https://doi.org/10.17576/apjitm-2020-0901-08
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 9 No. 1, June 2020: 94 - 109

Received: 9 April 2020
Accepted: 31 May 2020
Published: 15 June 2020

CORPUS DEVELOPMENT FOR MALAY SENTIMENT ANALYSIS USING SEMI SUPERVISED APPROACH

EZUANA SUKAWAI
NAZLIA OMAR

ABSTRACT

Research on sentiment analysis have gained so much interest currently. However, research on Malay sentiment analysis and the availability of the resources is still lacking. The aim of this study is to develop a Malay sentiment corpus using a semi-supervised approach. Data from Twitter have been used in this corpus development. The corpus is developed using the combination of lexicon and machine learning approach. The sentiment lexicon will be used to build the seed of training data from unlabelled data resources. In addition, sentiment emoticons are used to compare the accuracy of the lexicon-based approach. After preparation of training data set, the process of adding new training data instances will be carried out using the seed set and machine learning classification method. The process of classification using machine learning approach consists of pre-processing, feature extraction and classification. Five types of classifiers are considered for the classification task. Based on the experimental results, the lexicon-based approach and Multinomial Naïve Bayes algorithm is the best classifier for Malay sentiment corpus development.

Keywords: Corpus, Sentiment Analysis, Sentiment Lexicon, Classification, Semi Supervised, Twitter

PEMBANGUNAN KORPUS BAGI ANALISIS SENTIMEN DALAM BAHASA MELAYU SECARA SEPARA SELIA

ABSTRAK

Kebelakangan ini, kajian tentang analisis sentimen semakin mendapat tempat dan banyak dijalankan. Walau bagaimanapun, kajian berkenaan analisis sentimen dalam Bahasa Melayu masih kurang. Tujuan kajian ini adalah untuk mencadangkan suatu kajian tentang pembangunan korpus analisis sentimen dalam Bahasa Melayu dengan menggunakan kaedah separa selia dijalankan. Korpus analisis sentimen yang menggunakan data daripada Twitter ini dibangunkan dengan menggunakan dua gabungan pendekatan iaitu sentimen leksikon dan pembelajaran mesin. Pemprosesan bahasa tabii digunakan pada peringkat awal. Pada peringkat ini, leksikon sentimen digunakan untuk membuat pengelasan data yang akan dijadikan benih data latihan. Selain itu, emotikon sentimen turut digunakan untuk membandingkan ketepatan keputusan antara leksikon sentimen dan emotikon sentimen. Selepas benih data latihan telah disediakan, proses penambahan data latihan baru yang lebih besar kuantitinya akan dijalankan dengan menggunakan kaedah pengelasan menggunakan benih data dan pembelajaran mesin. Secara ringkasnya, proses pengelasan yang dijalankan dengan menggunakan pembelajaran mesin adalah merangkumi pra-pemprosesan, pengekstrakan fitur dan pengelasan. Perbandingan antara pengelasan juga dijalankan dengan menggunakan lima jenis algoritma. Berdasarkan keputusan ekperimen yang telah dijalankan, penggunaan pendekatan leksikon sentimen dan algoritma pengelas Bayes Naif Multinomial adalah pengelas yang terbaik untuk pembangunan korpus analisis sentimen dalam Bahasa Melayu ini.

Kata Kunci: Korpus, Analisis Sentimen, Leksikon Sentimen, Pengelasan, Separa Selia, Twitter

PENGENALAN

Pelbagai kajian yang menggunakan ulasan atau komen daripada media sosial seperti Facebook dan Twitter telah dijadikan sebagai sumber utama untuk mendapatkan maklumat terus daripada

orang awam atau pelanggan. Antara kajian yang telah dijalankan yang menggunakan data ulasan atau komen ini adalah kajian yang telah ditulis oleh Chumwatana (2018) tentang ulasan pengguna berkaitan produk dan perkhidmatan. Selain itu, Gohil et al. (2018) juga telah menjalankan kajian ulasan tentang penjagaan kesihatan. Manakala Sharma dan P. Shetty (2018) pula telah menjalankan kajian ulasan yang semakin popular pada masa kini iaitu ulasan atau komen orang ramai berkaitan politik.

