

ALGORITMA ANALISIS SENTIMEN YANG BERFOKUSKAN EMOJI

VISHNUPRIYA A/P RAVISHANKAR

UNIVERSITI KEBANGSAAN MALAYSIA

ALGORITMA ANALISIS SENTIMEN YANG BERFOKUSKAN EMOJI

VISHNUPRIYA A/P RAVISHANKAR

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEHI
IJAZAH SARJANAMUDA SAINS KOMPUTER DENGAN KEPUJIAN

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2020

PENGAKUAN

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

10 Julai 2020

VISHNUPRIYA A/P
RAVISHANKAR
A166014

PENGHARGAAN

Saya bersyukur kepada Tuhan kerana memberi saya kesihatan yang baik dan kebolehan untuk menyiapkan kajian ini dengan sempurna. Saya bersyukur kerana dapat menyiapkan kajian ini dalam masa yang tepat.

Selain itu, saya ingin merakam ribuan terima kasih kepada pensyarah penyelia saya, Prof. Dr. Shahrul Azman Mohd Noah atas bimbingan dan tunjuk ajar yang diberi sepanjang pelaksanaan kajian ini.

Penghargaan ini juga ditujukan kepada anggota keluarga saya kerana sentiasa memberi motivasi dan kata – kata semangat untuk menyiapkan kajian saya ini.

Akhir sekali, saya ingin mengucapkan terima kasih kepada rakan – rakan saya yang selalu meluangkan masa untuk membantu saya dalam menyiapkan kajian ini.

ABSTRAK

Kajian ini adalah berdasarkan analisis sentimen iaitu untuk mengenal pasti sentimen sesebuah teks media sosial yang mengandungi emoji. Selain itu, algoritma ini berfokuskan kepada penggunaan emoji yang digunakan semasa meluahkan perasaan atau pendapat. Hal ini kerana, emoji memainkan peranan penting dalam konteks sesebuah ayat / teks. Malah, emoji ini sukar untuk dikenalpasti dengan pengaturcaraan analisis sentimen yang biasa. Kesukaran ini dapat ditangani dengan melalui beberapa teknik. Teknik - teknik yang boleh digunakan adalah pengelas pembelajaran mesin. Pembangunan algoritma ini adalah berasaskan kaedah pembelajaran mesin. Sentimen ini terbahagi kepada tiga, iaitu positif, neutral dan negatif. Selepas menganalisis teks, algoritma ini menghasilkan output yang menunjukkan sentimen teks tersebut. Algoritma ini dibangunkan untuk kemudahan pengurusan perniagaan, individu terkenal dan pelbagai lagi bidang. Dengan menggunakan algoritma ini, mereka akan mendapat pendekatan yang terjamin daripada para masyarakat. Maklumat seperti ini sangat berguna untuk pembangunan sesuatu organisasi mahupun negara.

ABSTRACT

This research solely based on sentiment analysis, which in particular is to determine the sentiment of any provided text from social media platform given that the text or the sentence have the usage of emoji and emoticons. Besides, this algorithm focuses on the usage of emoji in sentences and texts that is giving opinions or venting out feelings in social media platform. This is mainly because, emoji plays a major role in the context of any particular sentence or text. The drawback in this research is that it is difficult to code the detection of emoji and emoticons. However, this drawback can be handled by the usage of machine learning techniques in the algorithm. The techniques that can be approached are machine learning classifiers. Sentiment is subdivided into three categories which are positive, neutral and also negative. After the whole process of analyzing text, the algorithm will produce an output that will give you the sentiment of the processed text. This algorithm is developed in order to be useful for business management, analysis of and individual / public figure and many other fields. By the usage of this algorithm, the users can benefit in knowing the public voice of their preferred topic/domain. Such information is undeniably useful in development of any particular organization or even the country.

KANDUNGAN

	Halaman
PENGAKUAN	ii
PENGHARGAAN	iii
ABSTRAK	iv
ABSTRACT	v
KANDUNGAN	vi
SENARAI JADUAL	ix
SENARAI SINGKATAN	xii
BAB I PENGENALAN	
1.1 Pendahuluan	1
1.2 Penyataan Masalah	2
1.3 Cadangan Penyelesaian Masalah	3
1.4 Matlamat dan Objektif Kajian	4
1.5 Skop	4
1.6 Kepentingan Kajian	5
1.6.1 Skop Pengurusan Perniagaan	5
1.6.2 Skop Individu	5
1.7 Cadangan Penyelesaian Masalah	5
1.8 Jadual Rancangan Pembangunan	8
BAB II KAJIAN LITERATUR	
2.1 Pengenalan	10
2.1.1 Definisi Analisis Sentimen	10
2.1.2 Proses Analisis Sentimen	12
2.1.3 Analisis Sentimen dan Analisis Emosi	13
2.1.4 Kepentingan Analisis Sentimen	22
2.1.5 Metodologi untuk Analisis Sentimen	24
2.1.6 Aplikasi Analisis Sentimen sedia ada	25
2.1.7 Perbandingan Aplikasi Analisis Sentimen Sedia Ada dan Algoritma Analisis	27
2.2 Kesimpulan	28

BAB III	SPESIFIKASI KEPERLUAN	
3.1	Pengenalan	29
3.2	Spesifikasi Keperluan Pengguna	29
3.3	Spesifikasi Keperluan Algoritma	30
	3.3.1 Keperluan Bukan Fungsian	30
	3.3.2 Keperluan Fungsian	31
3.4	Spesifikasi Keperluan Perkakasan dan perisian	31
	3.4.1 Keperluan Perkakasan dan Perisian Algoritma	31
3.5	Model Algoritma	32
3.6	Kesimpulan	34
BAB IV	SPESIFIKASI REKABENTUK	
4.1	Pengenalan	35
4.2	Rekabentuk Senibina	35
	4.2.1 Pendekatan Pembelajaran Mesin untuk Analisis Sentimen	39
4.3	Rekabentuk algoritma	40
	4.3.1 Ikon emosi / Emoji dan Sentimen	40
	4.3.2 Rekabentuk Algoritma	42
	4.3.3 Pseudokod Algoritma	50
	4.3.4 Kaedah Pembelajaran Mesin dan Model Pengelas	53
4.4	Kesimpulan	55
BAB V	IMPLEMENTASI DAN PENGUJIAN	
5.1	Pengenalan	56
5.2	Teknologi Pembangunan	56
5.3	Bahagian Kritikal Aturcara	58
	5.3.1 Kekutuban Emoji / Ikon emosi	58
	5.3.2 Sentimen Keseluruhan Teks	59
	5.3.3 Segmentasi data	59
	5.3.4 Vektorisasi perkataan	60
	5.3.5 Pembangunan Model Pegelas	61
5.4	Ketepatan Algoritma	62
	5.4.1 Menguji Model Pegelas	62
	5.4.2 Skor <i>Naive Bayes</i>	62
	5.4.3 Matriks <i>Confusion</i>	63
	5.4.4 Ketepatan dan Dapatan Balik	63
	5.4.5 Ketepatan Model	65
5.5	Lipatan K Rawak (<i>K- Fold Random Testing</i>)	65

5.6	Perbandingan analisis sentimen yang menggunakan emoji dan tanpa menggunakan emoji	68
5.7	Kesimpulan	69
BAB VI	KESIMPULAN	
6.1	Gambaran Keseluruhan	70
6.2	Kekangan	70
6.3	Pembangunan Kajian Pada Masa Depan	70
6.4	Kesimpulan	71
RUJUKAN		72
LAMPIRAN		
Lampiran A	Skor emoji yang digunakan	74

SENARAI JADUAL

No. Jadual		Halaman
Jadual 1.1	Carta gantt bagi pembangunan projek pada Semester 1	8
Jadual 2.1	Identifikasi emosi asas	14
Jadual 2.2	Ikon emosi barat	16
Jadual 2.3	Ikon emosi timur	20
Jadual 2.4	Ikon emosi 2channel	21
Jadual 2.5	Perbandingan Aplikasi Sedia Ada dan Algoritma Analisis Sentimen	27
Jadual 4.1	Skor kekutuban dan sentimen	51
Jadual 4.2	Jenis perkataan dan peta tag	51
Jadual 4.3	Jenis data dan peratusan	52
Jadual 4.4	Label dan sentimen	53
Jadual 4.5	Matriks Prestasi	54
Jadual 5.1	Lipatan K-Rawak	65
Jadual 5.2	Purata Lipatan K-Rawak Naive Bayes	66
Jadual 5.3	Perbandingan Analisis Sentimen yang menggunakan emoji dan tanpa menggunakan emoji	68

SENARAI ILUSTRASI

No. Rajah		Halaman
Rajah 1.1	Model Kitaran Hayat	6
Rajah 2.1	Langkah - langkah dalam Proses Analisis Sentimen	12
Rajah 2.2	Sistem TweetReach	26
Rajah 2.3	Sistem HootSuite	27
Rajah 3.1	Rajah Konteks untuk Algoritma Analisis Sentimen	32
Rajah 3.2	Rajah Kes Guna untuk Algoritma Analisis Sentimen	33
Rajah 3.3	Rajah Carta Alir Proses Algoritma Analisis Sentimen	34
Rajah 4.1	Rajah Senibina Lambda yang diubahsuai	38
Rajah 4.2	Kekutuban dan Neutraliti Emoji	41
Rajah 4.3	Rajah Carta Alir Lapisan Kelajuan	44
Rajah 4.4	Rajah data sebelum Lapisan Kelajuan	44
Rajah 4.5	Rajah data selepas Lapisan Kelajuan	44
Rajah 4.6	Rajah Carta Alir Lapisan <i>Batch</i> Bahagian Pertama	45
Rajah 4.7	Rajah data sebelum Lapisan <i>Batch</i> Bahagian Pertama	45
Rajah 4.8	Rajah data selepas Lapisan <i>Batch</i> Bahagian Pertama	46
Rajah 4.9	Rajah Carta Alir Lapisan <i>Batch</i> Bahagian Kedua	46
Rajah 4.10	Rajah data sebelum Lapisan <i>Batch</i> Bahagian Kedua	46
Rajah 4.11	Rajah data selepas Lapisan <i>Batch</i> Bahagian Kedua	47
Rajah 4.12	Rajah Carta Alir Lapisan <i>Batch</i> Bahagian Ketiga	47
Rajah 4.13	Rajah data sebelum Lapisan <i>Batch</i> Bahagian Ketiga	48
Rajah 4.14	Rajah hasil Lapisan <i>Batch</i> Bahagian Ketiga	48
Rajah 4.15	Rajah Carta Alir Lapisan Khidmat	49
Rajah 4.16	Rajah data Lapisan Khidmat	49
Rajah 4.17	Rajah hasil Lapisan Khidmat	50

Rajah 5.1	Model Kitar Hayat	57
Rajah 5.2	Kod Kekutuban Emoji / Ikon emosi	58
Rajah 5.3	Kod Sentimen Keseluruhan Teks	59
Rajah 5.4	Kod Segmentasi data	59
Rajah 5.5	Cara data disegmentasi	60
Rajah 5.6	Kod vektorisasi perkataan	60
Rajah 5.7	Kod model pengelas	61
Rajah 5.8	Sentimen teks	62
Rajah 5.9	Skor Naive Bayes	62
Rajah 5.10	Matriks Confusion	63
Rajah 5.11	Ketepatan dan Dapatan Balik	63
Rajah 5.12	Ketepatan dan Dapatan Balik	64
Rajah 5.13	Ketepatan Model	65
Rajah 5.14	Visualisasi Lipatan K-Rawak Naive Bayes	67
Rajah 5.15	Contoh tweet dan perbandingannya	68

SENARAI SINGKATAN

TF - IDF	Term Frequency-Inverse Document Frequency
----------	---

BAB I

PENGENALAN

1.1 PENDAHULUAN

Teknologi Web semakin meningkat dari masa ke semasa. Oleh kerana arus modenisasi ini, bilangan orang yang meluahkan perasaan dan pendapat melalui web semakin meningkat secara drastik. Maklumat ini berguna untuk pihak awam dan juga swasta. Maklumat dan pendapat sangat berguna dalam pengurusan perniagaan, kerajaan dan juga individu. Melalui maklumat-maklumat yang diperoleh pihak tertentu dapat mengetahui sentimen pendapat masyarakat terhadap mereka. Hal ini sangat bermanfaat untuk kejayaan mereka dalam masa akan datang.

Pada zaman dahulu, pihak awam dan swasta perlu bersusah payah untuk mengetahui pendapat masyarakat serata. Hal ini disebabkan, pihak awam dan swasta tidak mempunyai komunikasi yang efektif dengan masyarakat. Pendapat - pendapat masyarakat diambil secara manual. Contohnya, jika pengurusan jenama X ingin mengetahui pendapat masyarakat tentang produk mereka, mereka harus menjalankan kajian atau sesi soal selidik. Sebuah kumpulan harus membuat kajian dan soal-selidik. Kumpulan penyelidik tersebut harus pergi ke tempat-tempat tertentu untuk mengetahui pendapat masyarakat. Selain itu, data yang diperoleh daripada pengumpulan soal selidik hanya sedikit. Pendapat keseluruhan tidak dapat diketahui kerana sesi soal selidik hanya dijalankan kepada segelintir masyarakat.

Pengetahuan tentang pendapat masyarakat tidak lagi menjadi masalah dengan pembangunan web pada masa kini. Penggunaan web oleh pihak masyarakat meningkat sejak 20 tahun yang lepas. Pada zaman sekarang, semakin ramai orang meluahkan perasaan dan pendapat mereka terhadap produk, servis, individu, atau

peristiwa dalam blog atau media sosial mereka. Data ini sangat berguna untuk pembangunan organisasi. Data ini semua dikumpulkan dan dianalisis dengan cara analisis sentimen.

Peningkatan pendapat telah memberi peluang untuk bidang sains data dan analisis menjadi popular. Kaedah kecerdasan pengkomputeraan terbukti untuk menjadi alat persaingan yang penting dalam industri hari ini. Sebagai contoh, dalam membuat analisis sentimen untuk sesuatu perniagaan, corak pendapat masyarakat boleh ditimbang untuk memahami pelanggan, meningkatkan jualan dan pemasaran. Proses analisis sentimen boleh mengenal pasti corak dalam sesebuah data.

Data yang terdapat dalam internet boleh diproses kepada bentuk yang lebih senang untuk dibaca oleh proses analisis sentimen. Proses ini dapat mengenal pasti sentimen sesuatu teks. Terdapat tiga jenis sentimen utama iaitu, positif, neutral dan juga negatif.

1

1.2 PENYATAAN MASALAH

Analisis sentimen adalah topik kajian yang hangat sekarang. Teknik ini digunakan untuk mengenal pasti prestasi sesebuah perniagaan, produk atau sentimen sesebuah maklumat disebarkan, positif dan negatif. Analisis sentimen juga digunakan sebagai domain untuk kewartawanan untuk mengetahui topik yang digemari oleh masyarakat. Walaupun dunia aplikasi analisis sentimen ini besar, penggunaan analisis sentimen ini terhad kepada teks sahaja. Hal ini mungkin memberi konklusi yang tidak tepat dan salah kerana kegunaan alat atau ekspresi yang lain seperti emoji.

Pengguna media sosial seperti *Twitter* dan *Facebook* banyak menggunakan emoji. Terdapat impak yang besar kepada *tweets* dan pendapat yang diluahkan di media sosial dengan penggunaan emoji. Bukan sahaja teks, emoji juga harus diambil kira dalam analisis sentimen. Pada masa sekarang, tidak banyak model atau proses sentimen yang mengambil kira emoji untuk membuat analisis sentimen.

1.3 CADANGAN PENYELESAIAN MASALAH

Analisis sentimen melalui emoji wajar digunakan. Hal ini kerana, emoji menukar konteks sesuatu teks. Teks sahaja tidak mencukupi untuk mengetahui sentimen sesuatu pendapat atau luahan.

Proses sentimen analisis yang berasaskan emoji boleh direalisasikan melalui beberapa teknik. Teknik yang boleh digunakan adalah pengelas pembelajaran mesin seperti rangkaian neural, *Naive-Bayes* dan *Support Vector Machines(SVM)*. Analisis sentimen ini wajar menentukan teknik yang terbaik untuk mengetahui kekutuban yang tepat ayat yang mengandungi emoji.

Dalam analisis ini, teks dan emoji patut dilombong secara berasingan. Emoji ini harus diperinci dengan teliti untuk mengetahui kekutubannya. Selepas itu, pendapat ini dikumpulkan untuk mengetahui sentimen teks tersebut, sama ada positif, neutral atau negatif.

Kajian ini menggunakan kaedah pembelajaran mesin dalam mengelaskan sentimen. Pembelajaran mesin adalah kaedah yang digunakan dalam kajian ini. Kaedah ini akan mempunyai dua dataset yang berasingan, iaitu data untuk melatih mesin untuk belajar dan data yang ingin diuji.

Pembelajaran mesin terbahagi kepada tiga pendekatan iaitu terselia, pendekatan tidak terselia dan pendekatan yang separuh terselia. Pembelajaran mesin ini digunakan dalam banyak bidang seperti visi pengkomputeran, pengecaman pertuturan, pemprosesan bahasa tabii, pengenalan audio, penapisan rangkaian sosial, terjemaaahan mesin, bioinformatik, reka bentuk dalam perubatan, analisis imej perubatan dan pemeriksaan bahan.

1.4 MATLAMAT DAN OBJEKTIF KAJIAN

Terdapat banyak model yang boleh menentukan sentimen sesuatu teks, sama ada positif, neutral atau negatif. Tetapi tidak banyak model berfungsi untuk mengecam dan mengenal pasti konteks sesuatu emoji. Dalam sesuatu teks yang mengandungi penggunaan emoji, emoji juga harus diambil kira untuk memahami konteks teks tersebut dengan keseluruhan. Oleh itu, model analisis sentimen yang berfungsi untuk mengenal pasti sentimen emoji harus dibangunkan. Sehubungan dengan itu **matlamat utama kajian ini ialah untuk mengelaskan kekutuban sentimen teks yang mengandungi emoji.**

Bagi mencapai matlamat ini, objektif berikut digariskan:

- a. Mengenalpasti kekutuban emoji yang digunakan dalam teks media sosial.
- b. Membina set data ujian melibatkan teks yang mengandungi emoji.
- c. Menggunakan algoritma pembelajaran mesin yang sesuai untuk mengelaskan teks yang mengandungi emoji.

1.5 SKOP

Kajian yang dijalankan ini adalah di bawah bidang analisis sentimen. Kajian ini tertumpu kepada emoji yang digunakan dalam teks media sosial berbahasa Inggeris.

