan 77.30% mencapai ketepatan yang lebih tinggi berbanding set POS-1 yang memperolehi 75.40%. Keputusan ini membayangkan bahawa pemilihan penandaan POS yang lebih khusus berperanan memberikan hasil keputusan yang lebih baik berbanding banyak penandaan POS.

4.6 EKSPERIMEN 3 : SET POS-1 DAN SET POS-2 DENGAN SKOR SENTIMEN PMI

Eksperimen yang seterusnya iaitu eksperimen yang ke-3 melibatkan proses mendapatkan ketepatan sentimen dengan melibatkan penggunaan dan pengiraan PMI sebagi skor sentimen. PMI atau maklumat timbal balik titik diaplikasikan untuk mengukur kekuatan perkaitan antara dua istilah dalam teks ulasan restoran. Pengiraan PMI juga dilaksanakan selepas set data dibersihkan dan penandaan POS dijalankan. Hasil pengiraan skor bagi PMI yang diperolehi adalah seperti di Lampiran A. Jadual 4.10 memaparkan hasil pengiraan kejituan setelah eksperimen dijalankan.

Jadual 4.10 Perbandingan Antara Set POS-1+ PMI dan Set POS-2+PMI Mengikut Ketepatan

	POS-1 + PMI (Kata Nama, Kata Kerja, Kata Keterangan dan Kata Sifat)	**POS-2 + PMI (Kata Sifat)**
Ketepatan	0.484	0.549
Peratus Ketepatan	48.40%	54.90%

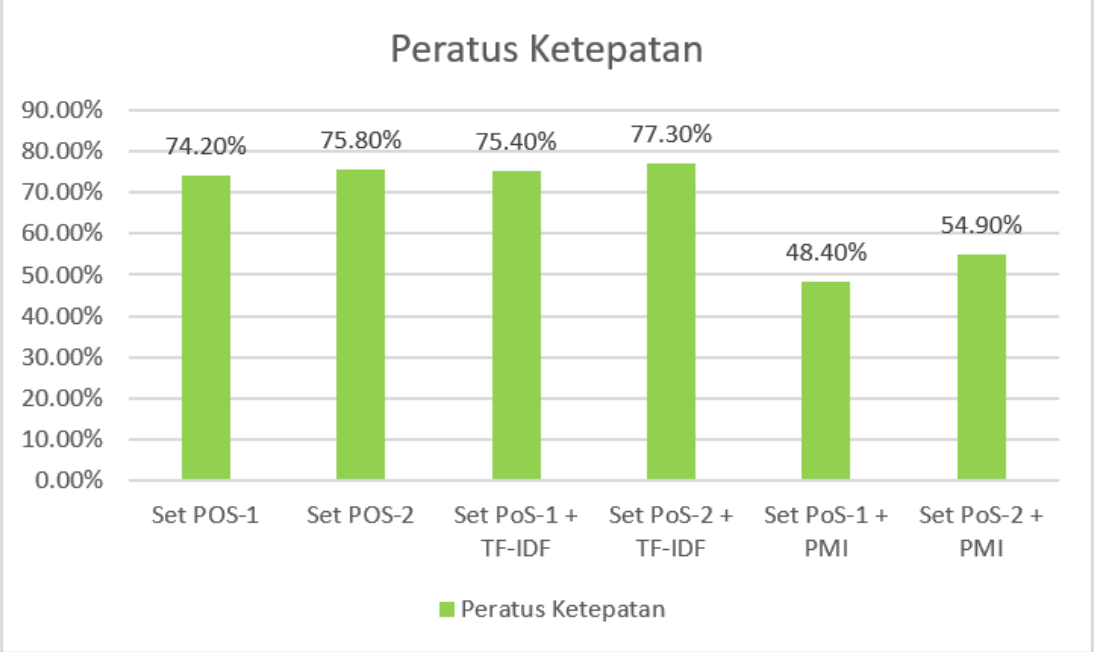
Jadual 4.10 menunjukkan bahawa keputusan set POS-2 masih lagi merekodkan keputusan kejituan yang lebih baik berbanding set POS-1 dimana set POS-2 memperolehi ketepatan 0.549 bersamaan 54.90% berbanding set POS-1, 0.484 yang bersamaan 48.40%. Bahagian seterusnya akan membincangkan penilaian hasil keputusan bagi eksperimen terlibat yang mencatatkan keputusan kejituan yang terbaik dan pemilihan set POS yang terbaik.