Media sosial kini telah dianggap sebagai alat untuk orang ramai melahirkan komen mereka secara positif atau negatif. Kajian yang dijalankan oleh Chumwatana menunjukkan sebanyak 510 komen dipaparkan setiap 60 saat di Facebook dan 4.75 bilion kandungan dikongsi setiap hari. Ulasan dalam talian ini membuka era baharu bagi perisikan perniagaan dan pemasaran dalam talian di dunia hari ini. Kata-kata yang dinyatakan di dalam media sosial yang terkandung dalam ulasan pelanggan boleh dianggap sebagai faktor utama untuk menghakimi kepuasan pelanggan.

Justeru itu, suatu kajian akan dijalankan dengan menggunakan sumber data daripada Twitter. Kajian yang akan dijalankan ini adalah pembangunan sebuah korpus analisis sentimen dalam Bahasa Melayu yang menggunakan data ulasan Twitter. Pembangunan ini akan menggunakan pendekatan separa selia iaitu dengan menggunakan gabungan leksikon sentimen dan pengelasan data.

KAJIAN LEPAS

Kajian analisis sentimen dalam Bahasa Inggeris dan lain-lain bahasa telah banyak dijalankan berbanding dengan kajian analisis sentimen dalam Bahasa Melayu. Oleh itu, beberapa kajian korpus analisis sentimen telah dijadikan sebagai sumber rujukan. Antara korpus tersebut adalah korpus analisis sentimen dalam bahasa Jerman yang dibangunkan oleh Flender dan Gips (2017). Dalam kajian mereka, mereka telah mencapai prestasi terbaik dengan menggunakan vektor fitur apabila klasifikasi SVM turut digunakan. Dalam kajian tersebut, didapati juga gabungan n hingga m-gram membawa kepada ketepatan yang lebih baik daripada hanya menggunakan n-gram sahaja. Selain itu, penggunaan langkah pra-pemprosesan, seperti penghapusan pangkasan dan kata henti tidak memberikan banyak kesan.

Brum dan Nunes (2018) pula telah membangunkan korpus analisis sentimen dalam bahasa Brazil Portugis dengan menggunakan domain novel untuk bahasa Brazil Portugis yang dapat dieksploitasi dengan pendekatan pembelajaran mesin. Korpus ini menjadi sumber baru untuk pendekatan linguistik dalam bahasa tabii dengan menjalankan pemerhatian ke atas ekspresi, tingkah laku media sosial atau pengesanan ucapan kebencian. Selain itu, korpus ini berbeza dengan yang lain kerana ia menggunakan polariti neutral.

Manakala kajian yang dijalankan oleh Adel et al. (2016) dalam pembangunan korpus analisis sentimen dalam dialek Arab Saudi adalah berbeza dengan yang lain. Ini adalah kerana kajian mereka lebih memfokus kepada cadangan kaedah untuk menganotasi secara manual oleh manusia untuk pembangunan korpus. Walau bagaimanapun, tujuan tetap sama iaitu pembangunan korpus adalah untuk digunakan dalam analisis sentimen dengan menggunakan ulasan yang bersumberkan Twitter. Korpus yang dibangunkan ini menjadi korpus yang pertama dikeluarkan secara terbuka.

Seterusnya adalah kajian pembangunan korpus analisis sentimen dalam Bahasa Indonesia oleh Wicaksono et al. (2014). Mereka telah mencadangkan kaedah pembangunan contoh latihan secara automatik untuk analisis sentimen dan perlombongan pendapat bagi twit Indonesia. Kaedah yang dicadangkan telah mengatasi sistem garis dasar yang hanya menggunakan emotikon sebagai ciri untuk membina korpus sentimen secara automatik. Model pengelas yang dilatih dalam data latihan menggunakan kaedah yang dicadangkan dapat mengekstrak twit pendapat dan pengelas polariti twit dengan prestasi yang tinggi. Selain itu,

kajian mereka turut membuktikan teknik pembangunan benih korpus adalah aspek penting dalam kaedah ini apabila penilaian menunjukkan bahawa pengetahuan sedia ada daripada leksikon pendapat dapat membantu membina contoh latihan yang lebih baik daripada hanya menggunakan teknik berasaskan pengklusteran.