Semua data yang digunakan untuk analisis sentimen adalah daripada pengguna media sosial atau blog. Luahan dalam media sosial dan blog adalah tidak formal. Hampir semua pengguna media sosial meluahkan perasaan mereka seperti pertuturan antara kalangan rakan sebaya mereka. Oleh itu, sindiran selalunya digunakan untuk meluahkan perasaan mereka. Mengikut para linguistik, pernyataan sindiran adalah sesuatu ayat atau luahan yang bertentangan dengan apa yang mereka sebenarnya rasa.

Kebanyakan masyarakat menggunakan sindiran untuk mengkritik. Walau bagaimanapun kajian ini tidak mempertimbangkan elemen sindiran dalam sentimen.

1.6 KEPENTINGAN KAJIAN

1.6.1 Skop Pengurusan Perniagaan

Pengurus dapat mengetahui cita rasa dan perasaan pengguna terhadap produk yang dihasilkan. Contohnya, jika pengguna tidak puas dengan bungkusan produk tersebut, pihak pengurus boleh menukar bungkusan tersebut kepada yang lebih menarik dan yang diinginkan oleh pengguna. Pihak pengurus boleh menggunakan analisis sentimen ini untuk mengetahui galakkan yang diperoleh oleh produk tersebut daripada masyarakat. Dengan pendekatan ini, perniagaan mereka akan menjadi lebih berjaya kerana pihak pengurusan boleh mendapat maklumat tentang produk mereka daripada pengguna sendiri.

1.6.2 Skop Individu

Individu-individu yang terkenal dalam bidang politik dan juga dunia hiburan boleh mengetahui apa pendapat masyarakat tentang mereka. Sebagai contoh, seorang pelakon boleh mengetahui pendapat masyarakat tentang filem yang baru dikeluarkan. Selain itu, pelakon- pelakon juga boleh mengenal pasti jenis filem yang diinginkan dan dijangka oleh peminat-peminat mereka.

1.7 CADANGAN PENYELESAIAN MASALAH

Metodologi yang akan digunakan untuk membangunkan algoritma ini adalah Model Kitar Hayat seperti yang ditunjukkan dalam Rajah 1.1. Model Kitar Hayat dipilih untuk mengurangkan risiko kegagalan semasa membangunkan algoritma yang

diingini. Selain itu, model ini mudah untuk difahami dan digunakan. Fasa pembangunan algoritma lebih mudah untuk dipantau.



Rajah 1.1 Model Kitaran Hayat

Model Kitaran Hayat Pembangunan ini mempunyai 5 fasa utama :-

(i) Fasa Pertama

Fasa ini adalah fasa perancangan dan analisis keperluan algoritma. Setiap model kitaran hayat ini bermula dengan analisis, dimana keperluan dan keperluan algoritma dibincang untuk produk akhir yang memenuhi keperluan algoritma. Kepentingan fasa ini adalah untuk mendapatkan definisi yang terperinci terhadap keperluan algoritma. Perancangan algoritma sangat penting untuk memastikan algoritma bertahan lama.

(ii) Fasa Kedua

Fasa kedua adalah untuk merancang seni bina algoritma. Dalam fasa ini, algoritma akan diberi seni bina yang menarik dan juga mudah difahami oleh pengguna atau

pengguna yang disasarkan. Skop kajian dan keperluan perisian dan maklumat diambil kira semasa mereka seni bina algoritma ini.

(iii) Fasa Ketiga

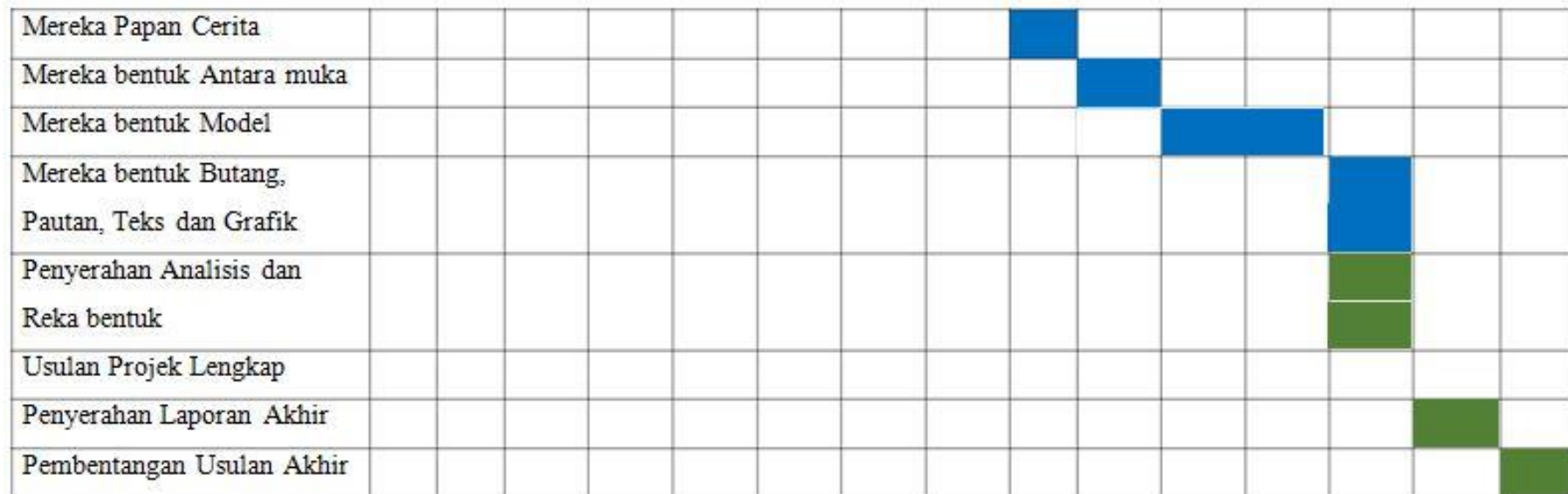
Fasa ini adalah fasa untuk pembangunan dan pengaturcaraan. Semua keperluan algoritma, keperluan perisian dan keperluan perkakasan diambil kira semasa membuat pengaturcaraan. Pada mulanya, perkembangan algoritma diberi keutamaan. Selepas itu, proses penulisan kod sumber dimulakan diikuti dengan penyusunan dan ujian.

(iv) Fasa Keempat

Fasa keempat adalah fasa implementasi dan pelaksanaan algoritma. Ini adalah fasa dimana algoritma diuji berulang kali. Selepas itu, algoritma yang direkabentuk akan dilaksanakan untuk kegunaan pengguna.

(v) Fasa Kelima

Fasa terakhir model kitaran hayat pembangunan algoritma. Proses penyahpejatan juga dilakukan berulang kali untuk memastikan algoritma tidak mempunyai masalah yang kritikal. Proses ini sangat penting untuk membangunkan algoritma yang stabil. Dalam fasa ini, semua masalah yang wujud dalam fasa sebelum ini akan diperbetulkan. Algoritma akan digunakan oleh para pengguna.



Jadual 1.1 Carta gantt bagi pembangunan projek pada Semester 1

BAB II

KAJIAN LITERATUR

2.1 PENGENALAN

2.1.1 Definisi Analisis Sentimen

Definisi perkataan analisis mengikut Kamus Bahasa Melayu adalah 1. Penghuraian atau pengupasan sesuatu perkara untuk mengetahui selok-beloknya (buruk baiknya dll); kajian. (Kamus Pelajar Edisi Kedua).

Definisi perkataan sentimen mengikut Kamus Bahasa Melayu adalah perasaan hati yang berlebih-lebihan terhadap sesuatu sehingga berlawanan dengan akal fikiran. (Kamus Pelajar Edisi Kedua).

Analisis sentimen adalah sesuatu proses untuk membuat pertimbangan yang teliti terhadap sesuatu teks atau pendapat, untuk mengetahui perasaan penulis teks atau pendapat tersebut. Proses ini mengandungi mengidentifikasikan, mengumpul dan mengelaskan teks/pendapat yang ditulis untuk mengetahui perasaan penulis terhadap produk, servis atau yang lain. Perasaan terutamanya terbahagi kepada tiga kategori iaitu, positif, neutral dan negatif.

Analisis sentimen, biasanya dikenali sebagai perlombongan pendapat, adalah satu bidang pengajian yang menganalisis pendapat, sentimen, penilaian, sikap dan emosi manusia terhadap sesuatu entiti, seperti produk, servis, organisasi, individu, topik hangat semasa, dan peristiwa / acara. Terdapat banyak variasi nama untuk analisis sentimen seperti perlombongan pendapat, perlombongan sentimen, pengekstrakan pendapat, analisis emosi, dan lain-lain. Walau bagaimanapun, semua istilah ini boleh dipermudahkan dengan nama analisis sentimen. Dalam industri, istilah analisis

sentimen digunakan secara berlebihan berbanding dengan istilah yang lain. (Bing Liu, 2012). Perkataan analisis sentimen muncul dahulu dalam (Nawasuka dan Yi, 2003).

Walaupun, para linguistik dan teknik pemprosesan bahasa tabii wujud sejak zaman dahulu lagi, hanya sedikit penyelidikan terhadap pendapat dan sentimen manusia dijalankan sebelum tahun 2000. Sejak tahun 2000, bidang ini telah menjadi lebih popular dan aktif. Hal ini disebabkan oleh pelbagai aplikasi dalam setiap domain. Industri ini berkembang disebabkan percambahan aplikasi komersial. Selain itu, sekarang adalah era pertama kita memperoleh pendapat dan perasaan manusia dengan jumlah yang banyak melalui data dalam media sosial. Hal ini, memberi motivasi untuk para penyelidik untuk menyelidik sentimen. Bukan itu sahaja, analisis ini banyak digunakan dalam kajian, perlombongan web dan teks. Penggunaan analisis sentimen telah merebak ke sektor lain daripada sektor sains komputer seperti sains pengurusan dan sains sosial disebabkan kepentingannya untuk perniagaan dan masyarakat.

2.1.2 Proses Analisis Sentimen



Rajah 2.1 Langkah- langkah dalam proses analisis sentimen

Rajah 2.1 menunjukkan aliran kerja yang umum untuk proses analisis sentimen yang mana perbincangannya seperti berikut:-

- a. Matlamat harus ditetapkan untuk memulakan proses analisis sentimen. Proses penetapan matlamat merangkumi penentuan misi analisis sentimen dan skop untuk kandungan teks.

- b. Pemprosesan teks harus dijalankan. Proses ini adalah proses yang penting dalam analisis sentimen. Sumber teks dan kandungan tersebut akan ditentukan, seperti data ini daripada web, blog atau media sosial. Selepas itu, teks akan dimuatkan dalam algoritma untuk pemprosesan, algoritma ini digunakan untuk membuat analisis. Perkataan – perkataan yang berulang dan tidak diperlukan dibuang. Simbol-simbol emosi atau emoji ditukarkan kepada perkataan dan seterusnya disusun dalam carta. Proses ini juga memperhatikan sentimen yang diluahkan, abjad – abjad dalam huruf besar seperti GEMBIRA/ CANTIK.
- c. Langkah seterusnya adalah mengkaji hurai kandungan yang merangkumi segmentasi perkataan berdasarkan kepada positif, neutral dan negatif. Penandaan segmen teks yang digunakan berdasarkan definisi dan konteks.
- d. Penapisan teks adalah untuk memastikan analisis ini betul. Proses ini untuk menidentifikasi perkataan yang dikeluarkan semasa melakukan pemprosesan bahasa tabii dan sinonimnya.
- e. Langkah terakhir untuk analisis sentimen adalah analisis dan kiraan mata. Proses ini merangkumi penentuan sentimen. Kiraan mata adalah proses yang ketinggian sentimen dalam teks yang dianalisis.

2.1.3 Analisis Sentimen dan Analisis Emosi

Analisis sentimen, yang juga dikenali sebagai perlombongan pendapat, bertumpukan kepada pengenalanpastian corak dalam teks untuk mengklasifikasikan kepada sentimen yang tepat. Analisis sentimen kebanyakannya dilaksanakan dalam perisian untuk mengekstrak emosi dan pendapat daripada teks secara autonomi. Analisis sentimen mempunyai banyak aplikasi dunia nyata seperti analisis sentimen membolehkan pihak pengurusan perniagaan menganalisis bagaimana produk atau jenama dirasakan oleh pengguna mereka dan ahli politik mungkin berminat untuk mengetahui bagaimana para masyarakat merancang untuk mengundi dalam pilihan raya. Analisis sentimen ini sukar untuk diklasifikasikan ke dalam satu bidang kerana bidang pembelajaran ini

merangkumi pelbagai bidang seperti bidang linguistik, pemprosesan bahasa tabii, pembelajaran mesin dan kecerdasan buatan. Kebanyakan sentimen yang dimuat naik dalam Internet adalah data yang tidak tersusun. Hal ini menyebabkan pemprosesan dan pengekstrakan maklumat yang bermakna sukar dilakukan. Kebanyakan algoritma pembelajaran mesin yang efektif seperti *Support Vector Machines* dan *Naïve Bayes* mengeluarkan keputusan yang tidak boleh dibaca dan difahami oleh manusia (Wieslaw Wolny).

Emosi berkait rapat dengan sentimen. Emosi ditakrifkan sebagai perasaan dan fikiran yang subjektif. Emosi seseorang boleh dikategorikan dalam beberapa kategori yang berbeza. Analisis emosi adalah salah satu lapisan analisis sentimen. Walau bagaimanapun, tiada set data yang mengklasifikasikan emosi dalam kalangan pengkaji.

Seseorang individu mempunyai enam emosi utama, iaitu, cinta, gembira, teraju, marah, sedih dan takut. Emosi- emosi utama ini boleh dibahagikan kepada emosi – emosi yang lain. Setiap emosi mempunyai tahap keamatan yang berbeza (W Parrott).

Emosi dalam dunia komunikasi maya berbeza daripada komunikasi bersemuka. Hal ini kerana, ciri-ciri komunikasi berasaskan komputer tidak mempunyai isyarat auditori dan visual yang biasanya dikaitkan dengan aspek emosi. Manakala, komunikasi yang berasaskan teks menyingkirkan isyarat audio dan visual kerana teks mempunyai cara lain untuk menunjukkan emosi. Emotikon dan ikon emosi, boleh digunakan untuk meluahkan pelbagai variasi emosi.

Jadual 2.1 Identifikasi emosi asas

Ahli Teori	Emosi Asas
Plutchik	Penerimaan, kemarahan, antisipasi, jijik, kegembiraan, ketakutan, kesedihan, kejutan
Arnold	Kemarahan, keengganan, keberanian, kecewa, keinginan, putus asa, ketakutan, benci, harapan, cinta, kesedihan



















Ekman, Friesen dan Ellsworth	Kemarahan, jijik, ketakutan, kegembiraan, kesedihan, kejutan
Frijda	Keinginan, kebahagiaan, minat, kejutan, keajaiban, kesedihan
Gray	Kemarahan dan ketakutan, kegelisahandan kegembiraan
Izard	Kemarahan, penghinaan, jijik, kesusahan, ketakutan, rasa bersalah, minat, kegembiraan, malu dan terkejut
James	Ketakutan, kesedihan, cinta, kemarahan
McDougall	Kemarahan, jijik, ketenangan, ketakutan, penundaan, emosi tender, tertanya-tanya
Mowrer	Kesakitan, keseronokan
Oatley dan Johnson-Laird	Kemarahan, jijik, kebimbangan, kebahagiaan, kesedihan
Panksepp	Antisipasi, ketakutan, kemarahan, panik
Parrott	Cinta, kegembiraan, kejutan, kemarahan, kesedihan, ketakutan
Tomkins	Kemarahan, minat, penghinaan, jijik, kesusahan, ketakutan, kegembiraan, rasa malu, kejutan
Watson	Ketakutan, cinta, kemarahan
Weiner dan Graham	Kebahagiaan, kesedihan


















Jadual 2.1 menunjukkan teori daripada pelbagai ahli teori untuk identifikasi emosi asas. Sesetengah konsep bermatlamat untuk mengurangkan jumlah emosi asas. Untuk kajian ini, emosi dapat diklasifikasikan menggunakan ikon emosi dan emotikon yang digunakan.

















Emotikon boleh dibahagikan kepada tiga jenis, iaitu, emotikon barat, emotikon timur dan juga emotikon gaya 2channel. Huraian yang didapati dalam perbezaan emotikon barat dan timur adalah emotikon timur berfokuskan ekspresi mata manakala emotikon barat berfokus kepada seluruh wajah muka (Wikipedia).






(i) Ikon Emosi Barat

Jadual 2.2 Ikon emosi barat

Sentimen	Emosi	Ikon Emosi	Emoji
Positif	Gembira	:-) :-] :] :) :-3 :3 :-> :> 8-) 8) :-} :} :o) :c) :^) =] =)	    
	Ketawa, Senyuman besar, Senyuman dengan mata hati	:~D :D x-D xD X-D XD =D =3	   
	Air mata kebahagiaan	:!-) :')	
	Mengenyit, Senyuman	;~) ;) *-) *) ;D :-, ;^)	  
	Cium	:-* :* :×	    

	Malaikat, Tidak Bersalah	O:-) O:) 0:-3 0:3 0:-) 0:) 0;^)	 
Neutral	Terkejut	:~O :O :-o :o :-0 8-0	  
	Bermain-main	:~P :P X-P XP x-p xp :-p :-P :P :p :-p :p :-b :b	   
	Ragu-ragu, Tidak malu, Tidak selesa, Teragak-agak	:~/ :/ :-> \ =/ =\	  
	Kebimbangan, Tanpa ekspresi	:~ :	 
	Rasa malu	://) ://3 :\$	  

	Tiada apa	:X :X :-# :# :-&:&	 
	Keliru	%-) %)	  
	Keraguan, Ketidak- percayaan	':- ':-l	
Negatif	Berkerut dahi, sedih, marah	:- (: (:-c :c :-< :< :-[:[:- >[:{ :@ :(       
	Menangis	:'- (:'( 

Seram, Jijik, D-' : D:< D: D8 D; D= DX Sedih, Kecewa	  
Sedang sakit :-### :###	 
Jahat >:-) :) } :-) } :) 3:-) 3 :)	
Bosan, -O Menguap	

Jadual 2.2 menunjukkan ikon emosi yang bergaya barat. Ikon emosi barat selalunya ditulis dari kiri ke kanan. Selalunya, bermula daripada mata daripada kiri, seterusnya dengan hidung. Hidung dalam ikon emosi selalunya tidak dimasukkan. Titik bertindih selalunya digunakan untuk mata, manakala jika mengenyit koma bertitik digunakan. Ada masanya 8 dan juga = digunakan untuk mata. Simbol yang digunakan untuk mulut adalah) untuk gembira dan (untuk kesedihan. Simbol } untuk janggut.