4.7 PENILAIAN

Jadual 4.11 dan Rajah 4.6 menunjukkan keputusan kejituan bagi setiap eksperimen 1, eksperimen 2 dan eksperimen 3. Secara keseluruhannya bagi eksperimen 1 dan eksperimen 2, keputusan yang diperolehi tidak begitu ketara antara satu sama lain berbanding eksperimen 3 yang melibatkan pengiraan maklumat timbal balik titik yang menunjukkan jurang yang besar. Ia akan dibincangakan lebih lanjut dalam bahagian perbincangan.

Jadual 4.11 Perbandingan Antara Eksperimen 1, Eksperimen 2 dan Eksperimen 3

Set	Ketepatan	Peratus Ketepatan
Set POS-1 + Skor 1/0	0.742	74.20%
Set POS-2 + Skor 1/0	0.758	75.80%
Set POS-1 + TF-IDF	0.754	75.40%
Set POS-2 + TF-IDF	0.773	77.30%
Set POS-1 + PMI	0.484	48.40%
Set POS-2 + PMI	0.549	54.90%



Rajah 4.6 Graf Perbandingan Antara Eksperimen 1, Eksperimen 2 dan Eksperimen 3

Antara keenam-enam set yang diuji, set POS-2 dengan skor sentimen TF-IDF mencapai ketepatan yang tertinggi iaitu 77.30% diikuti set POS-2 dengan skor 1/0

dengan peratusan 75.80%. Manakala dengan perbezaan hanya 0.4% set set POS-1 dengan skor sentimen TF-IDF memperolehi ketepatan 75.40% dan set POS-1 dengan skor 1/0 pula memperolehi 74.20%. Dapatan bagi set POS-1 dengan skor sentimen PMI dan set POS-2 dengan skor sentimen PMI ialah 48.40% dan 54.90% masing-masing dicatatkan.

4.7.1 Penilaian Set Leksikon

Untuk menilai set leksikon sentimen yang dihasilkan, satu pengujian dengan menggunakan set data yang sama namun hanya menggunakan 500 ulasan pelanggan yang dipilih secara rawak telah dijalankan. Set pengujian ini telah diambil dan diuji menggunakan model yang mencatatkan ketepatan yang tertinggi iaitu set POS-2 dengan skor sentimen TF-IDF. Keputusan yang diperolehi adalah seperti Jadual 4.12.

Jadual 4.12 Perbandingan Keputusan Ketepatan Menggunakan Keseluruhan Set Data Ulasan Restoran dan Sebahagian Set Data Ulasan Restoran

	Keseluruhan Set Data Ulasan Restoran	**Sebahagian Set Data Ulasan Restoran**
Ketepatan	0.773	0.768
Peratus Ketepatan	77.30%	76.80%

Berdasarkan keputusan yang diperolehi, pengujian dengan menggunakan set data pengujian mencatatkan ketepatan sebanyak 76.8%. Hasil ketepatan ini tidak begitu ketara perbezaan di mana hanya berbeza 0.5 berbanding menggunakan set data latihan dan menunjukkan model yang dihasilkan boleh diyakini.

4.8 PERBINCANGAN

Hasil daripada eksperimen yang dijalankan boleh disimpulkan bahawa set POS-2 yang mengandungi penandaan golongan kata, kata sifat beserta skor sentimen frekuensi songsang jangka kekerapan dokumen adalah set golongan kata dan skor sentimen yang terbaik dalam meramal sama ada ulasan terhadap restoran itu positif atau negatif. Didapati perbezaan antara set POS-1 dengan skor 1/0, set POS-2 dengan skor 1/0, set POS-1 dengan skor sentimen TF-IDF dan set POS-2 dengan skor sentimen TF-IDF tidak menunjukkan jurang perbezaan yang besar. Namun begitu dapat diperhatikan bahawa hasil kajian yang melibatkan set POS-1 dengan skor sentimen PMI dan set POS-2 dengan skor sentimen PMI menghasilkan jurang perbezaan yang begitu ketara iaitu 48.40% dan 54.90% masing-masing. Perbezaan yang ketara ini perlu dijelaskan dan diperhalusi.