Kajian lepas yang terakhir sekali adalah kajian yang dijalankan oleh Pak dan Paroubek (2010) dalam pembangunan korpus analisis sentimen dalam Bahasa Inggeris. Kajian mereka adalah untuk menunjukkan kaedah pengumpulan korpus secara automatik yang boleh digunakan untuk melatih pengelas sentimen. Selain itu, pengelas tersebut boleh menentukan polariti dokumen sama ada positif, negatif dan neutral. Berdasarkan kajian mereka, pengelasan yang digunakan adalah berdasarkan pengelas Bayes Naif Multinomial yang menggunakan N-gram dan tag POS sebagai fitur. Kajian mereka ini juga kerap menjadi garis dasar untuk perbandingan dengan kajian terbaru.

METODOLOGI

Secara umumnya, korpus dalam kajian ini akan dibangunkan dengan menggabungkan dua kaedah iaitu kaedah leksikon sentimen dan kaedah pengelasan yang menggunakan pendekatan pembelajaran mesin. Oleh itu, setiap kaedah di dalam kajian ini akan diterangkan secara terperinci yang akan melibatkan beberapa proses. Proses-proses yang terlibat adalah seperti pengekstrakan data, pengumpulan data, pra-pemprosesan, pembinaan benih data latihan, penambahan data latihan baru serta penilaian dan pengujian. Selain itu, proses terakhir yang akan dijalankan dalam kajian ini adalah penggabungan antara benih data latihan dan data latihan baru.

PENGEKSTRAKAN DATA

Pengekstrakan data telah dijalankan dengan menggunakan aplikasi RapidMinner yang dihubungkan kepada aplikasi Twitter. Domain data yang telah dipilih untuk kajian ini merupakan domain politik di Malaysia dan bahasa yang telah ditetapkan di dalam skop kajian ini adalah Bahasa Melayu. Data bagi domain ini telah diambil dengan menggunakan carian kata kunci berdasarkan parti politik, ahli politik dan isu semasa berkaitan politik di Malaysia.

PENGUMPULAN DATA

Selepas pengekstrakan data dijalankan, data dikumpulkan dan dibahagikan kepada empat set data. Kaedah pembahagian data ini adalah berdasarkan kajian yang dijalankan oleh Wicaksono et al. (2014). Setiap set data yang dibahagikan ini mempunyai tujuannya seperti yang ditunjukkan di Jadual 1.

JADUAL 1. Pembahagian set data awal

Bil.	Set Data	Jumlah Ulasan	Penerangan	Penghasilan Data
1	Set data 1	5000	Set data contoh yang tidak mempunyai label	Data latihan baru
2	Set data 2	2299	Set data yang akan digunakan untuk menjadi benih data	Benih data latihan
3	Set data 3	5000	Set data yang dianotasi secara manual	Set data pengujian
4	Set data 4	5000	Set data untuk analisis sentimen	Set data analisis sentimen
	Jumlah	17,299	Jumlah keseluruhan data ulasan yang telah diekstrak	

Set data 1 merupakan set data yang tidak berlabel yang akan digunakan untuk menjalankan proses pengelasan bagi data baru. Set data 2 juga merupakan set data yang tidak berlabel tetapi ia akan digunakan untuk proses pengelasan yang akan menggunakan kaedah pembenihan data dan akan menjadi set benih data.

Manakala set data 3 merupakan set data yang akan digunakan untuk melakukan pengujian dan penilaian. Set data 3 ini dianotasi secara manual oleh pakar dalam Bahasa Melayu. Selepas anotasi dijalankan, hanya 2862 data sahaja yang dapat digunakan dalam kajian ini. Ini adalah kerana hanya 299 sahaja ulasan yang mempunyai polariti positif dan 2563 ulasan yang mempunyai polariti negatif. Selebihnya adalah neutral yang mana tidak digunakan dalam kajian ini. JADUAL 2 menunjukkan polariti positif dan negatif bagi set data 3 selepas dianotasi secara manual.