(ii) Ikon Emosi Timur

Jadual 2.3 Ikon emosi timur

Sentimen	Emosi	Ikon Emosi
Positif	Mengenyit	(^_-) (^_-)-☆
	Gembira	^_^ (°o°) (^_^)/ (^O^)/ (^o^)/ (^^)/ (≥∇≤)/ (/●ㄣ●)/ (^o^) J ∩(·ω·)∩ (·ω·) ^ω^
	Ketawa	^m^
	Gelak	>^_^< <^!^> ^/^ (*^_^*) §^.^§ (^<^) (^.^) (^Λ^) (^.^) (^.^) (^_.) (^_^) (^^) (^J^) (*^.^*) ^_^ (#^.^#) (^—^)
	Teruja	\(~o~)/ \ (^o^)/ \ (-o-) / \ (^_ ^) J \ (^o^) J (*^0^*)
	Kagum	(*_*) (*_*; (+_+)) (@_@)(@_@(@_@;) \(◎o◎)/ !
	Gelak, Gembira	(*^^)v (^^)v (^_^)v ('_')* (^ v ^) (^ ∇ ^) (· ∇ ·) (^ v ^) (∩ ∇ ∩)
	Gembira	(● ^ o ^ ●) (^ v ^) (^ u ^) (^ ◇ ^) (^) o (^) (^ O ^) (^ o ^) (^ o ^) ^ o ^ (* ^ ∇ ^ *) (☼ ^ _ ^)
Neutral	Senyum, Suka, Gembira	(*^∇`*) (*°∇°) =3 Uwu UwU
	Cium	(*^3^)/~☆
	Bertanya	\(°□\)(/□°)/
	Letih	(= _ =)

Negatif	Terkejut	($\overline{\square}$;) °o° °O° :O o_O o_0 o.O (o.o) oO (° ㄣ°) (◊◊°)
	Tidak Pasti	ヽ('—`) ㄱ ㄴ (ツ) ㄴ
	Bermasalah	(>_<) (>_<)>
	Takut, Malu, Berpeluh	(^^ ㄹ (^_ ^;) (-_-;) (~_~;) (. . .;) (. _ .;) (. . .;) ^.^; ^_ ^; (#.^.#) (^^;)
	Bosan	(-_-)zzz
	Keliru	((+_+))(+o+) (°°) (°-°) (°.°) (°_°) (°_°>) (° ㄴ°)
	Sedih, Menangis	('_) (/_;) (T_T) (;_;) (;_; (;_;) (;O;) (:_;) (ToT) (ㄸ ㄷ ㄸ) ;_; ;-; ;n; ;; Q.Q T.T TnT QQ Q_Q
	Tidak puas	(* ㄴ m ㄴ)

Jadual 2.3 menunjukkan ikon emosi yang bergaya timur. Ikon emosi ini diputar seperti ikon emosi barat. Ikon – ikon ini mula-mulanya digunakan di Jepun. Ikon emosi ini dipanggil kaoemoji yang membawa maksud ikon emosi muka.

(iii) Ikon Emosi 2channel

Jadual 2.4 Ikon emosi 2channel

Sentimen	Emosi	Ikon Emosi
Positif	Menghormat	m(_ _)m
	Gembira	(`ω´) (° ㄣ°) ♪ ㄱ (. o .) ㄴ ♪ ㄴ (. o .) ㄴ d(* ㄴ ㄷ ㄴ *)b ヽ (ㄷ ㄴ) / ^ ㄴ ^
	Tabik	(` - ´)>

Neutral	Tenang	$\backslash(' - \text{`})/$
	Rasa bebas	$(\forall \text{`})$
	Peramah	$\backslash(' - \text{`})\text{人}(\nabla \text{`})\text{人}(\text{D}')/$
	Terkejut	$\Sigma(^{\circ} \text{D}^{\circ} ;) (^{\circ} \text{D}^{\circ}) \Sigma(^{\circ} \text{D}^{\circ})$
	Tidak tahu jawapan	$\neg(' \sim \text{`} ;) \Gamma$
	Gerak hati/ Intuisi	$m9(\cdot \forall \cdot)$
	Berfikir	$(\text{`} - \text{`}) \text{.oO}(\dots)$
	Tidak Sabar	$(^{\circ} \text{D}^{\circ} ; \equiv ; ^{\circ} \text{D}^{\circ})$
	Tiada ekspresi	$(\overline{\text{`} - \text{`}})$
	Tidak peduli	$(\text{`} \cdot \omega \cdot \text{`}) [^{\circ} \text{D}^{\circ}]$
Negatif	Sedih	$(\text{`} ; \omega ; \text{`}) (\text{つD`})$
	Marah / Geram	$\backslash(\text{D}')/ \quad (\#^{\circ} \text{D}^{\circ})$ $\backslash(\text{o`III'o})/ \quad ($ $\text{D`}) (\geq \square \leq)$
	Takut / Tekanan Emosi	$(((((; ^{\circ} \text{D}^{\circ}))) (\text{A`})(\text{`} - \text{`}))$
	Tidak bagus	$(\cdot \forall \cdot) \quad (\cdot \text{A} \cdot)$
	Bosan	$(\Theta \varepsilon \Theta ;)$

Ikon emosi dalam Jadual 2.4 mula-mula digunakan dalam 2channel, iaitu halaman diskusi Jepun.

2.1.4 Kepentingan Analisis Sentimen

Kepentingan analisis sentimen adalah seperti berikut:-

- Melaraskan strategi pemasaran

Daripada perspektif pengurusan, media sosial adalah platform untuk mempromosikan servis. Media sosial bukan sahaja tempat untuk mempromosikan servis tetapi tempat pengguna dan pelanggan bergaul dan berinteraksi tentang jenama tersebut. Pergaulan dan perinteraksian pelanggan adalah maklumat tentang persepsi pelanggan tentang servis tersebut. Maklumat sebegini diberi daripada proses analisis sentimen. Maklumat-maklumat yang didapati selepas proses analisis sentimen boleh digunakan untuk melaraskan dan mengoptimumkan strategi pemasaran servis.

Daripada perspektif taktikal, pengurusan jenama boleh membuat kempen pemasaran jangka masa pendek yang mengikuti cita rasa pelanggan. Melalui analisis sentimen yang berterusan, kempen tersebut boleh diselaraskan untuk menarik perhatian lebih banyak pelanggan.

b) Mengukur ROI untuk kempen pemasaran

Kejayaan kempen pemasaran bukan sahaja diukur melalui bilangan likes, comment, share dan post di media sosial. Kejayaan kempen pemasaran juga diukur melalui bilangan diskusi dan pergaulan positif. Melalui analisis sentimen, pergaulan dan interaksi positif atau yang negatif boleh dipantau dalam kalangan masyarakat. Dengan menggabungkan ukuran kualitatif dan kuantitatif, ROI dapat diketahui. ROI adalah *Return of Investment* yang bermaksud ukuran yang digunakan untuk mengukur dan menilai kecekapan pelaburan.

c) Menambahbaik kualiti produk

Analisis sentimen boleh digunakan untuk membuat kajian pemasaran melalui mengetahui pendapat pengguna tentang produk/servis, bagaimana kualiti dapat diperbaiki dan ditambahbaik mengikut citarasa pengguna. Produk bukan hanya dinilai melalui fungsi produk tersebut dan adakah produk tersebut memenuhi fungsi tersebut. Produk dinilai daripada semua perspektif seperti kecantikan pakej pembungkusan, promosi-promosi yang diberikan untuk produk tersebut dan

harga produk yang tidak melampau. Kualiti produk dapat diperbaiki daripada pendapat yang diberi oleh pengguna.

2.1.5 Metodologi untuk Analisis Sentimen

Terdapat dua jenis metodologi yang boleh digunakan untuk proses analisis sentimen. Pendekatan tak terselia dan pendekatan terselia adalah dua metod yang paling digunakan untuk proses ini.

a. Pendekatan tak terselia

Pendekatan tak terselia adalah algoritma pembelajaran mesin yang digunakan untuk mendapat inferens daripada data tanpa melatih mesin dengan set data. Antara contoh penggunaan pendekatan tak terselia adalah dalam kamus/leksikon.

Metod ini menggunakan kombinasi perkataan yang diberi penjelasan dan dikategorikan sebagai positif, neutral dan negatif. Dengan menggunakan perkataan-perkataan ini metod ini akan menentukan konteks dan sentimen teks tersebut. Pendekatan ini tidak memerlukan data untuk melatih mesin.

b. Pendekatan terselia

Pendekatan terselia adalah algoritma pembelajaran mesin yang digunakan untuk mendapat konklusi selepas melatih mesin dengan set data. Pendekatan terselia menggunakan teknik pembelajaran mesin khususnya melibatkan penghasilan pengelas yang berciri untuk mentafsir dan mengenali sentimen teks. Penghasilan pengelas yang tepat memerlukan data untuk melatih mesin untuk menimba ilmu. Pengelas yang boleh digunakan untuk pendekatan pembelajaran mesin ini adalah *Support Vector Machine (SVM)*, *Neural Network*, *Maximum Entropy* dan *Naïve Bayes*.

- *Naïve Bayes* - *Naïve Bayes* adalah model yang menggunakan kebarangkalian tetapi dalam masa yang sama mempunyai kebebasan yang tinggi untuk merumuskan andaian.
- *Support Vector Machine (SVM)* - SVM adalah model pendekatan terselia yang melibatkan pembelajaran algoritma yang menganalisa data untuk analisis klasifikasi dan regrasi.
- *Neural Network* - *Neural Network* adalah deretan algoritma yang mengidentifikasi hubungan antara set data yang wujud melalui proses yang selari dengan cara otak manusia berfungsi.
- *Maximum Entropy* - *Maximum Entropy* adalah model yang mengklasifikasikan data menggunakan kebarangkalian.

2.1.6 Aplikasi Analisis Sentimen sedia ada

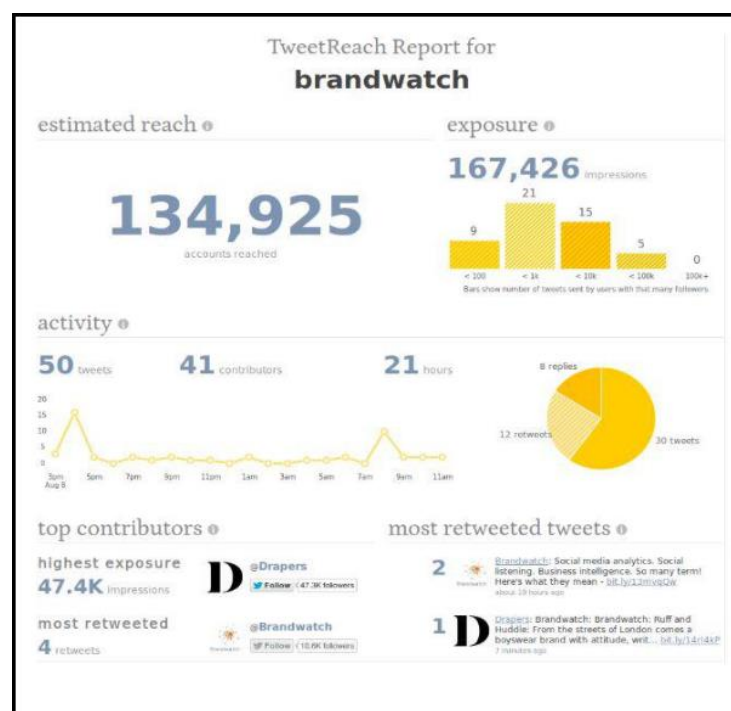
Aplikasi utama analisis sentimen adalah *Social Media Monitoring (SMM)*. Media sosial adalah tempat pengguna servis/produk meluahkan perasaan dan pendapat mereka. Kekurangan dalam media sosial adalah mesej, status, tweet banyak menggunakan kata-kata singkatan dan emoji. Kebanyakan servis analisis sosial tidak dapat mentafsir pendapat/luahan sebegini. Oleh itu, pemprosesan bahasa tabii dan pembelajaran mesin digunakan untuk menukar teks yang mengandungi dialek, tatabahasa yang tidak betul, kata-kata singkatan kepada data yang berstruktur, dapat difahami dan bermanfaat.

Aplikasi analisis sentimen adalah dalam *People Analytics & Voice of Employee & Voice of Employee*. Analisis sentimen digunakan untuk meningkatkan penglibatan pekerja dan meningkatkan produktiviti. Program-program dalam lingkungan rangka kerja ini mengumpul, menganalisis dan mentafsir maklum balas dan pendapat pekerja untuk mengetahui asas masalah pekerja dan sebab pekerja hilang semangat.

Aplikasi analisis sentimen seterusnya adalah *Voice of Customer (VoC) and Customer Experience Management*. Aplikasi ini adalah untuk mentransformasikan maklum balas/ pendapat pengguna yang tidak berstruktur kepada data yang berstruktur dan bermanfaat. Hal ini, dapat memberi pendekatan kepada syarikat- syarikat yang menghasilkan produk atau servis tentang pendapat pengguna mereka terhadap produk-produk yang dikeluarkan dan servis yang diberi. Pendekatan sebegini dapat menghasilkan produk dan servis yang lebih berkualiti dan memenuhi cita rasa pengguna pada masa yang akan datang.

TweetReach

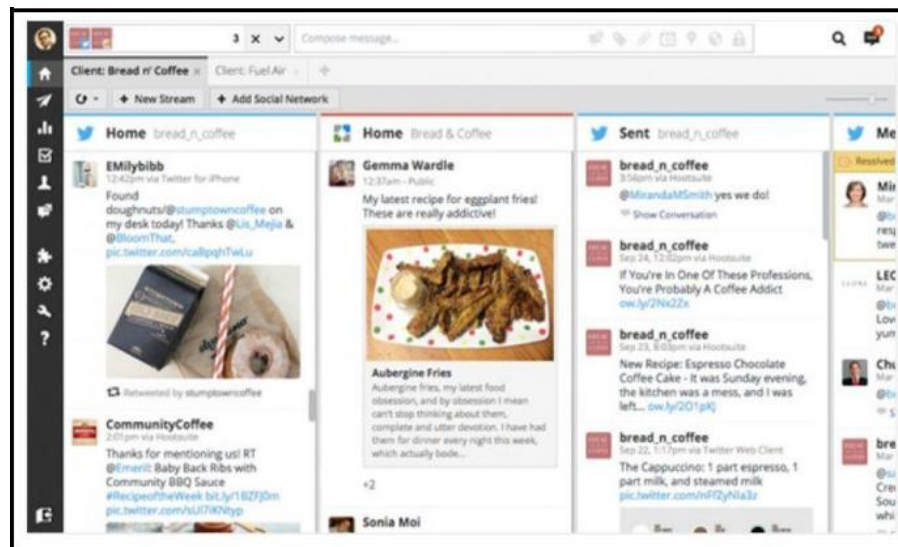
TweetReach adalah alat pemantauan untuk perniagaan anda jika berminat untuk sejauh mana laman-laman dapat sambutan. TweetReach mengukur kesan dan implikasi sebenar perbincangan media sosial.



Rajah 2.2 Sistem TweetReach

HootSuite

HootSuite adalah salah satu alat pendengaran media sosial dan meliputi pelbagai rangkaian sosial seperti *Facebook*, *Twitter*, *Instagram*, *LinkedIn*, *WordPress*, *FourSquare* dan *Google+*. HootSuite terkenal dengan fungsi pengurusan media sosial.



Rajah 2.3 Sistem HootSuite

2.1.7 Perbandingan Aplikasi Analisis Sentimen Sedia Ada dan Algoritma Analisis

Jadual 2.5 Perbandingan Aplikasi Sedia Ada dan Algoritma Analisis Sentimen

Aplikasi	TweetReach	HootSuite	Algoritma yang ingin dibangunkan
Kegunaan	Berfokuskan kepada Twitter sahaja.	Berfokuskan kepada media sosial.	Berfokuskan mana-mana domain yang ingin dianalisis sentimen.
Kos	Mempunyai pelan berbayar untuk pengurus sosial, pemasar sosial, pengurus pemasaran.	Semua pengguna boleh menggunakan aplikasi ini untuk 30 hari. Penggunaan selanjutnya harus dibayar.	Semua pengguna boleh muat turun algoritma ini. Algoritma ini boleh digunakan dengan percuma.

Kemudahan penggunaan	Mesra pengguna. Aplikasi ini mudah difahami oleh pengguna.	Mesra pengguna dan boleh membuat analisis sentimen sebaik sahaja mendaftar dalam platform ini.	Mesra pengguna dan boleh digunakan sebaik sahaja algoritma ini diperoleh.
Kelebihan	Mempunyai khidmat secara langsung dengan <i>Twitter</i> .	Mempunyai perkhidmatan yang efektif untuk analisis sentimen.	Mempunyai algoritma yang boleh membuat analisis sentimen kepada mana-mana teks media sosial.
Kekurangan	Terhad kepada media sosial <i>Twitter</i> sahaja.	Mempunyai fungsi-fungsi yang mungkin tidak dapat difahami oleh pengguna.	Tidak mempunyai khidmat secara langsung dengan mana-mana media sosial.

2.2 KESIMPULAN

Dalam bab ini, analisis sentimen teks dan analisis sentimen ikon emosi dan emoji diterangkan dengan lanjut. Analisis sentimen ayat-ayat media sosial selalunya mengandungi emoji. Emoji memainkan peranan penting dalam penentuan konteks dan kekutuban ayat tersebut sama ada ia positif, negatif atau neutral. Oleh itu, setiap jenis ikon diterangkan dalam bab ini secara mendalam supaya setiap ikon emosi dan emoji dapat ditangkap dan difahami semasa membuat latihan pembelajaran mesin.

BAB III

SPESIFIKASI KEPERLUAN

3.1 PENGENALAN

Dalam bab ini, spesifikasi keperluan dibincangkan dengan lanjut. Spesifikasi keperluan terbahagi kepada dua, iaitu, spesifikasi keperluan pengguna dan spesifikasi keperluan algoritma. Selain itu, spesifikasi keperluan dan perkakasan membincangkan tentang keperluan yang tersedia untuk membangunkan algoritma ini. Bukan itu sahaja, rajah – rajah yang berkaitan dengan pembangunan algoritma ini terdapat dalam bab ini.