Terdapat kelemahan dengan teknik PMI ini melalui kajian yang dijalankan oleh (Bos & Frasincar 2022), penyelidik tersebut telah menambahbaik teknik PMI yang digunakan dengan menggunakan NPMI iaitu Wajaran Dinormalisasi Maklumat Timbal Balik Titik atau dalam Bahasa Inggeris Weighted Normalized Pointwise Mutual Information untuk mendapatkan hasil yang lebih baik bagi membina sentimen leksikon domain kewangan. Kelemahan yang ditemui adalah istilah yang hanya muncul dalam mesej salah satu daripada dua kelas sentimen dan menyebabkab kiraan dalam ukuran PMI kelas sentimen yang lain menjadi sama dengan sifar. Logaritma tidak ditentukan untuk sifar dan pengiraan ukuran PMI yang sepadan tidak dapat dibuat.

4.9 PENUTUP

Bab ini menjelaskan hasil kajian yang diperolehi selepas eksperimen-ekperimen yang dirangka dijalankan. Secara kesimpulannya, bahagian ini telah menunjukkan bahawa set golongan kata POS-2 dengan skor sentimen TF-IDF telah menghasilkan prestasi yang lebih tepat dalam pembinaan leksikon sentimen secara automatik menggunakan set data ulasan restoran ini.

BAB V

RUMUSAN

5.1 PENGENALAN

Bab ini menjelaskan perkara berkaitan dengan rumusan kajian, pencapaian objektif dan penerangan berkenaan dengan sumbangan kajian serta kajian yang relevan pada masa hadapan untuk proses meningkatkan hasil penyelidikan analisis sentimen melibatkan domain perniagaan makanan dan retoran. Hasil daripada perbandingan antara set golongan kata beserta skor sentimen yang digunakan dirumuskan kepada pengetahuan baharu terhasil yang boleh dimanfaatkan oleh pengusaha restoran dan makanan serta pelanggan yang ingin mencuba sesuatu yang baharu.

5.2 RINGKASAN KAJIAN

Dengan penggunaan media sosial seperti Twitter dan Facebook yang merangkumi berbilion-bilion pengguna dari seluruh dunia berkongsi data, berinteraksi, berkongsi perasaan dan pendapat tentang pelbagai perkara tanpa mengira batasan jarak dan tempat. Sejajar dengan itu jua, terdapat sejumlah besar data media sosial yang dijana dalam bentuk yang tidak berstruktur oleh pengguna boleh diterokai dan menganalisis emosi yang terkandung di dalam teks tersebut. Media sosial menjadi platform yang penting masa kini sebagai alat untuk mendapatkan maklumat yang berguna juga sebagai saluran untuk berhubung dengan dengan dunia luar. Dalam kajian ini fokus utama adalah untuk mengenal pasti gabungan set golongan kata (POS) yang terbaik dan juga ukuran skor sentimen dalam pembinaan sentimen leksikon yang melibatkan ulasan restoran. Set data diperolehi dari *Kaggle Dataset Respository* untuk mengkaji secara umum tanpa melibatkan secara khusus negara, negeri atau mana-mana kawasan.

Kajian ini menerangkan beberapa metodologi kajian yang bermula dengan pengumpulan data, prapemprosesan, penandaan PoS, pembinaan leksikon, pengukuran dan penentuan skor sentimen yang melibatkan TF-IDF dan PMI, klasifikasi sentimen dan pengiraan sentimen seterusnya membuat pemilihan gabungan set golongan kata (POS) dan skor sentimen yang terbaik berdasarkan penilaian analisis sentimen yang dijalankan. Antara set golongan kata yang diuji, set POS-2 yang mengandungi golongan kata (POS) iaitu kata sifat beserta skor sentimen frekuensi songsang jangka kekerapan dokumen adalah set golongan dan skor sentimen yang terbaik untuk membina sentimen leksikon bagi set data ulasan restoran.