JADUAL 2. Polariti bagi Set data 3

Positif	Negatif	Jumlah
229	2563	2862

PRA-PEMPROSESAN

Pra-pemprosesan merupakan suatu proses awal yang penting dalam menjalankan kajian berkaitan pembelajaran mesin. Pra-pemprosesan membantu dalam meningkatkan ketepatan membuat pengelasan atau ramalan. Oleh itu, pra-pemprosesan yang dijalankan dalam kajian ini dibahagikan kepada tiga proses utama iaitu proses pembersihan, pentokenan dan normalisasi.

Pada permulaan, proses pembersihan data ulasan dijalan terlebih dahulu. Proses ini hanya menggunakan aplikasi Microsoft Excel sahaja. Aktiviti yang dilakukan dalam proses pembersihan ini adalah termasuk memeriksa ulasan sepintas lalu dan menghapuskan ulasan sekiranya terdapat bahasa selain Bahasa Melayu. Selain itu, pautan URL, akaun pengguna Twitter yang bermula dengan tanda '@', maklumat retweet (RT), perkataan bermula dengan tanda #, selang baris dan ayat pendua juga dihapuskan.

Selepas proses pembersihan, proses seterusnya adalah proses pentokenan. Menurut Lourdusamy dan Abraham (2018), pentokenan adalah satu kaedah untuk membahagikan sekumpulan teks ke dalam elemen bermakna atau token seperti kata, frasa, dan simbol. Senarai koleksi token digunakan sebagai teks input untuk pemprosesan selanjutnya. Oleh itu, setiap kajian berkenaan dengan teks akan menggunakan kaedah pentokenan untuk memisahkan perkataan dan kebanyakan kajian akan menggunakan ruang kosong untuk membuat pentokenan seperti yang dilakukan oleh Wicaksono et al. (2014) dan Pak dan Paroubek (2010).

Akhir sekali, proses yang terlibat dalam pra-pemprosesan adalah proses normalisasi. Proses normalisasi yang digunakan dalam kajian ini untuk kedua-dua proses pembenihan data dan penambahan data latihan baru adalah sama dan ringkas iaitu hanya melibatkan proses menyeragamkan semua perkataan dalam data ulasan kepada huruf kecil.