3.2 SPESIFIKASI KEPERLUAN PENGGUNA

Pengguna algoritma ini adalah sesiapa yang ingin mengetahui sentimen sesuatu teks, iaitu, pendapat pengguna terhadap produk, pendapat untuk filem-filem terbaharu, dan perasaan orang lain terhadap produk/ servis yang diberi. Dengan penggunaan algoritma ini, pengguna boleh mendapat sentimen tentang domain yang dianalisis dengan tepat iaitu, positif, neutral atau negatif. Antara keperluan pengguna adalah :-

- a) Pengguna harus mempunyai algoritma dan pangkalan data untuk sentimen emoji ini sebelum melaksanakan analisis sentimen.
- b) Pengguna harus mempunyai data sedia ada yang mereka ingin mengetahui sentimen.

- c) Seterusnya, teks akan dianalisis dan algoritma akan menunjukkan sentimen teks tersebut mengikut kekutuban teks dan juga kekutuban ikon emosi atau emoji yang digunakan.

3.3 SPESIFIKASI KEPERLUAN ALGORITMA

3.3.1 Keperluan Bukan Fungsian

Keperluan bukan fungsian adalah keperluan algoritma yang menentukan kriteria yang boleh digunakan untuk operasi algoritma. Berikut adalah keperluan bukan fungsian :-

Keberkesanan

- Algoritma ini akan mencapai tugas atau menghasilkan hasil yang diperlukan. Pengguna boleh mengetahui sentimen sesebuah domain/teks, algoritma ini akan menganalisis kekutuban teks dan juga kekutuban emoji / ikon emosi dan akan memaparkan sentimen setiap teks yang dianalisis.

Kecekapan

- Algoritma ini akan melaksanakan tugas dengan sebaik mungkin. Tumpuan diberi kepada hasil yang maksimum dalam masa yang singkat. Dengan menggunakan model pengelas yang sesuai, algoritma ini membolehkan pengguna mendapat analisis dengan cepat dan dalam masa nyata.

Kebolegunaan

- Algoritma ini mudah difahami dan mesra pengguna. Pengguna yang menggunakan algoritma ini pada kali pertama juga dapat memahami algoritma ini. Setiap bahagian algoritma akan mempunyai komen untuk kemudahan pengguna.

3.3.2 Keperluan Fungsian

Keperluan fungsian sangat penting untuk pembangunan. Keperluan fungsian memastikan algoritma berfungsi berasaskan keperluan yang ingin dicapai. Keperluan fungsian juga memastikan algoritma berjalan dengan lancar dan menepati matlamat.

a. Pengguna

Pengguna utama algoritma ini akan menggunakan algoritma dan pangkalan data untuk sentimen emoji / ikon emosi untuk menganalisis domain yang ingin dianalisis sentimen. Keperluan fungsian pengguna adalah seperti berikut:-

Memuat naik domain/dokumen yang ingin dianalisis

- Algoritma ini memerlukan domain / dokumen / teks yang ingin ditentukan sentimen oleh pengguna untuk membuat analisis sentimen. Pelbagai jenis teks boleh dianalisis melalui algoritma ini, teks yang biasa dan juga teks yang mempunyai emoji / ikon emosi.

Memuat turun domain/dokumen yang telah dianalisis

- Selepas menganalisis dengan sepenuhnya, domain / dokumen / teks yang mempunyai sentimen setiap teks tersebut boleh dimuat turun jika perlu.

3.4 SPESIFIKASI KEPERLUAN PERKAKASAN DAN PERISIAN

Pembangunan sesebuah algoritma yang boleh berfungsi dengan lancar memerlukan dua spesifikasi utama iaitu perkakasan dan perisian yang tepat.

3.4.1 Keperluan Perkakasan dan Perisian Algoritma

Keperluan perkakasan pembangunan algoritma adalah seperti berikut :-

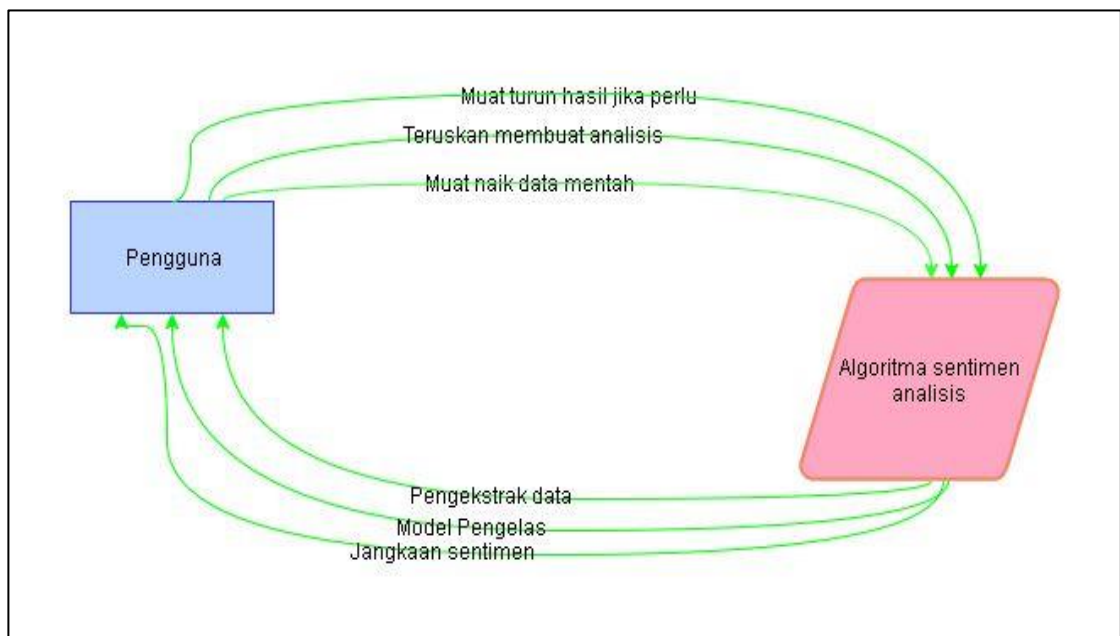
- CPU yang mempunyai 1.6GHz atau ke atas.
- Intel Core i7 6700HQ
- Ruang ingatan dalaman 8GB atau keatas
- 8 GB RAM

Keperluan perisian pembangunan algoritma adalah seperti berikut :-

- Windows 8 – Sistem pengoperasian perkakasan komputer
- Python 3
- Google Colaboratory

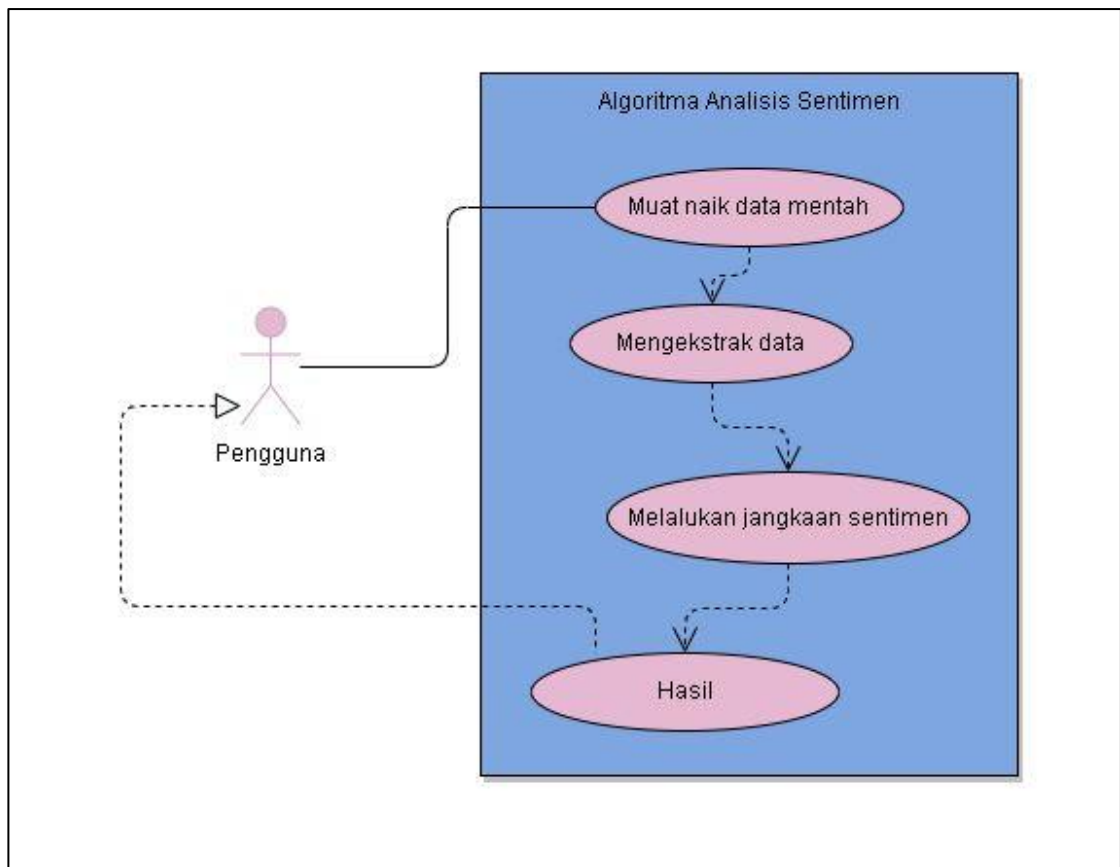
3.5 MODEL ALGORITMA

1. Rajah konteks untuk algoritma analisis sentimen yang berfokuskan emoji.



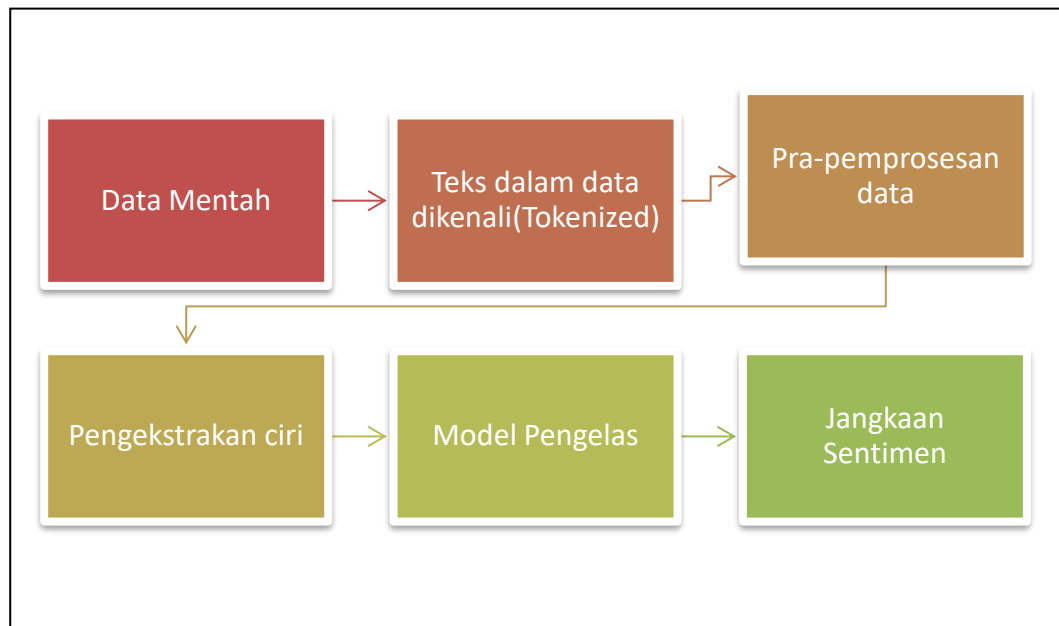
Rajah 3.1 Rajah Konteks untuk Algoritma Analisis Sentimen

2. Rajah kes guna untuk algoritma analisis sentimen yang berfokuskan emoji



Rajah 3.2 Rajah Kes Guna untuk Algoritma Analisis Sentimen

3. Carta alir proses analisis sentimen secara ringkas untuk algoritma ini.



Rajah 3.3 Rajah Carta Alir Proses Algoritma Analisis Sentimen

3.6 KESIMPULAN

Dalam bab ini, keperluan bagi aspek perkakasan dan juga perisian dinyatakan. Gambar rajah konteks, gambar rajah kes guna dan gambar rajah carta aliran proses untuk algoritma analisis sentimen ini ditunjukkan. Kedua-dua rajah konteks dan kes guna memudahkan pemahaman terhadap pembangunan algoritma ini yang berasaskan analisis sentimen yang mengambil kira kekutuban ikon emosi / emoji.

BAB IV

SPESIFIKASI REKABENTUK

4.1 PENGENALAN

Bab ini membincangkan tentang spesifikasi reka bentuk dengan terperinci. Reka bentuk seni bina dan reka bentuk algoritma yang sesuai digunakan untuk kajian ini dibincang dan ditentukan. Reka bentuk – reka bentuk ini amat penting untuk pembangunan. Reka bentuk perisian adalah antara satu proses fasa analisis keperluan. Aspek pembangunan perisian yang penting adalah reka bentuk. Hal ini kerana, idea asal dan cita rasa yang diperlukan dilakukan dalam peringkat reka bentuk dan bukan semasa peringkat pembinaan perisian. Reka bentuk yang teliti dan tepat akan menghasilkan sebuah sistem yang memenuhi keperluan dan yang berkualiti.

4.2 REKABENTUK SENIBINA

Takrifan dan pemodelan seni bina yang khusus untuk proses analisis data besar, seperti yang dihasilkan oleh rangkaian sosial, kini masih pada peringkat awal pembangunan dan penyatuannya. Tidak seperti gudang data tradisional atau sistem perisian perniagaan, seni bina yang direka untuk data berstruktur, algoritma yang didedikasikan untuk kerja data besar dan bukannya data separa berstruktur, atau "data mentah" yang disebut, tanpa struktur tertentu. Ia juga harus dinyatakan bahawa algoritma sepatutnya membolehkan proses pemprosesan dan analisis data bukan sahaja dalam mod batch, tetapi juga dalam cara sebenar masa sebenar.

Kini, sejumlah besar data yang dihasilkan oleh rangkaian sosial setiap hari boleh diproses dan dianalisis untuk tujuan yang berbeza. Data ini disediakan dengan beberapa ciri, antaranya ialah :-

- Dimensi
- Keanehan
- Sumber
- Kebolehpercayaan

Keperluan untuk memperoleh maklumat dan cara maklumat ini mesti diproses. Data harus diproses terlebih dahulu dan kemudiannya tersedia, tanpa mengira aspek masa. Jenis pemprosesan ini biasanya disebut pemprosesan batch. Kini, jumlah data meningkat secara eksponen dan pemprosesan masa nyata diperlukan untuk mendapatkan kelebihan maksimum daripada data ini.

Sebenarnya model kelompok tidak membenarkan untuk bekerja dengan data dalam cara masa sebenar, kerana masa yang lama diperlukan oleh operasi pemprosesan. Terhadap pelaksanaan seni bina pemprosesan masa nyata boleh membawa kepada ketepatan yang lebih rendah. Satu penyelesaian yang mungkin adalah untuk menggabungkan dua konsep menjadi satu seni bina tunggal, yang mampu mengendalikan data besar, tetapi juga dengan ciri pemprosesan yang berskala dan cepat.

Penyelesaian yang mungkin untuk masalah ini ialah senibina Lambda (Marz dan Warren, 2015), senibina perisian yang dibuat oleh 3 tahap yang berbeza seperti ditunjukkan dalam Rajah 4.1:

- Lapisan kelajuan
- Lapisan *Batch*
- Lapisan khidmat

Fungsi setiap lapisan senibina lambda diubahsuai untuk memenuhi kehendak kajian dan algoritma ini

(a) Lapisan Kelajuan

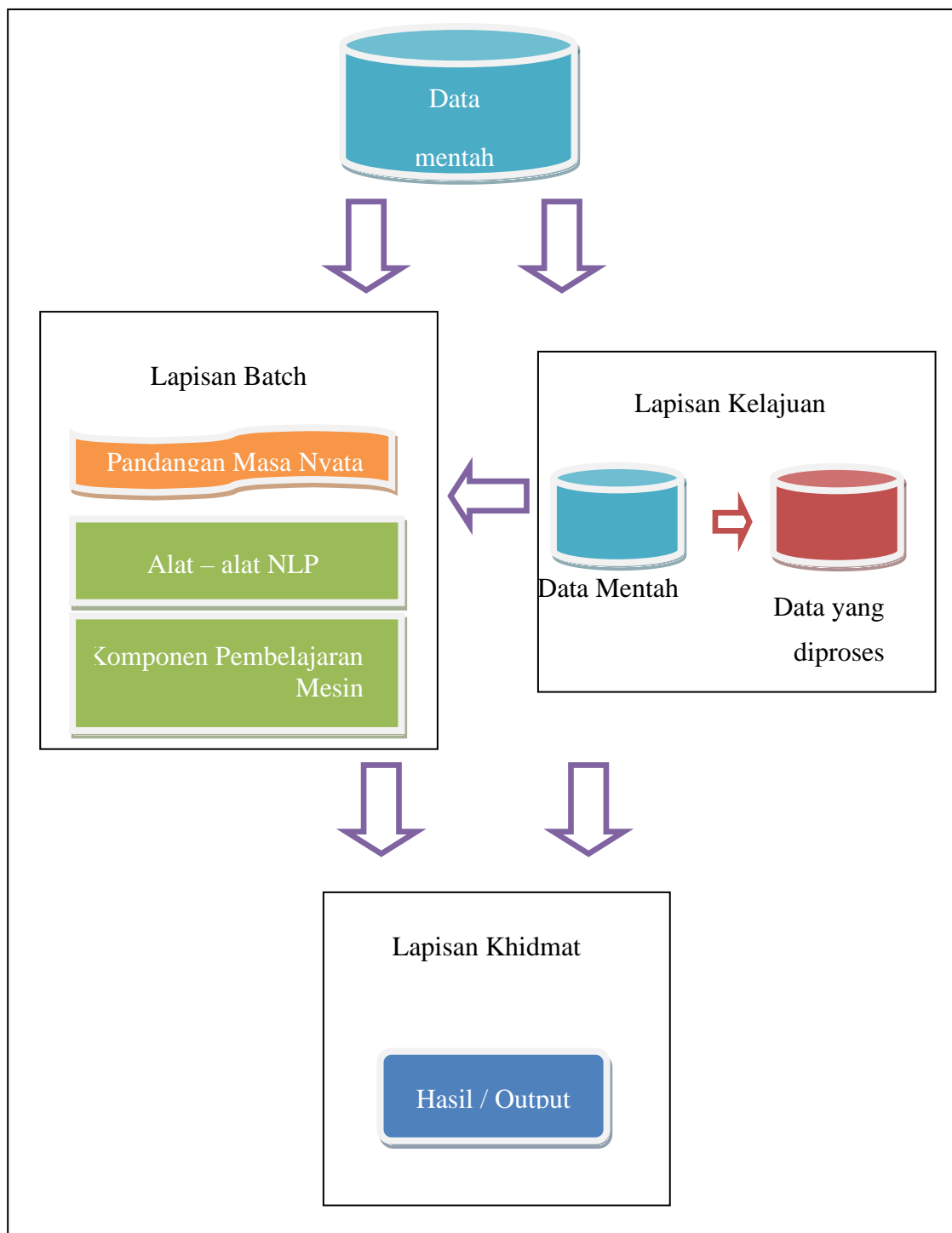
Pada dasarnya, lapisan kelajuan bertanggungjawab untuk mengisi "jurang" antara lapisan *Batch* dan lapisan khidmat dalam memberikan hasil yang dihendaki. Lapisan ini bertugas untuk membantu lapisan *Batch* dan lapisan khidmat

(b) Lapisan Batch

Tahap ini bertanggungjawab untuk menyimpan data input dalam struktur dataset induk. Data ini diproses secara berkala, secara amnya dengan pendekatan *Map Reduce* (Berlinska, 2011). Lapisan *Batch* mempercepat keputusan menggunakan algoritma pemprosesan yang diedarkan yang dapat mengendalikan jumlah data yang sangat besar. Lapisan *Batch* itu bertujuan dengan ketepatan yang sempurna dengan dapat memproses semua data yang ada ketika menghasilkan pandangan masa nyata. Lapisan ini menyempurnakan sebahagian besar kehendak algoritma ini. Output biasanya disimpan dalam pangkalan data.