5.3 PENCAPAIAN OBJEKTIF

Kajian ini telah berjaya mencapai objektif kajian sepertimana yang dijelaskan dalam Bab 1 untuk mengkaji set golongan dan skor sentimen terbaik untuk membina sentimen leksikon secara automatik bagi set data ulasan restoran.

Pencapaian objektif utama iaitu mencadangkan set golongan yang sesuai untuk model analisis sentimen untuk set data ulasan restoran telah berjaya dicapai dengan beberapa eksperimen yang telah dijalankan. Dua jenis set golongan kata (POS) telah ditetapkan iaitu set POS-1 yang mengandungi kata nama, kata kerja, kata keterangan dan kata sifat serta set POS-2 yang terdiri daripada satu golongan kata (POS) iaitu kata sifat. Untuk mencari perbandingan antara kedua-dua set POS ini, tiga jenis skor sentimen digunakan iaitu skor 1/0, TF-IDF dan PMI turut diuji supaya hasil yang lebih optimum diharapkan dapat dicapai. Analisis sentimen ini dibangunkan dengan menggunakan metologi sepertimana yang dijelaskan pada Bab 3.

Pencapaian objektif yang kedua adalah mencadangkan ukuran skor sentimen yang sesuai untuk model analisis sentimen untuk set data ulasan restoran juga telah berjaya dipenuhi dengan pemilihan tiga jenis skor sentimen digunakan iaitu skor 1/0, TF-IDF dan PMI.

Seterusnya pencapaian objektif yang ketiga adalah membina leksikon sentimen yang sesuai berdasarkan gabungan output objektif satu dan objektif dua bagi model analisis sentimen untuk set data ulasan restoran turut dipenuhi. Dengan menggunakan

pendekatan pembinaan leksikon sentimen secara automatik telah mengenal pasti leksikon sentimen yang sesuai dengan memperolehi perbandingan ketepatan yang tinggi iaitu set golongan POS-2 menggunakan skor sentimen TF-IDF.

5.4 BATASAN KAJIAN

Kajian ini menggunakan set data yang sedia ada dari Kaggle dan mengandungi 1,000 ulasan dalam Bahasa Inggeris. Set data yang tersedia ini tidak memfokuskan kepada mana-mana demografi dan dianggap sebagai maklumat umum.

5.5 SUMBANGAN KAJIAN

Sumbangan utama dalam kajian ini dilihat dari sudut capaian objektif satu dan dua iaitu mencadangakan set golongan dan ukuran skor sentimen yang sesuai untuk model analisis sentimen bagi set data ulasan restoran. Dapatan daripada eksperimen yang dijalankan dapat dijadikan sebagai rujukan kepada komuniti penyelidik pemprosesan bahasa tabii (NLP) untuk menggunakan pendekatan yang sama serta menambahbaik teknik dan pendekatan yang digunakan untuk mendapatkan hasil yang lebih baik dalam pembinaan leksikon sentimen yang lebih canggih dan efektif. Di samping itu, penyelidik boleh membangunkan model klasifikasi yang lebih tepat dan boleh dipercayai untuk mengenal pasti polariti sentimen.

Selain itu, dapatan daripada objektif tiga iaitu membina leksikon sentimen yang sesuai berdasarkan gabungan output objektif 1 dan objektif 2 bagi model analisis sentimen untuk set data ulasan restoran ia juga dapat dijadikan sebagai platform rujukan dalam masa nyata kepada pihak industri terutama industri makanan dan restoran. Para pengusaha restoran dapat meningkatkan serta menambahbaik produk dan perkhidmatan restoran hasil daripada kefahaman perasaan dan emosi pelanggan selepas mendapatkan perkhidmatan daripada mereka. Seterusnya, dapatan pembentukan leksikon sentimen yang dihasilkan dapat diaplikasikan ke dalam industri restoran secara amnya.Para pelanggan juga dapat membuat keputusan dan pemilihan yang tepat sebelum mengunjungi serta menggunakan perkhidmatan atau produk sesebuah restoran.