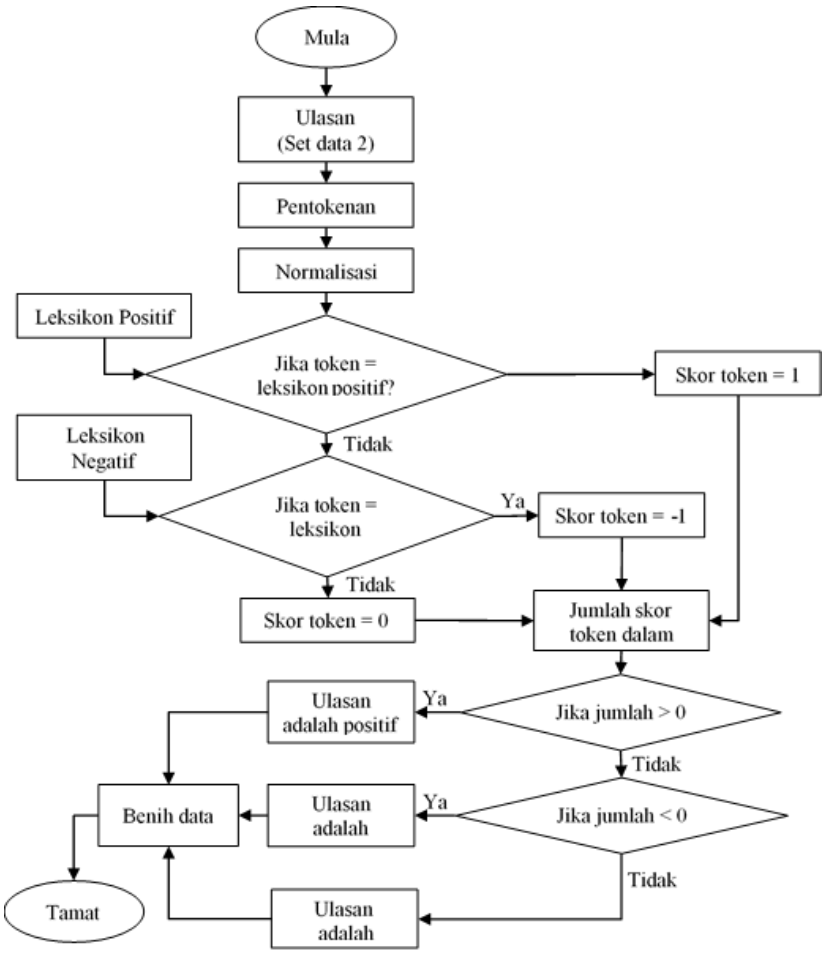
MEMBINA BENIH DATA LATIHAN

Pembenihan data ini adalah proses untuk mewujudkan benih data yang akan menjadi data awal untuk membangunkan sebuah korpus. Kaedah seperti ini telah dijalankan oleh Wicaksono et al. (2014) dalam kajiannya. Bagi tujuan kajian ini, proses pembenihan data ini adalah dengan menggunakan kaedah pemprosesan bahasa tabii. Data yang akan digunakan untuk menjadi benih data adalah set data 2. Manakala leksikon sentimen dan kata adjektif digunakan untuk menentukan label benih data.

Leksikon sentimen yang digunakan dalam kajian ini telah diambil daripada hasil kajian Nur Sharmini dan Nazlia (2017). Selain daripada penggunaan leksikon sentimen, penggunaan kata adjektif turut digunakan. Kata adjektif telah digunakan dalam beberapa kajian seperti Brum dan Nunes (2018), Sharma et al. (2015) dan Modak dan Mondal (2014). Leksikon sentimen dan kata adjektif digabungkan dan dibahagikan mengikut polariti positif dan negatif.

Setiap data ulasan dalam set data 2 akan dipadankan dengan leksikon sentimen. Sekiranya terdapat padanan perkataan yang sama dalam set data 2 dengan leksikon sentimen, perkataan tersebut akan diberikan skor. Padanan untuk leksikon positif adalah 1 dan padanan untuk lesikon negatif adalah -1. Setelah itu, jumlah keseluruhan skor untuk sesuatu data ulasan akan dikira dan dijumlahkan. Proses dan kaedah ini juga adalah berdasarkan kaedah yang dilakukan oleh Wicaksono et al (2014) dalam kajian mereka.

Keseluruhan proses yang terlibat dalam membina benih data latihan ini ditunjukkan seperti di Rajah 1. Rajah 1 merupakan carta aliran proses yang berlaku dalam membina benih data latihan.

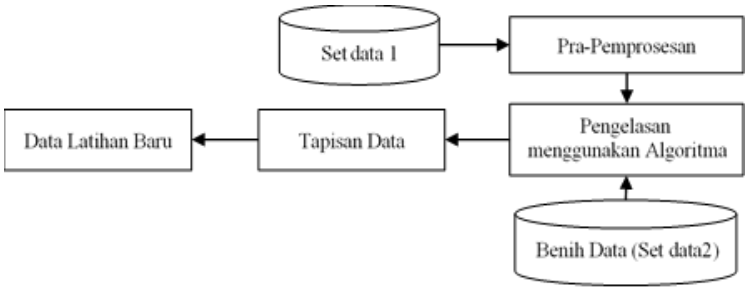


RAJAH 1. Aliran proses bagi pembenihan data latihan

PENAMBAHAN DATA LATIHAN BARU

Proses seterusnya adalah proses untuk menambahkan atau mengembangkan set data latihan yang baru yang lebih besar untuk dijadikan korpus. Set data yang digunakan untuk penambahan data latihan baru adalah set data 1. Set data 1 ini masih belum dianotasi atau dikelaskan ke dalam mana-mana polariti positif atau negatif. Proses penambahan data latihan baru turut

digunakan oleh Wicaksono et al. (2014) dalam kajian mereka. Proses penambahan data latihan baru adalah dengan menggunakan kaedah pengelasan pembelajaran mesin. Proses ini ditunjukkan seperti dalam Rajah 2.