(c) Lapisan Khidmat

Lapisan ini mengurus tugas output akhir, menyertakan hasil lapisan *Batch* dan Kelajuan, untuk mendapatkan satu pandangan data yang sempurna. Data yang dihasilkan oleh lapisan khidmat adalah output yang akhir algoritma ini. Tugas penting di peringkat ini adalah untuk mengintegrasikan output daripada lapisan batch dan lapisan kelajuan dengan sempurna.



Rajah 4.1 Rajah Senibina Lamda yang diubahsuai

4.2.1 Pendekatan Pembelajaran Mesin untuk Analisis Sentimen

Analisis sentimen, juga dikenali sebagai pertambahan pendapat, adalah bidang kajian yang menganalisis pendapat, sentimen, penilaian, penilaian, sikap, dan emosi orang terhadap entiti seperti produk, perkhidmatan, organisasi, individu, isu, peristiwa, topik, dan sifat mereka (Bing Liu, 2012). Terdapat dua teknik penerapan utama untuk aktiviti analisis sentimen: pembelajaran berasaskan komputer dan berasaskan leksikon. Beberapa kajian penyelidikan juga telah menggabungkan dua kaedah ini untuk mendapat prestasi dan hasil yang lebih baik. Walau bagaimanapun, kedua-dua pendekatan yang digunakan dalam sastera setakat ini telah menunjukkan dua masalah asas.

(i) Dalam konteks teknik pembelajaran mesin bersama-sama dengan sokongan teknik yang diperolehi dari NLP (Pemprosesan Bahasa Tabii), algoritma yang dibangunkan biasanya berfungsi pada data yang diberikan dalam format yang disesuaikan dengan algoritma tertentu yang dibangunkan. Contohnya, dalam bidang klasifikasi teks, satu penyelesaian lazim yang digunakan adalah menggunakan senarai perkataan yang tidak disusun (disebut *bag of words*), mengabaikan dengan itu hubungan antara kata-kata kalimat dan struktur tata bahasa kalimat itu sendiri. Terutama dalam bidang analisis sentimen, korelasi di antara kata-kata yang berdekatan dapat secara dramatik mengubah makna kalimat dalam konteks.

(ii) Satu lagi masalah asas berkaitan dengan perwakilan ciri-ciri. Raksia banyak sistem analisis bergantung pada jenis perwakilan ciri-ciri yang digunakan, seperti pilihan entiti yang dinamakan (iaitu orang atau organisasi), atau penggunaan penandaan Pos (POS). Penggunaan ciri-ciri *ad hoc* melibatkan penggunaan sumber-sumber komputasi dan menjadikan sistem atau algoritma berkembang kurang fleksibel untuk tujuan selain daripada yang ia direalisasikan. Oleh itu, adalah wajar untuk mengamalkan perwakilan umum ayat-ayat, tetapi tanpa kehilangan maklumat mengenai struktur tata bahasa di mana ayat-ayat ini dibentangkan. (Irsoy & Cardie, 2014)

(iii) Pendekatan pembelajaran mesin ini mempunyai dua dataset yang akan digunakan. Data set yang didapati daripada internet ini adalah data set yang sebenar

dan diperoleh daripada media sosial. Teks data tersebut akan mempunyai emoji. Data set ini dibahagi kepada dua iaitu, data set yang digunakan untuk melatih mesin dan juga dataset yang digunakan untuk menguji algoritma. Data set yang digunakan untuk melatih mesin ini akan mempunyai kekutuban / kekutuban yang betul untuk setiap teks dan juga setiap emoji. Emoji yang digunakan akan mempunyai kekutuban sendiri. Kekutuban teks dan juga kekutuban emoji ini diambilkira untuk melatih mesin. Kemudian, data set yang mentah digunakan untuk menguji mesin.

4.3 REKABENTUK ALGORITMA

4.3.1 Ikon emosi / Emoji dan Sentimen

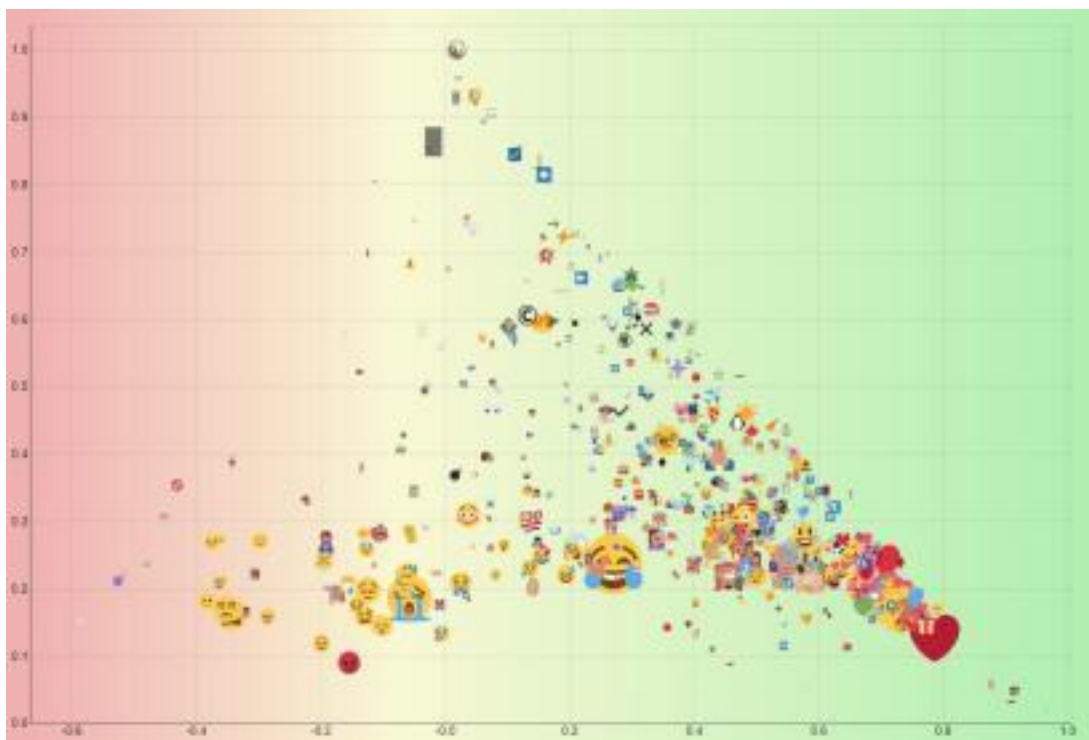
Dalam era baru komunikasi visual dalam talian, ikon emosi dan emoji semakin relevan sebagai bahasa utama yang membolehkan kita berkomunikasi dengan sesiapa sahaja di seluruh dunia. Mengumpul dan menganalisis data mengenai ikon emosi dan emoji memberikan syarikat maklumat yang berguna tentang bagaimana pelanggan merasa terhadap produk baru, kempen baru, atau mengenai jenama itu sendiri. Ikon emosi dan emoji juga dapat membantu mengenal pasti di mana terdapat keperluan untuk meningkatkan penglibatan pengguna dengan menggambarkan perasaan, sikap, dan pendapat pengguna. Oleh itu, syarikat perlu melaksanakan kajian analisis sentimen.

Ikon emosi dan emoji dianggap sebagai petunjuk sentimen yang berguna dan boleh dipercayai, dan oleh itu boleh digunakan sama ada untuk secara automatik menghasilkan korpus latihan atau bertindak sebagai ciri bukti untuk meningkatkan klasifikasi sentimen. Ikon emosi diperkenalkan sebagai komponen ekspresif, tanpa lisan ke dalam bahasa bertulis, mencerminkan peranan yang dimainkan oleh ekspresi muka dalam ucapan. Peranan mereka terutama pragmatik: ikon emosi memberikan rasa positif atau negatif kepada ayat-ayat bertulis oleh ungkapan visual. Menurut pertimbangan ini, terdapat hubungan antara orientasi sentimen emoji, ikon emosi dan mesej. Ikon emosi dan emoji telah dibezakan dalam dua kategori utama, iaitu positif dan negatif. Contoh-contoh ikon emosi positif adalah :-), :), =),: D, sementara contoh-

contoh negatif adalah :-), :(, = (, ; (.) Hal ini menunjukkan ikon emosi dan emoji pasti merupakan sumber maklumat penting untuk klasifikasi kutub. Malah, pada media sosial mesej positif dan negatif mempunyai peratusan tinggi ikon emosi dan emoji.

Maklumat yang terkandung dalam media sosial, yang menjadi salah satu jenis komunikasi utama, menjadikannya sumber data yang menarik untuk analisis sentimen. Bukan sahaja teks tetapi juga emoji, yang mewakili elemen linguistik yang biasanya digunakan dalam media sosial untuk mendapatkan mesej yang diberikan, boleh digunakan untuk meningkatkan analisis sentimen.

Ketepatan pengiktirafan emosi boleh meningkat dengan analisis ikon emosi dan emoji. Mereka menyediakan sekeping maklumat penting. Oleh itu, sebuah pangkalan data yang mempunyai kekutuban ikon emosi dan emoji akan digunakan untuk meningkatkan ketepatan output kajian ini. Kajian ini akan menggunakan perpustakaan NLTK daripada *Python 3*. Rajah 4.2 menunjukkan kekutuban emoji.



Rajah 4.2 Kekutuban dan Neutraliti Emoji

4.3.2 Rekabentuk Algoritma

Sebelum proses analisis sentimen dimulakan, data mentah perlu sedia ada. Data mentah yang digunakan untuk kajian ini adalah dicipta untuk kegunaan analisis sentimen kajian ini dan sebahagian daripada data diambil daripada internet untuk melatih mesin untuk belajar. Pendapat / luahan yang mempunyai emoji dimasukkan dalam satu dokumen .txt. Satu teks diletakkan dalam satu baris. Setiap baris seterusnya adalah teks seterusnya.

Selepas memperolehi data mentah, setiap emoji dalam teks tersebut harus ditukar kepada simbol atau dalam perkataan yang boleh dibaca untuk membolehkan algoritma analisis sentimen ini mengenalpasti dan membaca emoji berkenaan. Emoji harus ditukar kepada simbol atau perkataan kerana emoji ini berbentuk gambar dan sukar untuk dibaca oleh algoritma jika tidak ditukar kepada simbol atau perkataan.

Selepas mempunyai data mentah yang lengkap dan boleh dibaca oleh mesin, proses analisis sentimen boleh dijalankan menggunakan NLTK dalam *Python 3*. Antara proses yang akan dijalankan untuk menjalankan analisis sentimen adalah pra-pemprosesan data. Proses ini mengeluarkan perkataan-perkataan yang tidak bermakna daripada teks dan meninggalkan perkataan-perkataan yang bermakna sahaja. Semua perkataan dalam teks ditukarkan kepada huruf kecil. Simbol – simbol yang tidak bermakna seperti #, \$, @, %, ^, * dikeluarkan daripada teks. Bukan itu sahaja, nama-nama dikeluarkan daripada teks. Jika sumber data ini daripada Twitter, teks akan mengandungi nama akaun seperti '@PriyaVishnu16'. Nama – nama sebegini tidak bermanfaat dalam menentukan sentimen teks tersebut. Oleh itu, nama akaun dikeluarkan daripada teks.

Semua dataset yang diperoleh dan dicipta ini akan digunakan dalam dua bahagian iaitu, dataset untuk melatih mesin dan dataset untuk menguji mesin. Bagi data set yang digunakan untuk melatih mesin ini, teks yang mengandungi emoji ini akan diklasifikasikan kepada tiga kategori iaitu positif, neutral dan negatif mengikut kekutuban / kekutuban teks dan kekutuban / kekutuban emoji dengan menggunakan algoritma yang sesuai.

Dalam algoritma ini, kekutuban / kekutuban teks ditentukan dengan menggunakan *VADER Lexicon*, *Sentimen Intensity Analyzer* dan NLTK. Kekutuban dan kekutuban ikon emosi dan emoji ditentukan dengan menggunakan pangkalan data. Pangkalan data ini mempunyai ikon emosi dan emoji dan kekutuban setiap emoji tersebut. Kemudian mesin akan dilatih menggunakan set data ini.

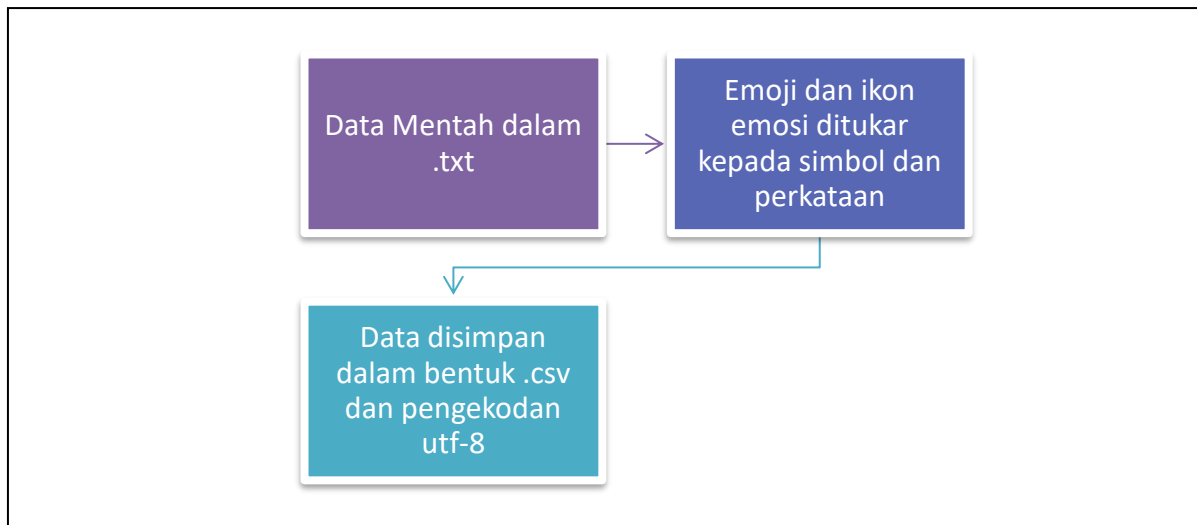
Seterusnya, semua pendapat positif, negatif dan neutral akan dimasukkan dalam satu fail. Susunan setiap teks yang akan dianalisis ini adalah rawak. Data set ini yang disusun secara rawak ini akan digunakan untuk fasa menguji mesin. Sebelum menguji, teks tersebut harus dibuat pra-pemprosesan. Teks dalam data set ini tidak mempunyai kekutuban dan kekutuban.

Model pengelas yang sesuai digunakan untuk menguji dan melatih untuk algoritma analisis sentimen yang berfokuskan emoji ini. Model pengelas yang dilatih ini akan disimpan untuk kegunaan masa depan.

Algoritma ini berfokuskan kepada ketepatan dan dapat memberi sentimen dalam masa nyata. Oleh itu, model pengelas harus memberi sentimen yang tepat dalam masa yang tersingkat.

Model pengelas yang disimpan ini digunakan untuk menentukan sentimen setiap teks yang hendak ditentukan sentimen. Sebelum menggunakan model pengelas untuk menentukan sentimen, teks yang hendak ditentukan sentimen harus dibuat pra-pemprosesan data seperti meninggalkan perkataan-perkataan yang bermakna dan emoji sahaja. Setiap teks diberi sentimen, iaitu positif, neutral atau negatif. Proses-proses analisis sentimen yang berfokuskan emoji ditunjukkan dalam rajah-rajah berikut.