5.6 CADANGAN KAJIAN MASA DEPAN

Kajian ini masih mempunyai banyak peluang yang lebih cerah dan meluas dengan mengaplikasikan pendekatan-pendekatan yang pelbagai. Ia boleh ditambah baik oleh para penyelidik lain pada masa hadapan. Antara cadangan kajian yang boleh dilaksanakan pada masa hadapan boleh digambarkan seperti berikut:

1. Menggunakan set data yang lebih khusus iaitu set data yang memfokuskan kepada demografi tertentu supaya kajian lebih tertumpu kepada penduduk, keadaan, kawasan dan situasi setempat;

2. Menggunakan set data bersaiz lebih besar sekiranya kajian yang dijalankan menggunakan pendekatan yang memerlukan latihan dan pengujian data;

3. Kajian perbandingan antara set golongan kata (POS) ini boleh menggunakan pendekatan berasaskan pembelajaran mesin dan pendekatan hibrid untuk melihat perbezaan sama ada dapat meningkatkan hasil pencapaian atau sebaliknya;

4. Menggunakan golongan kata (POS) yang berbeza selain kata nama, kata kerja, kata keterangan dan kata sifat untuk melihat hasil yang lebih bervariasi.

5. Menambah polariti neutral kerana kajian ini hanya mempertimbangkan sentimen berpolariti positif dan negatif sahaja.

6. Dapatan daripada eksperimen yang melibatkan skor sentimen PMI tidak memberikan hasil yang memberangsangkan. Untuk meningkatkan keputusan ketepatan, cadangan supaya skor sentimen PMI digabungkan dengan lain-lain teknik yang bersesuaian.

5.7 PENUTUP

Bab ini menerangkan rumusan tentang segala kajian yang telah dijalankan bermula dari Bab I sehingga Bab V. Bab ini juga menjelaskan tentang pencapaian objektif yang disenaraikan sepertimana Bab I, batasan dalam kajian ini, sumbangan kajian kepada pihak yang berkaitan serta cadangan penambahbaikan untuk kajian masa depan.

RUJUKAN

Addiga, A. & Bagui, S. 2022. Sentiment Analysis on Twitter Data Using Term Frequency-Inverse Document Frequency. *Journal of Computer and Communications 10*(08): 117–128.

Ali, I. & Hameed, N. 2017. Hybrid Tools and Techniques for Sentiment Analysis: A Review. *International Journal of Multidisciplinary Sciences and Engineering 8*(4) https://www.researchgate.net/publication/318351105.

Alqadi, R., Al-Nojaidi, H., Alabdulkareem, L., Alrazgan, M., Alghamdi, N. & Kamruzzaman, M.M. 2020. How Social Media Influencers Affect Consumers' Restaurant Selection: Sta 2020 *2nd International Conference on Computer and Information Sciences, ICCIS 2020*. Institute of Electrical and Electronics Engineers Inc.

Alrumayyan, N., Bawazeer, S., AlJurayyad, R. & Al-Razgan, M. 2018. Analyzing User Behaviors: A Study of Tips in Foursquare. *Advances in Intelligent Systems and Computing* Vol. 753, hlm. 153–168. Springer Verlag.

Amira Sumitro, P., Iskandar Mulyana, D., Saputro, W., Teknologi Informasi, J., Cipta Karya Informatika, S., Teknik Informatika, J. & Eresha, S. 2021. Analisis Sentimen Terhadap Vaksin Covid-19 di Indonesia pada Twitter Menggunakan Metode Lexicon Based. *Jurnal Informatika dan Teknologi Komputer 2*(2): 50–56. https://developer.twitter.com.

Aye, Y.M. & Aung, S.S. 2017. Sentiment Analysis for Reviews of Restaurants in Myanmar Text. *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*

Aye, Y.M. & Aung, S.S. 2018. Enhanced Sentiment Classification for Informal Myanmar Text of Restaurant Reviews. *2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications (SERA)*. IEEE.

Benamara, F., Cesarano, C., Picariello, A., Reforgiato, D. & Subrahmanian, V.S. 2007. Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone. *2007 International Conference on Weblogs and Social Media, ICWSM 2007 - Boulder, CO, United States*

Bijoyan, D. & Sarit, C. 2018. An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation.

Bos, T. & Frasincar, F. 2022. Automatically Building Financial Sentiment Lexicons While Accounting for Negation. *Cognitive Computation 14*(1): 442–460.