RAJAH 2. Proses pengelasan yang berlaku ke atas set data 1

ALGORITMA

Pemilihan penggunaan pengelas adalah berdasarkan kajian yang dilakukan oleh Brum dan Nunes (2018), Pak dan Paroubek (2010) dan Wicaksono et al. (2014) . Proses pengelasan ini melibatkan empat jenis algoritma pengelasan dan satu algoritma pengklusteran. Algoritma-algoritma tersebut adalah Bayes Naif, Bayes Naif Multinomial, Mesin Sokongan Vektor, Regresi Logistik dan Pemaksimuman Jangkaan.

Bayes Naif adalah teknik pengelasan kebarangkalian. Pengelas ini berfungsi dengan baik apabila digunakan untuk set data yang besar. Pengelas Bayes Naif ini mengira kebarangkalian *posterior* dengan menggunakan formula (1) dan (2) yang diterangkan oleh Ankit dan Saleena (2018) seperti berikut:

$$Kebarangkalian\ posterior = \frac{kemungkinan \times kebarangkalian\ sebelum}{bukti} \tag{1}$$

bersamaan,

$$P(Class_i \mid z) = \frac{p(z \mid Class_i)\ x\ P(Class_i)}{p(z)} \tag{2}$$

dengan z mewakili vektor fitur dan $Class_i$ mewakili kelas ke-i. Pengelas Bayes Naif membuat andaian bahawa fitur adalah kebebasan bersyarat.

Manakala Bayes Naif Multinomial menurut Lohar et al. (2017) adalah contoh khusus pengelas Bayes Naif yang menggunakan pengagihan multinomial untuk setiap fitur dan bukannya merujuk kepada ketidak bergantungan bersyarat bagi setiap fitur dalam model. Dalam kaedah pengelasan ini, pengagihan dianggarkan dengan mempertimbangkan prinsip Bayes Naif generatif, yang mengandaikan bahawa fitur-fitur tersebut diagihkan secara multinomial untuk mengira kebarangkalian dokumen bagi setiap label dan mengekalkan pemaksimuman kebarangkalian label. Dengan mengandaikan kebarangkalian ciri $P(x_i \mid c_j)$ adalah bebas diberikan kelas *c* dan untuk dokumen *d* diwakili sebagai fitur-fitur $x_1, x_2, \ldots, x_n$. Persamaan untuk Bayes Naif Multinomial yang ditunjukkan oleh Lohar et al. (2017) adalah seperti formula (3) berikut:

$$C_{NB} = \arg\max_{c \in C} P(c_j) \prod_{x \in X} P(x|c) \tag{3}$$

Lohar et al. (2017) turut menerangkan persamaan penggunaan pengelas Bayes Naif Multinomial dalam pengelasan teks diwakili seperti formula (4) berikut:

$$C_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_{i \in positions} P(x_i|c_j) \quad (4)$$

Menurut Ankit dan Saleena (2018), Mesin Sokongan Vektor memerlukan data latihan untuk melatih model. Ia juga dikenali sebagai pengelas kebarangkalian. Mesin Sokongan Vektor menggunakan pemetaan bukan linear yang bertujuan untuk mencari margin yang besar antara kelas yang berbeza. Walaupun tempoh masa latihan Mesin Sokongan Vektor boleh menjadi lambat tetapi ia sangat tepat. Mesin Sokongan Vektor mencuba untuk mencari sempadan keputusan yang memaksimumkan jurang pemisahan antara kelas. Tidak seperti pengelas Bayes Naif, Mesin Sokongan Vektor tidak membuat andaian kebebasan bersyarat bagi kelas. Mesin Sokongan Vektor menghasilkan hasil yang baik untuk tugasan bagi permasalah analisis sentimen Twitter. Nilai parameter pengelas Mesin Sokongan Vektor ditunjukkan sebagai C = 0.1, kernel = linear.