- Lapisan Kelajuan



Rajah 4.3 Rajah Carta Alir Lapisan Kelajuan

```

Line 1: I really love flying to North Korea !! @tashmin9 🇰🇷
Line 2: I dislike being in crowded place 🤔 @cillian5
  
```

Rajah 4.4 Rajah Data sebelum Lapisan Kelajuan

Rajah 4.4 menunjukkan contoh data sebelum proses lapisan kelajuan. Emoji dalam data tersebut berbentuk gambar.

```

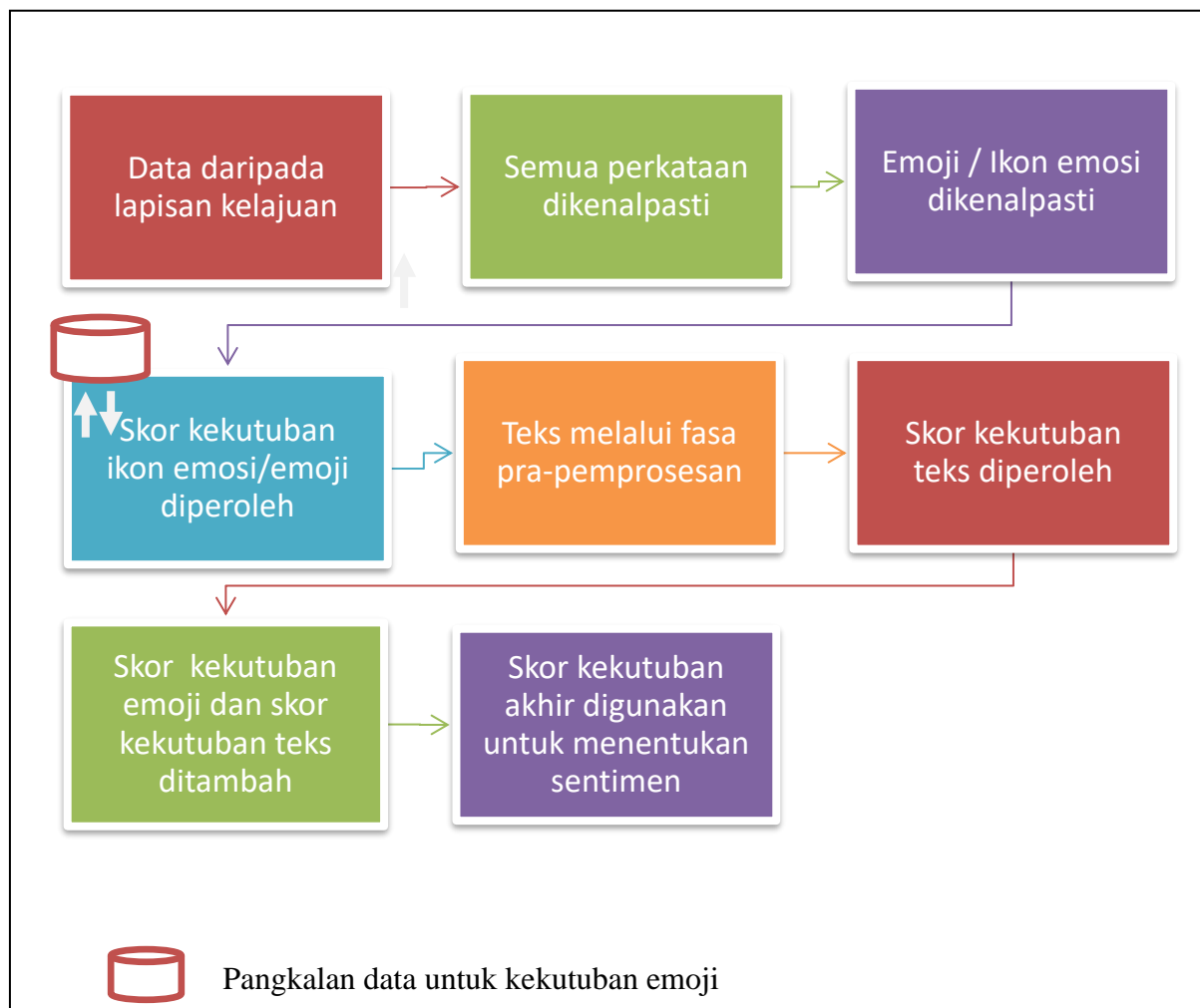
Line 1: I really love flying to North Korea !! @tashmin9 \xF0\x9F\x98\x8D
Line 2: I dislike being in crowded place \xF0\x9F\x98\xA9 @cillian5
  
```

Rajah 4.5 Rajah Data selepas Lapisan Kelajuan

Rajah 4.5 menunjukkan contoh data selepas proses lapisan kelajuan. Emoji dalam data mentah ditukarkan kepada bentuk *bytes*, *unicode* atau perkataan yang boleh dibaca oleh mesin.

- Lapisan *Batch*

(i) Bahagian pertama – Penentuan sentimen teks



Rajah 4.6

Rajah Carta Alir Lapisan *Batch* bahagian pertama

Line 1: I really love flying to North Korea !! @tashmin9 \xF0\x9F\x98\x8D
 Line 2: I dislike being in crowded place \xF0\x9F\x98\xA9 @cillian5

Rajah 4.7

Rajah data sebelum Lapisan *Batch* bahagian pertama

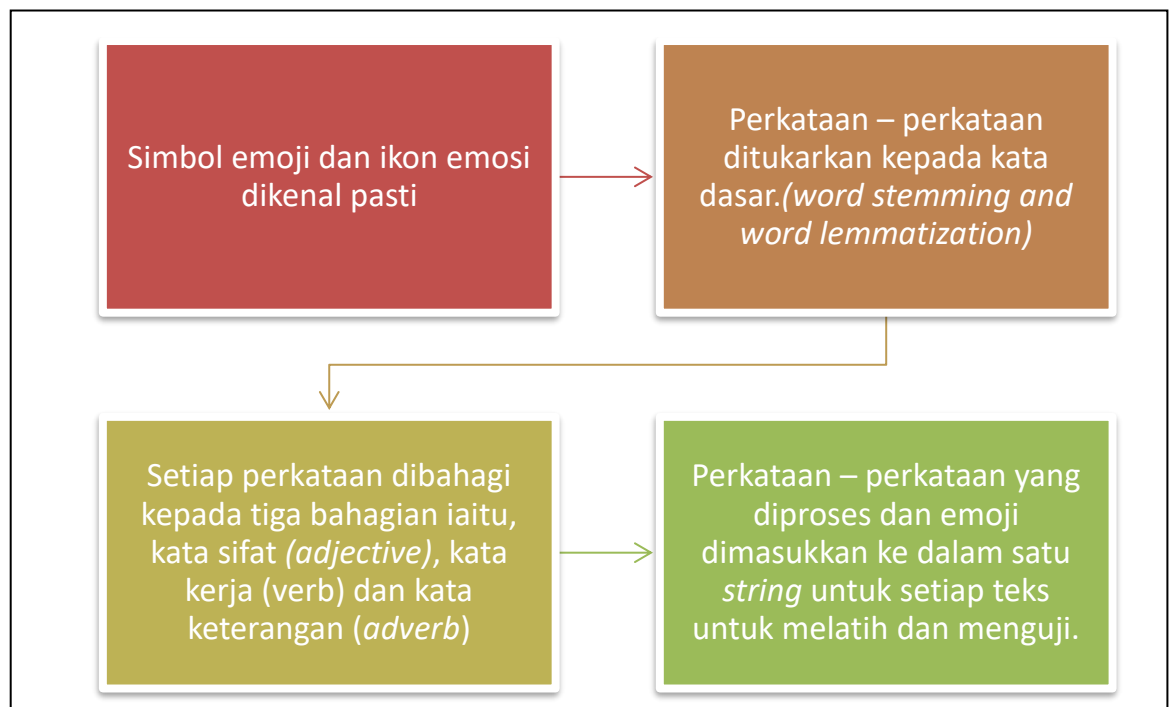
Rajah 4.7 menunjukkan contoh data sebelum proses lapisan *Batch* yang pertama. Data yang diproses daripada lapisan kelajuan digunakan untuk lapisan *Batch*.

emoji	emoji_score	preprocessed_text	sentiment_score	final_sentimentvalue	predicted sentiment
\xF0\x9F\x98\x8D	0.678	[really, love, flying, north, korea]	0.6369	1.3149	positive
\xF0\x9F\x98\xA9	-0.368	[dislike, crowded, place]	-0.3818	-0.7498	negative

Rajah 4.8 Rajah data selepas Lapisan *Batch* bahagian pertama

Rajah 4.8 menunjukkan data selepas lapisan Batch yang pertama. Semua perkataan yang tidak bermakna dikeluarkan. Skor kekutuban emoji dan skor kekutuban teks ditambah. Skor kekutuban terakhir digunakan untuk menentukan sentimen setiap teks.

(ii) Bahagian kedua – Penghasilan data set ujian dan data set latihan



Rajah 4.9 Rajah Carta Alir Lapisan *Batch* bahagian kedua

Line 1: I really love flying to North Korea !! @tashmin9 \xF0\x9F\x98\x8D
 Line 2: I dislike being in crowded place \xF0\x9F\x98\xA9 @cillian5

Rajah 4.10 Rajah data sebelum Lapisan *Batch* bahagian kedua

```

traintestset
['really', 'love', 'fly', 'north', 'korea'], \xF0\x9F\x98\x8D
['dislike', 'crowd', 'place'], \xF0\x9F\x98\xA9

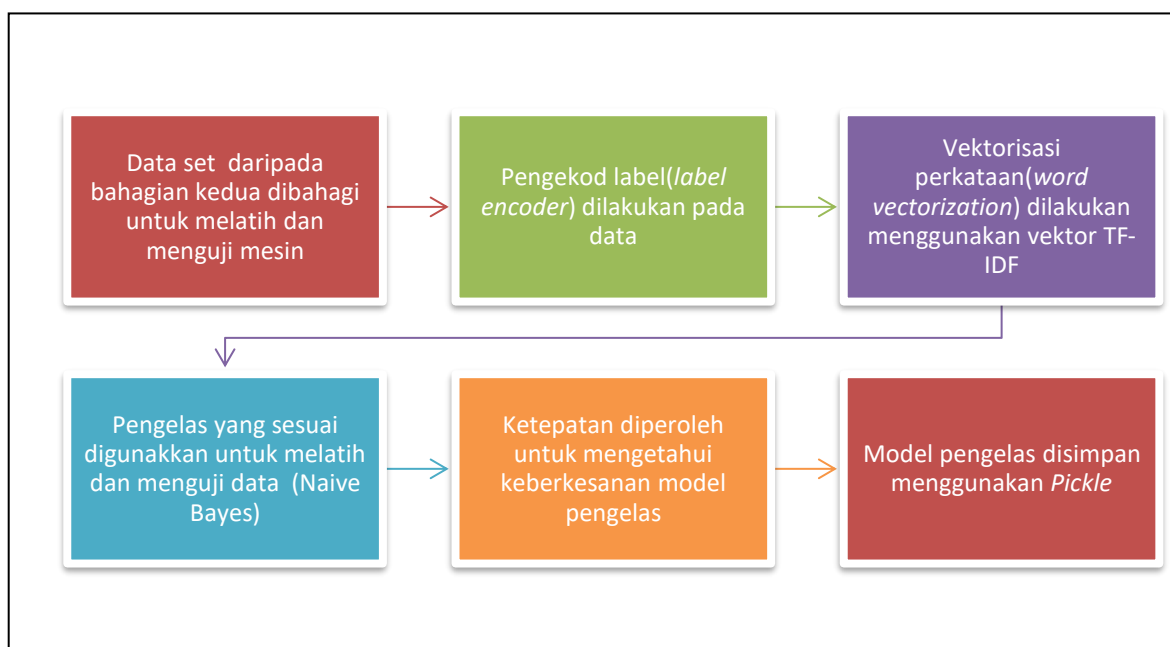
```

Rajah 4.11 Rajah data selepas Lapisan *Batch* bahagian kedua

Rajah 4.11 menunjukkan hasil lapisan *Batch* kedua iaitu penghasilan set data ujian dan set data latihan. Perkataan- perkataan bermakna dan emoji ditambah dan disimpan dalam satu bahagian. Set data ini berbeza dengan set data bahagian pertama kerana perkataan- perkataan dalam set data ini ditukarkan kepada kata dasar. Contohnya :-

flying - fly
crowded - crowd

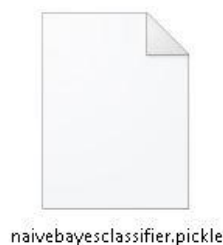
(ii) Bahagian ketiga – Penghasilan model pengelas



Rajah 4.12 Rajah Carta Alir Lapisan *Batch* bahagian ketiga

traintestset				
['really', 'love', 'fly', 'north', 'korea'],\xF0\x9F\x98\x8D				
['dislike', 'crowd', 'place'],\xF0\x9F\x98\xA9				

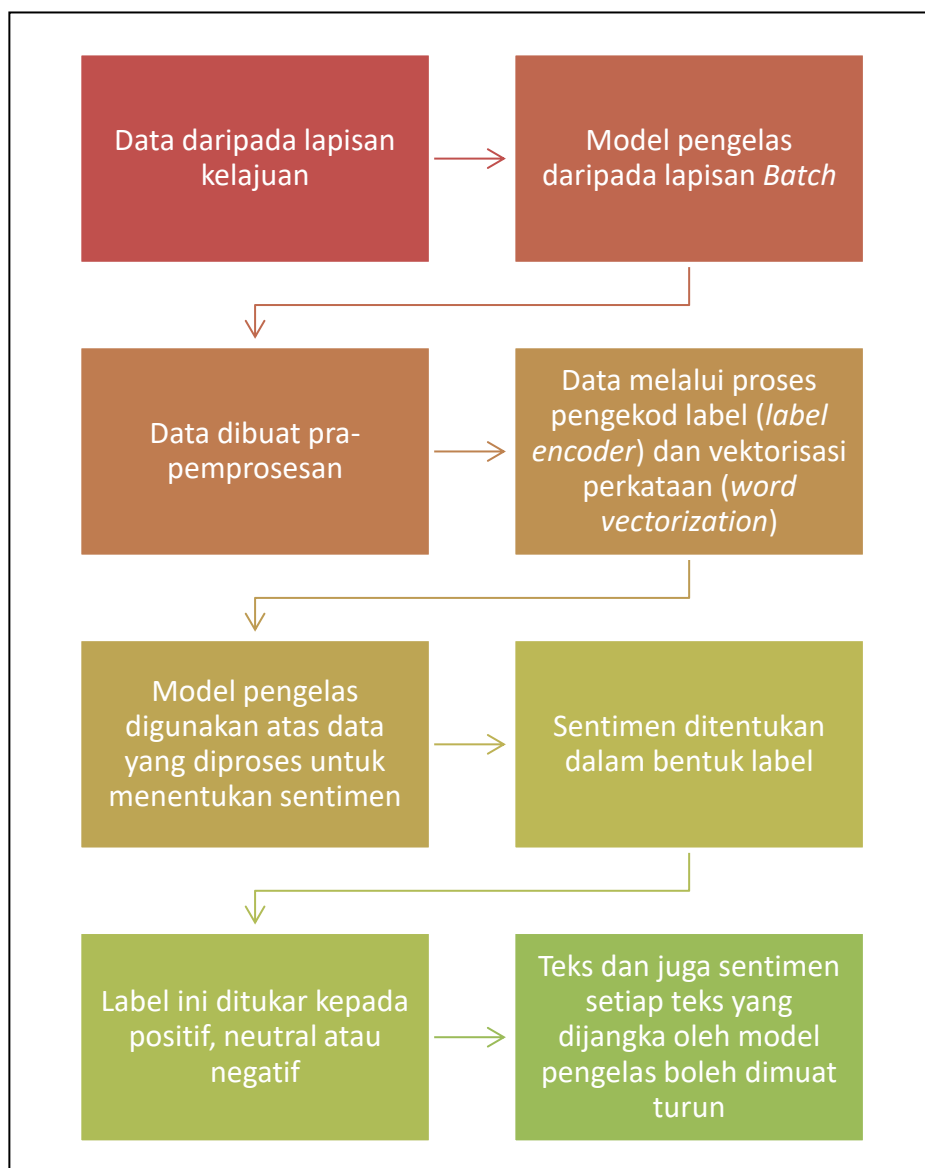
Rajah 4.13 Rajah data sebelum Lapisan *Batch* bahagian ketiga



Rajah 4.14 Rajah hasil Lapisan *Batch* bahagian ketiga

Hasil bahagian ketiga Lapisan Batch ini adalah model pengelas. Model pengelas yang digunakan dalam algoritma analisis sentimen yang berfokuskan emoji ini adalah *Naive bayes*. Rajah 4.13 adalah contoh format set data yang digunakan. Penghasilan model pengelas memerlukan minima 200 bilangan teks yang mengandungi emoji untuk penghasilan model pengelas yang efektif. Model ini akan disimpan dan digunakan untuk lapisan khidmat. Model pengelas baharu perlu dihasilkan jika teks yang hendak ditentukan sentimen ini daripada domain yang lain. Hal ini kerana, domain yang berbeza akan mempunyai variasi dalam penggunaan perkataan dan penentuan sentimen.

- Lapisan Khidmat



Rajah 4.15

Rajah Carta Alir Lapisan khidmat

Line 1: I really love flying to North Korea !! @tashmin9 \xF0\x9F\x98\x8D
 Line 2: I dislike being in crowded place \xF0\x9F\x98\xA9 @cillian5

Rajah 4.16

Rajah data sebelum Lapisan khidmat

text	sentiment	predicted sentiment
I really love flying to North Korea !! @tashmin9 \xF0\x9F\x98\x8D	2	positive
I dislike being in crowded place \xF0\x9F\x98\xA9 @cillian5	0	negative

Rajah 4.17 Rajah hasil Lapisan khidmat

Rajah 4.17 menunjukkan jangkaan sentimen yang dihasilkan oleh model pengelas Algoritma ini menggunakan sebagai model pengelas *Naive bayes*.

4.3.3 Pseudokod Algoritma

A Lapisan Kelajuan

- (i) Data mentah disimpan dalam fail jenis .txt.
- (ii) Emoji dan ikon emosi dalam data mentah ditukar kepada simbol dan perkataan yang boleh dibaca oleh mesin
- (iii) Jenis fail ditukar pengekodan utf-8 dan disimpan dalam fail .csv.

B Lapisan *Batch*

- **Bahagian Pertama – Penentuan Sentimen Teks**

- (i) Data daripada lapisan kelajuan dibaca
- (ii) Setiap perkataan dan simbol dikenal pasti (*Tokenized*)
- (iii) Emoji / Ikon emosi dikenal pasti
- (iv) Kekutuban emoji / ikon emosi dalam setiap teks didapati daripada pangkalan data emoji sentimen yang disediakan terlebih dahulu
- (v) Teks melalui fasa pra pemprosesan
- (vi) Kekutuban teks didapati melalui *Vader Lexicon*, *Sentimen Intensity Analyzer* dan NLTK.
- (vii) Kekutuban emoji dan kekutuban teks ditambah untuk mendapatkan kekutuban akhir teks.

(viii) Kekutuban akhir teks digunakan untuk menentukan sentimen setiap teks tersebut

Jadual 4.1 Skor kekutuban dan sentimen

Kekutuban	Sentimen
Kurang atau bersamaan dengan -0.1	Negatif
Lebih daripada -0.1 dan kurang daripada 0.1	Neutral
Lebih atau bersamaan dengan 0.1	Positif

- **Bahagian Kedua – Penghasilan set data ujian dan set data latihan**

- (i) Simbol emoji dan ikon emosi dikenal pasti
- (ii) Teks melalui pra-pemprosesan data
- (iii) Perkataan – perkataan yang tinggal ditukarkan kepada katar dasar.