Selain itu, Ankit dan Saleena (2018) turut menerangkan tentang Regresi Logistik yang mana Regresi Logistik adalah model regresi yang digunakan untuk tujuan pengelasan. Regresi Logistik biasanya digunakan untuk mengaitkan pembolehubah kebergantungan kategori tunggal kepada satu atau lebih pembolehubah bebas. Regresi Logistik cuba mencari hiper satah yang memaksimumkan jurang pemisahan antara kelas. Nilai parameter pengelas Regresi Logistik ditunjukkan sebagai C = .01, maksimum lelaran = 100.

Algoritma yang terakhir adalah Pemaksimuman Jangkaan. Adebisi et al. (2012) telah menerangkan, Pemaksimuman jangkaan adalah pendekatan berasaskan model untuk menyelesaikan masalah clustering. Ia adalah algoritma iteratif yang digunakan dalam masalah di mana data tidak lengkap atau dianggap tidak lengkap. Tidak seperti algoritma berdasarkan jarak seperti Purata K, Pemaksimuman Jangkaan dikenali sebagai algoritma pengoptimuman yang sesuai untuk membina model statistik data yang betul. Pemaksimuman Jangkaan digunakan secara meluas dalam aplikasi seperti penglihatan komputer, pemprosesan ucapan dan pengenalan corak. Pemaksimuman Jangkaan bermatlamat mencari kluster sedemikian sehingga kemungkinan maksimum setiap parameter kluster diperoleh.

PENGELASAN

Hasil pengelasan yang dijalankan oleh kelima-lima algoritma ini akan dibuat tapisan terlebih dahulu sebelum pengujian dan penilaian dibuat. Hasil keputusan ketepatan akan dibuat perbandingan untuk memilih model yang mempunyai ketepatan yang tertinggi. Proses pengelasan data bagi kajian ini adalah dengan menggunakan aplikasi WEKA. Pengelasan yang dijalankan ini juga menggunakan set data 2 yang merupakan benih data latihan sebagai data latihan untuk melabelkan set data ulasan baru yang tidak berlabel daripada set data 1.

Jadual 3 di bawah menunjukkan keputusan pengelasan dan pengklusteran bagi set data 1 dengan menggunakan lima jenis model algoritma. Setiap model algoritma menjalankan empat kali proses pengelasan atau pengklusteran. Setiap proses ini mempunyai ketetapan yang berbeza-beza yang melibatkan kaedah pengekstrakan fitur. Hasil daripada pengelasan ini menunjukkan, bilangan polariti yang berbeza dihasilkan untuk positif dan negatif. Secara keseluruhan, bilangan data yang mempunyai polariti positif adalah kurang daripada negatif dan bilangan polariti positif ini semakin berkurangan apabila bigram digunakan dalam proses pengelasan atau pengklusteran.

JADUAL 3 Keputusan pengelasan polariti bagi setiap algoritma

Bil	Algoritma	Polariti Positif	Polariti Negatif	Jumlah
1.	Bayes Naif	1554	3446	5000
2.	Bayes Naif Multinomial	1576	3424	5000
3.	Regresi Logistik	2115	2885	5000
4.	Mesin Sokongan Vektor	1517	3483	5000
5.	Pemaksimuman jangkaan	1182	3818	5000

Bagi kaedah pengelasan, aplikasi WEKA digunakan untuk mengkategorikan ulasan mengikut polariti sama ada positif atau negatif. Setiap ulasan akan diberikan juga nilai polarisasi, yang mana nilainya adalah $0 < \in \leq 1$. Walau bagaimanapun, untuk kaedah pengklusteran menggunakan WEKA, nilai polarisasi bagi setiap ulasan tidak dinyatakan.

TAPISAN DATA

Menurut kajian yang dijalankan oleh Wicaksono et al. (2014), nilai empirik adalah $0 < \epsilon < 1$ yang nilai ϵ adalah 0.98. Oleh itu, hanya data yang mempunyai nilai polarisasi sebanyak 0.98 atau lebih akan dipilih. Tujuannya adalah untuk mendapatkan ulasan dalam perkataan sentimen yang sangat berpolarisasi sebagai contoh data latihan baru mereka.