(word stemming and word lemmatization)

- (iv) Setiap perkataan dibahagi kepada tiga bahagian iaitu, kata sifat (*adjective*), kata kerja (*verb*) dan kata keterangan (*adverb*)

Jadual 4.2 Jenis perkataan dan peta tag

Jenis Perkataan	Peta Tag (<i>Tag Map</i>)
Kata sifat (<i>adjective</i>)	J
Kata kerja (<i>verb</i>)	V
Kata keterangan (<i>adverb</i>)	R

- (v) Perkataan – perkataan yang diproses dan emoji dimasukkan ke dalam satu *string* untuk setiap teks untuk melatih dan menguji.

- **Bahagian Ketiga – Penghasilan Model Pengelas**

- (i) Data set daripada bahagian kedua dibahagi untuk melatih dan menguji mesin

Jadual 4.3 Jenis data dan peratusan

Jenis data set	Peratusan
Data set untuk melatih	70%
Data set untuk menguji	30%

- (ii) Pengekod label(*label encoder*) dilakukan pada data
- (iii) Seterusnya, vektorisasi perkataan(*word vectorization*) dilakukan menggunakan vektor TF-IDF
- (iv) Pengelas yang sesuai digunakkan untuk melatih dan menguji data
- (v) Ketepatan(*accuracy*), ketepatan(*precision*) dan dapatan balik(*recall*) digunakan untuk mengetahui keberkesanan model pengelas
- (vi) Model pengelas disimpan menggunakan *Pickle*.

B Lapisan Khidmat

- (i) Dataset daripada lapisan Kelajuan digunakan untuk menganalisis sentimen
- (ii) Model pengelas daripada lapisan *Batch* digunakan untuk menentukan sentimen.
- (iii) Data dibuat pra-pemprosesan dan pemprosesan seperti *Word Stemming and Word Lemmatization*
- (iv) Seterusnya, data melalui proses pengekod label (*label encoder*) dan vektorisasi perkataan (*word vectorization*)

- (v) Model pengelas digunakan atas data yang diproses untuk menentukan sentimen
- (vi) Sentimen ditentukan dalam bentuk label. Label ini ditukar kepada perkataan untuk memudahkan kefahaman.

Jadual 4.4 Label dan sentimen

Label yang diberi oleh model pengelas	Sentimen
0	Negatif
1	Neutral
2	Positif

- (vii) Teks dan juga sentimen setiap teks yang dijangka oleh model pengelas boleh dimuat turun

4.3.4 Kaedah Pembelajaran Mesin dan Model Pengelas

Analisis sentimen adalah bidang kajian yang mengenalpasti dan mengeluarkan maklumat subjektif teks. Dua pendekatan utama boleh digunakan dalam algoritma ini adalah pendekatan mesin dan pendekatan leksikon.

Pendekatan pembelajaran mesin melihat kekutuban klasifikasi sebagai masalah pengkategorian teks. Teks biasanya diwakili sebagai vektor ciri-ciri, dan bergantung pada ciri yang digunakan oleh algoritma untuk membolehkan mencapai hasil yang lebih baik. Sekiranya set latihan berlabel diperlukan, pendekatan ditakrifkan sebagai pendekatan terselia dan jika tidak, ia ditakrifkan sebagai pendekatan tidak terselia pembelajaran. Algoritma ini menggunakan pendekatan terselia. Pendekatan ini berfungsi dengan baik dalam domain dilatih, tetapi prestasi menurun ketika pengelas yang sama digunakan dalam domain yang berbeza. Oleh itu, jika ingin analisis sentimen untuk domain lain pengelas harus dilatih menggunakan domain baharu terlebih dahulu sebelum digunakan.

Pengelas *Naive Bayes* adalah salah satu yang paling mudah dan paling kaedah klasifikasi yang biasa digunakan dalam penyelidikan. Ini klasifikasi mengira kebarangkalian kategori berdasarkan bilangan pendedaran perkataan dalam dataset. Klasifikasi analisis sentimen ini menggunakan teorem *Bayes* untuk meramalkan kemungkinan ciri-ciri yang telah dilabelkan mengikut yang telah ditetapkan kategori.

Klasifikasi teks dilakukan menggunakan kaedah *Naive Bayes*. Nilai penilaian prestasi mesti diambil kira dalam kaedah klasifikasi ini supaya model pengelas boleh dipercayai semasa menggunakannya. Pengukuran boleh dilakukan dengan kaedah matriks prestasi seperti Jadual 4.5 berikut:

Jadual 4.5 Matriks Prestasi

		Data selepas analisis sentimen (Data yang diuji)	
		Positif	Negatif
Data yang dilabel (Data yang dilatih)	Positif	Positif Benar (TP)	Negatif Palsu (FN)
	Negatif	Positif Palsu (FP)	Negatif Benar (TN)

- Positif Benar (TP) - klasifikasi kelas adalah positif dan ramalan model pengelas adalah positif.
- Negatif Benar (TN) - klasifikasi kelas adalah negatif dan ramalan model pengelas adalah negatif.
- Positif Palsu (FP) - klasifikasi kelas adalah negatif tetapi ramalan model pengelas adalah positif.

- d. Negatif Palsu (FN) - klasifikasi kelas adalah positif tetapi ramalan model pengelas adalah negatif.

Cara mengira ketepatan (*accuracy*), ketepatan (*precision*) dan dapatan semula (*recall*) adalah seperti berikut:-

$$(i) \quad \text{Ketepatan (accuracy)} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$(ii) \quad \text{Ketepatan (precision)} = \frac{TP}{TP + FP}$$

$$(iii) \quad \text{Dapatan semula (recall)} = \frac{TP}{TP + FN}$$

4.4 KESIMPULAN

Reka bentuk adalah satu langkah dan gambaran awal yang amat penting untuk pembangunan sesebuah algoritma. Reka bentuk seni bina dan reka bentuk algoritma adalah antara yang paling diperlukan sebelum membangunkan algoritma ini iaitu analisis sentimen yang berfokuskan ikon emosi dan emoji. Pembangunan sesebuah algoritma yang berfungsi haruslah merujuk kepada reka bentuk – reka bentuk ini untuk memastikan proses pembangunan berjalan dengan lancar dan semua matlamat dan objektif kajian dapat dicapai.

BAB V

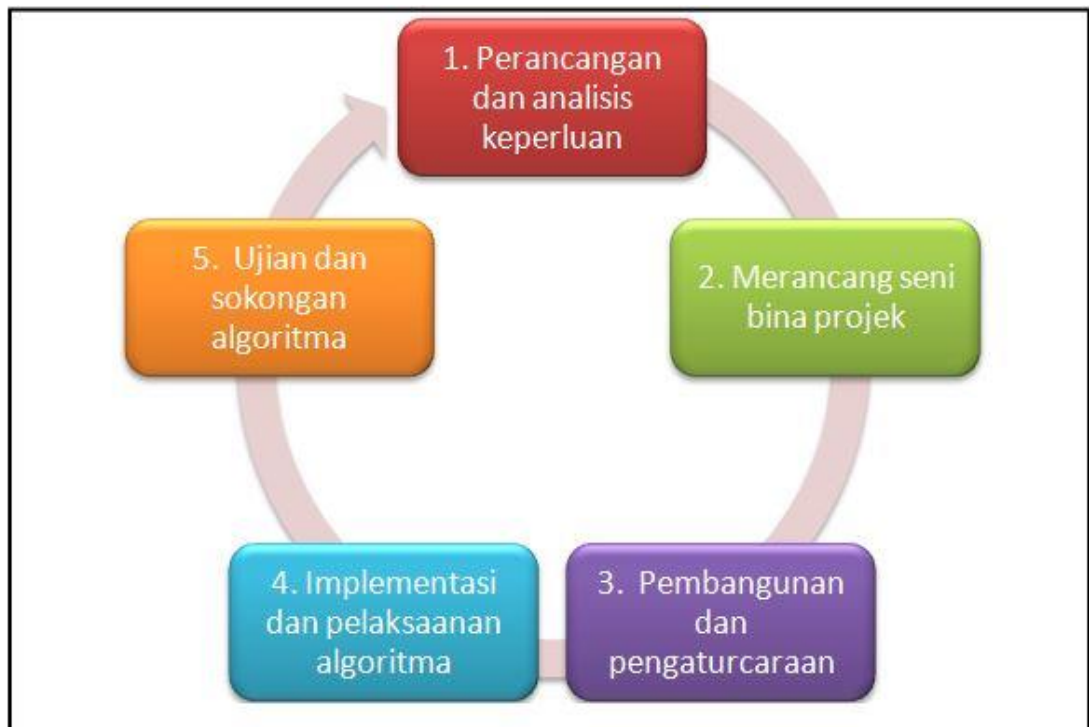
IMPLEMENTASI DAN PENGUJIAN

5.1 PENGENALAN

Bab ini membincangkan pembangunan algoritma dan membincangkan kod-kod yang kritikal untuk pembangunan algoritma ini. Algoritma analisis sentimen yang berfokuskan emoji ini menggunakan bahasa pengaturcaraan *Python 3*. Pengujian dijalankan dalam beberapa tahap untuk memastikan algoritma berfungsi seperti diinginkan dan mencapai objektifnya.

5.2 TEKNOLOGI PEMBANGUNAN

Pembangunan algoritma seperti yang ditunjukkan dalam Rajah 5.1. Model Kitar Hayat dipilih untuk mengurangkan risiko kegagalan semasa membangunkan algoritma yang diinginkan. Selain itu, model ini mudah untuk difahami dan digunakan. Fasa pembangunan algoritma lebih mudah untuk dipantau.



Rajah 5.1 Model Kitar Hayat

5.3 BAHAGIAN KRITIKAL ATURCARA

5.3.1 Kekutuban Emoji / Ikon emosi

```
emojisentiment = dict(emojisentiment.values.tolist()) #Converting to dicts
analyser = SentimentIntensityAnalyzer()
all_tweets = list(tweets['text'])
emoji_sentimental_score = []
for tweet in all_tweets:
    nltk_tokens = nltk.word_tokenize(str(tweet))
    print(nltk_tokens)
    avrg_sentimental_score = 0.0
    score = 0
    counter = 0
    for token in nltk_tokens:
        if token in emojisentiment.keys():
            score += emojisentiment[token]
            counter += 1
    if counter > 0:
        avrg_sentimental_score = score/counter
    emoji_sentimental_score.append(avrg_sentimental_score)

tweets['emoji_score'] = emoji_sentimental_score
print(tweets[tweets['emoji_score'] > 0])
```

Rajah 5.2 Kod Kekutuban Emoji / Ikon emosi

Pangkalan data mempunyai simbol / perkataan yang menentukan emoji dan skor kekutuban kepada emoji tersebut. Data dari *Twitter* diambil untuk menganalisis sentimen. Skor kekutuban emoji yang digunakan untuk setiap *tweet* yang terdapat dalam fail .csv dipaparkan. Jika terdapat beberapa emoji, purata skor kekutuban emoji / ikon emosi tersebut dipaparkan.

5.3.2 Sentimen Keseluruhan Teks

```
i = 0
final_sentiment = [ ] #empty series to hold our final sentiment values

while(i<len(tweets)):
    y = 0
    y = (tweets.iloc[i]['sentiment_score']) + (tweets.iloc[i]['emoji_score'])
    final_sentiment.append(y)
    i = i+1

tweets['final_sentimentvalue'] = final_sentiment
tweets.head(20)
```

Rajah 5.3 Kod Sentimen Keseluruhan Teks

Skor kekutuban teks dikira dan ditambahkan dengan skor kekutuban emoji daripada pangkalan data. Jika tiada emoji digunakan dalam sesuatu teks, skor kekutuban teks adalah skor kekutuban yang akhir. didapati dan dipaparkan.

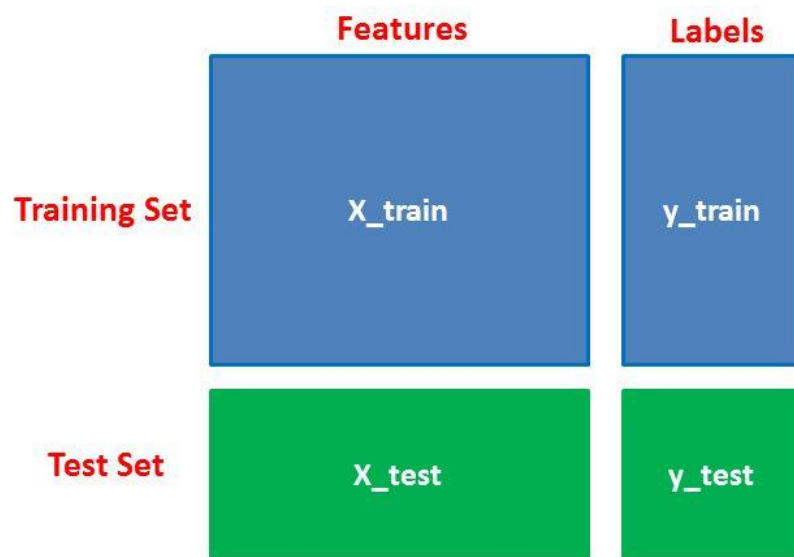
5.3.3 Segmentasi data

```
# train and test dataset split
train_x, test_x, train_y, test_y = model_selection.train_test_split(tweets['text_final'],
                                                                    tweets['predicted sentiment'],
                                                                    test_size = 0.3)

# encoding the dataset
encoder = LabelEncoder()
train_y = encoder.fit_transform(train_y)
test_y = encoder.fit_transform(test_y)
```

Rajah 5.4 Kod Segmentasi data

Data dibahagi kepada data set untuk dilatih dan diuji data untuk menguji model pengelas *Naïve Bayes*. 70% daripada data ini digunakan untuk melatih dan 30% daripada data ini digunakan untuk menguji. *Label Encoder* digunakan untuk pengekodan data yang ingin diuji dan dilatih.



Rajah 5.5 Cara data disegmentasi

Data untuk set pengujian dan set ujian dibahagi seperti yang dipaparkan dalam Rajah 5.5.

5.3.4 Vektorisasi perkataan

```
# word vectorization

Tfidf_vect = TfidfVectorizer(max_features=5000)
Tfidf_vect.fit(tweets['text_final'])
Train_X_Tfidf = Tfidf_vect.transform(train_x)
Test_X_Tfidf = Tfidf_vect.transform(test_x)
```

Rajah 5.6 Kod vektorisasi perkataan

Perkataan divektorisasi dengan menggunakan TF-IDF (*term frequency-inverse document frequency*) adalah ukuran statistik yang menilai betapa relevannya kata dengan dokumen dalam kumpulan dokumen. Ini dilakukan dengan mengalikan dua metrik: berapa kali perkataan muncul dan kekerapan perkataan sebaliknya muncul dalam dokumen.

5.3.5 Pembangunan Model Pegelas

```
# classification using Naive Bayes classifier

Naive = naive_bayes.MultinomialNB()
Naive.fit(Train_X_Tfidf,train_y)

# predict the labels on validation dataset
predictions_NB = Naive.predict(Test_X_Tfidf)
print(Test_X_Tfidf)
print(predictions_NB)
```

Rajah 5.7 Kod model pengelas

Model Pengelas *Naïve Bayes* digunakan untuk menentukan sentimen teks yang digunakan untuk menguji model. *Naïve bayes* adalah algoritma pembelajaran yang menggunakan peraturan *Bayes* bersama dengan anggapan bahawa atribut tidak mempunyai syarat berdasarkan kelas. *Naïve Bayes* memberikan ketepatan klasifikasi yang kompetitif.

5.4 KETEPATAN ALGORITMA

5.4.1 Menguji Model Pegelas

(0, 3167)	0.5089294070981548	(1559, 3597)	0.25623084161682674
(0, 2939)	0.4987153261541142	(1559, 3553)	0.2757577973169548
(0, 1685)	0.45676071703474996	(1559, 3143)	0.1446615061292319
(0, 1676)	0.34142891831113825	(1559, 2393)	0.17162405388748297
(0, 1249)	0.4087417560813958	(1559, 2056)	0.2458085938344187
(1, 4755)	0.2429305846160643	(1559, 1960)	0.2960317558403455
(1, 4404)	0.3353183237748329	(1559, 1804)	0.11077833603057861
(1, 4332)	0.30909833947742216	(1559, 682)	0.2960317558403455
(1, 3118)	0.18407662135461134	(1559, 356)	0.15935083934978728
(1, 2873)	0.3834893583961743	(1559, 317)	0.2395827419977168
(1, 2366)	0.28792950609863455	(1560, 4778)	0.44457537375176126
(1, 2155)	0.26170952180122387	(1560, 4170)	0.24860607435453763
(1, 1804)	0.11566799521194154	(1560, 4020)	0.3198591918018586
(1, 1314)	0.30909833947742216	(1560, 3859)	0.2876683219380392
(1, 719)	0.16862341004377485	(1560, 2941)	0.31181118490702237
(1, 476)	0.28792950609863455	(1560, 2483)	0.28126851028227323
(1, 412)	0.2708320653536076	(1560, 950)	0.34874391160709284
(1, 366)	0.2862035966900892	(1560, 101)	0.5074685064397358
(1, 175)	0.15855716524988223	(1561, 3936)	0.4102807750538735
(2, 4728)	0.291548310322206	(1561, 2542)	0.22588002394501155
(2, 3151)	0.3010463707608327	(1561, 2481)	0.3657702979769018
(2, 1959)	0.36005371177650813	(1561, 2395)	0.34505857379495275
(2, 1673)	0.24574265152871763	(1561, 1044)	0.4168462460361467
(2, 1650)	0.0995444815249653	(1561, 753)	0.49412117234535247
(2, 825)	0.22029583967926464	(1561, 545)	0.3314785730298448
:	:	[2 2 2 ... 2 2 0]	

Rajah 5.8 Sentimen teks

Rajah 5.7 menunjukkan hasil ujian menggunakan klasifier *Naïve Bayes*.

5.4.2 Skor Naive Bayes

```
#Testing Naive Bayes Model
Naive.score(Test_X_Tfidf,test_y)

0.7272727272727273
```

Rajah 5.9 Skor *Naive Bayes*

Skor ini menunjukkan ketepatan model dalam mengelaskan sentimen teks yang mengandungi emoji dengan betul. Model akan mengelaskan setiap teks yang mengandungi emoji ini kepada positif, neutral atau negatif.

5.4.3 Matriks *Confusion*

```
#Formulation of confusion matrix
from sklearn.metrics import confusion_matrix
y_pred = Naive.fit(Train_X_Tfidf,train_y).predict(Test_X_Tfidf)
cm = confusion_matrix(test_y, y_pred)
print (cm)

[[30  6  4]
 [ 6 34  6]
 [ 6 14 48]]
```

Rajah 5.10 Matriks *Confusion*

Matriks *Confusion* adalah jadual yang sering digunakan untuk menggambarkan prestasi model klasifikasi. Dalam kajian ini, model pengelas *Naïve Bayes* digunakan untuk sekumpulan data ujian yang mana nilai yang dikenal pasti.

5.4.4 Ketepatan dan Dapatan Balik

```
#Precision and recall for each label
print("label precision recall")
for label in range(3):
    print(f"{label:5d} {precision(label, cm):9.3f} {recall(label, cm):6.3f}")

label precision recall
0      0.714  0.750
1      0.630  0.739
2      0.828  0.706
```

Rajah 5.11 Ketepatan dan Dapatan Balik

Label 0 menunjukkan ketepatan dan dapatan semula untuk sentimen yang negatif. Label 1 menunjukkan ketepatan dan dapatan semula untuk sentimen yang neutral. Label 2 menunjukkan ketepatan dan dapatan semula untuk sentimen yang positif.

```
#Precision and Recall of the Naive Bayes Model
print("precision total:", precision_macro_average(cm))

print("recall total:", recall_macro_average(cm))

precision total: 0.723833850270632
recall total: 0.7316709292412616
```

Rajah 5.12 Ketepatan dan Dapatan Balik

Cara mengira ketepatan (*precision*) dan dapatan semula (*recall*) adalah seperti berikut:-

$$(i) \quad \text{Ketepatan (precision)} = \frac{TP}{TP + FP}$$

$$(ii) \quad \text{Dapatan semula (recall)} = \frac{TP}{TP + FN}$$

- a. Positif Benar (TP) - klasifikasi kelas adalah positif dan ramalan model pengelas adalah positif.
- b. Negatif Benar (TN) - klasifikasi kelas adalah negatif dan ramalan model pengelas adalah negatif.
- c. Positif Palsu (FP) - klasifikasi kelas adalah negatif tetapi ramalan model pengelas adalah positif.
- d. Negatif Palsu (FN) - klasifikasi kelas adalah positif tetapi ramalan model pengelas adalah negatif.

5.4.5 Ketepatan Model

```
#Get the accuracy of the model
print("Naive Bayes Accuracy Score -> ",accuracy_score(predictions_NB, test_y)*100)

Naive Bayes Accuracy Score -> 72.72727272727273
```

Rajah 5.13 Ketepatan Model

Cara mengira ketepatan (*accuracy*)

$$(i) \quad \text{Ketepatan (accuracy)} = \frac{TP + TN}{TP + TN + FP + FN}$$

Selepas penghasilan model, periksa ketepatan menggunakan nilai sebenar dan nilai ramalan. Ketepatan *Naïve Bayes* dalam algoritma analisis sentimen yang berfokuskan emoji adalah 72.73%.

5.5 LIPATAN K RAWAK (*K- FOLD RANDOM TESTING*)

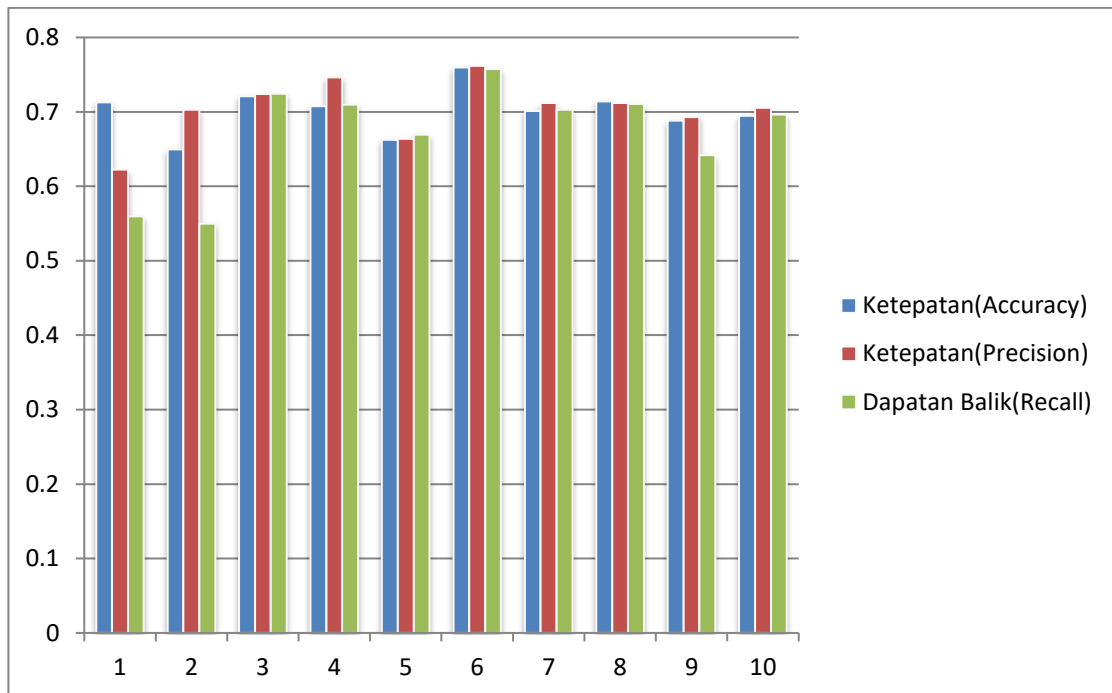
Jadual 5.1 Lipatan K-Rawak

Iterasi	1	2	3	4	5
Ketepatan (<i>Accuracy</i>)	0.7143	0.6494	0.7208	0.7078	0.6523
Ketepatan (<i>Precision</i>)	0.6227	0.7028	0.7242	0.7463	0.6614
Dapatan semula (<i>Recall</i>)	0.5598	0.6497	0.7245	0.7098	0.6696

Iterasi	6	7	8	9	10
Ketepatan <i>(Accuracy)</i>	0.7597	0.7013	0.7143	0.6883	0.6948
Ketepatan <i>(Precision)</i>	0.7618	0.7118	0.7118	0.6929	0.7056
Dapatan semula <i>(Recall)</i>	0.7574	0.7030	0.7108	0.6417	0.6963

Jadual 5.2 Purata Lipatan K-Rawak Naive Bayes

	Purata
Ketepatan <i>(Accuracy)</i>	0.7005
Ketepatan <i>(Precision)</i>	0.7041
Dapatan semula <i>(Recall)</i>	0.6849



Rajah 5.14 Visualisasi Lipatan K-Rawak Naive Bayes

Ketepatan algoritma ini diperoleh daripada perbandingan hasil lapisan *batch*, bahagian pertama dengan model yang dihasilkan, iaitu *Naive Bayes*. Model yang dihasilkan ini boleh digunakan untuk penggunaan masa depan jika teks yang hendak ditentukan sentimen ini daripada domain yang sama. Setiap domain mempunyai penentuan sentimen yang berbeza. Oleh itu, jika ingin menentu sentimen untuk teks yang mengandungi emoji daripada domain lain, model ini harus dilatih dan dihasilkan terlebih dahulu. Dengan penggunaan model setiap sentimen teks yang mengandungi emoji ini dapat ditentukan dengan betul dan tepat dalam masa yang sangat singkat. Walaupun jumlah teks yang hendak ditentukan sentimen ini besar dengan penggunaan model ini, masa yang diperuntukkan untuk penghasilan sentimen setiap teks singkat.

K-fold random testing adalah salah satu cara untuk mendapatkan purata ketepatan model yang dihasilkan. Ketepatan (*accuracy*) adalah 0.70, ketepatan (*precision*) adalah 0.70 dan dapatan balik (*recall*) untuk model *Naive Bayes* model adalah 0.68.


5.6 PERBANDINGAN ANALISIS SENTIMEN YANG MENGGUNAKAN EMOJI DAN TANPA MENGGUNAKAN EMOJI

Jadual 5.3 Perbandingan Analisis sentimen yang menggunakan emoji dan tanpa menggunakan emoji

	Ketepatan (Precision)	Dapatan Semula (Recall)	Ketepatan (Accuracy)
Analisis sentimen menggunakan emoji	0.7041	0.6849	0.7005
Analisis sentimen tanpa menggunakan emoji	0.5689	0.4411	0.5676

Ketepatan analisis sentimen yang menggunakan emoji menunjukkan ketepatan, 0.70, yang lebih tinggi berbanding dengan analisis sentimen tanpa menggunakan emoji, 0.57. Dengan ini analisis sentimen yang menggunakan emoji terbukti lebih tepat berbanding kepada analisis sentimen yang tanpa menggunakan emoji. Hal ini kerana, data daripada media sosial sangat bergantung kepada emoji yang digunakan.

Contohnya :-

 Vishnupriya R @PriyaVishnu16 Cheese cake 🍷 7:27 AM · Jun 22, 2020 · Twitter Web App		Sentimen
	Analisis sentimen menggunakan emoji	Positif
	Analisis sentimen tanpa menggunakan emoji	Neutral

Rajah 5.15 Contoh tweet dan perbandingannya

Analisis sentimen tanpa menggunakan emoji akan menjangka sentimen tweet tersebut neutral hal ini disebabkan analisis ini menggunakan hanya menggunakan ['cheese',

‘cake’] untuk menjangka sentimen. Skor tweet tersebut adalah 0 jika menggunakan analisis sentimen tanpa emoji.

Analisis sentimen yang menggunakan lebih tepat kerana analisis ini menggunakan [‘cheese’, ‘cake’, ‘\xF0\x9F\x98\x8D’] untuk menjangka sentimen. Emoji tersebut mempunyai skor sentimen yang positif iaitu 0.678. Skor tweet tersebut adalah 0.678 jika menggunakan analisis sentimen emoji. Oleh itu, jangkaan sentimen tweet tersebut adalah positif.

5.7 KESIMPULAN

Konklusinya, algoritma analisis sentimen yang berfokuskan emoji ini memberi ketepatan yang lebih tinggi daripada analisis sentimen yang menganalisis teks sahaja. Dengan perbandingan ini, kepergantungan sentimen teks kepada emoji dapat dikenal pasti dan analisis sentimen terhadap teks media sosial ini harus menggunakan algoritma analisis sentimen yang mengambilkira emoji yang digunakan. Algoritma analisis sentimen yang berfokuskan emoji ini memberi ketepatan 70.05% dengan penggunaan model pengelas *Naive Bayes*. Selain itu, dengan penggunaan model setiap sentimen teks yang mengandungi emoji ini dapat ditentukan dengan betul dan tepat dalam masa yang sangat singkat. Walaupun jumlah teks yang hendak ditentukan sentimen ini besar dengan penggunaan model ini, masa yang diperuntukkan untuk penghasilan sentimen setiap teks singkat.

BAB VI

KESIMPULAN

6.1 GAMBARAN KESELURUHAN

Kajian ini berdasarkan analisis sentimen. Pembangunan algoritma ini adalah untuk kegunaan analisis sentimen untuk teks yang mengandungi emoji. Terdapat banyak algoritma yang telah dibangunkan dan sedang digunapakai untuk analisis sentimen. Tetapi algoritma analisis sentimen yang dapat menunjukkan kekutuban untuk teks yang mengandungi emoji masih kurang. Oleh itu, pembangunan algoritma ini adalah untuk teks yang mengandungi emoji. *Python 3* digunakan untuk pembangunan algoritma ini dan kaedah yang digunakan adalah kaedah pembelajaran mesin dan pembelajaran mesin terselia.

6.2 KEKANGAN

Kekangan dalam kajian ini adalah unsur sindiran. Pembangunan algoritma ini tidak dapat menganalisis teks yang berunsur sindiran. Hal ini kerana setiap teks yang dianalisis mengikut makna literalnya. Algoritma ini tidak diaturcara untuk mengenalpasti unsur sindiran. Tetapi, teks dan ayat di media sosial banyak menggunakan unsur sindiran untuk meluahkan perasaan.

6.3 PEMBANGUNAN KAJIAN PADA MASA DEPAN

Algoritma yang akan dibangunkan ini akan menunjukkan sentimen sesebuah teks yang mengandungi emoji iaitu mengeluarkan output sama ada positif, neutral atau negatif mengikut sentimen sesebuah teks. Unsur yang boleh ditambah untuk keberkesanan output ini adalah menunjukkan jumlah perkataan dalam ayat tersebut.

Selain itu, dalam teks media sosial terdapat banyak perkataan yang bermula dengan #. Dalam algoritma ini, semua perkataan yang bermula dengan *hashtag* tidak diambilkira. Unsur ini boleh ditambah untuk meningkatkan ketepatan algoritma.

6.4 KESIMPULAN

Secara konklusinya, bab ini menghuraikan batasan yang dihadapi sepanjang kajian ini dan cadangan untuk menambahbaikkan algoritma ini untuk penggunaan ramai. Algoritma ini boleh digunapakai oleh pengguna yang ingin menganalisis teks yang mengandungi emoji dengan tepat dan efisien. Setiap domain yang ingin dianalisis sentimen ini mempunyai penggunaan tatabahasa dan penentuan sentimen berbeza. Dengan algoritma ini, model yang berbeza untuk setiap domain dapat dihasilkan untuk penentuan sentimen. Model-model yang dihasilkan ini boleh disimpan untuk penggunaan masa depan. Walaupun dataset yang ingin dianalisis sentimen ini besar, algoritma analisis sentimen ini akan memberi output dengan ketepatan yang tinggi dalam masa yang singkat.

















RUJUKAN












- A. Pak and P. Paroubek, 2010. Twitter as a Corpus for Sentimen Analysis and Opinion Mining. In 7th Conference on International Language Resources and Evaluation (LREC 2010), pages 1320–1326.
- Bing Liu, 2012. Sentimen analysis and opinion mining. Morgan & Claypool Publishers: 7 -8.
- C. Manning, T. Grow, T. Grenager, J. Finkel, and J. Bauer. Stanford Tokenizer, 2010. Available online, <http://nlp.stanford.edu/software/tokenizer.shtml>
- De Smet, W., & Moens, M. F. (2007). Generating a topic hierarchy from dialect texts. In DEXA Workshops (pp. 249–253). IEEE Computer Society.
- Emoji Sentimen Ranking. http://kt.ijs.si/data/Emoji_sentimen_ranking/
- Finn, A., & Kushmerick, N. ,2003. Learning to classify documents according to genre. Journal of the American Society for Information Science, 57, 1506–1518. Special issue on Computational Analysis of Style
- Freund, Y., Seung, H. S., Shamir, E., & Tishby, N. ,1997. Selective sampling using the query by committee algorithm. Machine Learning, 28, 133–168.
- Godsay, M. 2015. The process of sentimen analysis: A study. International journal of computer application (0975 – 8887) , 126(7) : 26 – 30.
- J. Read. Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentimen Classification. In Student Research Workshop at the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pages 43–48. Association for Computational Linguistics, 2005.
- K. Liu, W. Li, and M. Guo, 2012. Emoticon Smoothed Language Models for Twitter Sentimen Analysis. In 26th AAAI Conference on Artificial Intelligence (AAAI 2012), pages 1678–1684. Association for the Advancement of Artificial Intelligence, 2012
- List of emoticons. https://en.wikipedia.org/wiki/List_of_emoticons , 04 2016.
- Michele Di Capua, Alfredo Pertosino, Emanuel Di Nardo, September 2015. An architecture for sentimen analysis in Twitter. International Conference on E-Learning: 3–7.
- Nasukawa, Tetsuya and Jeonghee Yi. Sentimen analysis: Capturing favorability using natural language processing. in Proceedings of the KCAP-03, 2nd Intl. Conf. on Knowledge Capture. 2003.
- Ozan Irsoy and Claire Cardie. 2014. Opinion mining with deep recurrent neural networks. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-14), pages 720-728.

- R. Shepard. Recognition Memory for Words, Sentences, and Pictures. Journal of Verbal Learning and Verbal Behavior, 6(1):156–163, 1967.
- Shiho Hashimoto (1 June 2017). 7 Benefits of sentiment analysis. <https://blog.insightsatlas.com/7-benefits-of-sentiment-analysis-you-cant-overlook>
- Vaalmeekam Karthik, Dheeraj Nair, Anuradha J, 2018. Opinion mining on emojis using deep learning techniques. Procedia Computer Science 132: 167-173.
- W Parrott, 2001. Emotions in social psychology. Essential readings. Psychology Press.
- Wieslaw Wolny, 2016. Emotion analysis of twitter data that use emoticons and emoji ideograms. 25TH International Conference On Information Systems Development.

LAMPIRAN A

SKOR EMOJI YANG DIGUNAKAN

Char	Image [twemoji]	Unicode codepoint	Occurrences [5...max]	Position [0...1]	Neg [0...1]	Neut [0...1]	Pos [0...1]	Sentiment score [-1...+1]
😂		0x1f602	14622	0.805	0.247	0.285	0.468	0.221
♥		0x2764	8050	0.747	0.044	0.166	0.790	0.746
♥		0x2665	7144	0.754	0.035	0.272	0.693	0.657
😍		0x1f60d	6359	0.765	0.052	0.219	0.729	0.678
😭		0x1f62d	5526	0.803	0.436	0.220	0.343	-0.093
😘		0x1f618	3648	0.854	0.053	0.193	0.754	0.701
😊		0x1f60a	3186	0.813	0.060	0.237	0.704	0.644
👉		0x1f44c	2925	0.805	0.094	0.249	0.657	0.563
😞		0x1f629	1808	0.826	0.591	0.186	0.223	-0.368
🙏		0x1f64f	1539	0.794	0.081	0.421	0.498	0.417
✌		0x270c	1534	0.790	0.113	0.310	0.576	0.463
😊		0x1f60f	1522	0.765	0.112	0.444	0.444	0.332
😊		0x1f609	1521	0.845	0.100	0.337	0.563	0.463
🙌		0x1f64c	1506	0.791	0.101	0.238	0.661	0.559
🙈		0x1f648	1456	0.739	0.164	0.241	0.596	0.432
👊		0x1f4aa	1409	0.807	0.072	0.301	0.627	0.555

		0x1f633	846	0.797	0.327	0.327	0.345	0.018
		0x1f497	836	0.800	0.051	0.241	0.708	0.657
★	★	0x2605	828	0.353	0.031	0.655	0.314	0.283
■	■	0x2588	798	0.634	0.090	0.853	0.057	-0.032
*		0x2600	786	0.546	0.028	0.479	0.493	0.465
		0x1f621	756	0.862	0.532	0.108	0.360	-0.173
		0x1f60e	754	0.766	0.106	0.297	0.597	0.491
		0x1f622	749	0.814	0.384	0.225	0.391	0.007