Berdasarkan kajian tersebut, maka kajian ini akan turut menggunapakai kaedah yang digunakan oleh Wicaksono et al. (2014) untuk memilih data ulasan yang mempunyai polarisasi 0.98 dan lebih. Walau bagaimanapun, kaedah ini tidak digunapakai ke atas data daripada pengklusteran memandangkan nilai polarisasi tidak dinyatakan. Jadual 4 menunjukkan bilangan positif dan negatif bagi setiap model algoritma yang mempunyai nilai polarisasi 0.98 atau lebih.

JADUAL 4. Bilangan polariti yang mempunyai nilai polarisasi 0.98 atau lebih

Bil	Pengelas	Polariti Positif	Polariti Negatif	Jumlah
1.	Bayes Naif	728	1142	1870
2.	Bayes Naif Multinomial	273	989	1262
3.	Regresi Logistik	2100	2866	4966
4.	Mesin Sokongan Vektor	1517	3483	5000
5.	Pemaksimuman jangkaan	1182	3818	5000

Merujuk kepada Jadual 4, didapati bahawa majoriti pengelasan data menghasilkan data yang tidak seimbang selain daripada Regresi Logistik. Dalam kajian yang dijalankan oleh Jagdale, Shirsat dan Deshmukh (2016), mereka telah menghasilkan beberapa set data daripada ulasan Twitter yang mempunyai polariti yang tidak seimbang. Begitu juga dengan Brum dan Nunes (2018) yang telah menghasilkan set data yang tidak seimbang dan telah mengambil semua set data tersebut untuk proses pengujian. Oleh itu, semua data yang terhasil daripada pengelas yang berbeza dalam kajian ini akan diambil untuk pengujian.

PENYEIMBANGAN DATA

Sebelum pengujian dijalankan, set data ini diseimbangkan terlebih dahulu kerana ketidak seimbangan pengagihan data antara polariti boleh menjadi masalah untuk kaedah pembelajaran mesin. Oleh itu, dengan merujuk kepada kajian Brum dan Nunes (2018) juga, data yang tidak seimbang akan diseimbangkan dengan mengurangkan bilangan polariti yang banyak mengikut

bilangan polariti yang sedikit. Selepas pengurangan, bilangan data yang telah diseimbangkan ditunjukkan seperti di JADUAL 5.

JADUAL 5. Bilangan polariti yang telah diseimbangkan

Bil	Pengelas	Polariti Positif	Polariti Negatif	Jumlah
1.	Bayes Naif	728	728	1456
2.	Bayes Naif Multinomial	273	273	546
3.	Regresi Logistik	2100	2100	4200
4.	Mesin Sokongan Vektor	1517	1517	3034
5.	Pemaksimuman jangkaan	1182	1182	2364

PENGUJIAN DAN PENILAIAN

Setelah proses penambahan data latihan baru yang menggunakan set data 1 telah berjaya dihasilkan, proses pengujian akan dijalankan ke atas set data tersebut. Pengujian yang dijalankan ini akan menggunakan empat jenis algoritma untuk mencari keputusan ketepatan yang tertinggi iaitu Bayes Naif, Bayes Naif Multinomial, Mesin Sokongan Vektor dan Regresi Logistik. Pengujian ini turut dibahagikan kepada dua tugasan pengujian iaitu:-

1. Pengujian bagi pengelas untuk set data 1 dan
2. Pengujian antara leksikon sentimen dan emotikon sentimen.

PENGUJIAN BAGI PENGELAS UNTUK SET DATA 1

Hasil keputusan pengujian yang ditunjukkan di dalam Jadual 6 hingga Jadual 9 di bawah adalah dengan menggunakan empat jenis algoritma pengelas. Keempat-empat algoritma pengelas yang digunakan untuk pengujian tersebut terdiri daripada Bayes Naif, Bayes Naif Multinomial, Mesin Sokongan Vektor dan Hutan Rawak. Jadual 6 menunjukkan keputusan pengujian set data 1 yang menggunakan algoritma Bayes Naif sebagai pengelas untuk pengujian.

JADUAL 6 Keputusan pengujian set data 1 menggunakan pengelas Bayes Naif

Bil	Pengelas Data Latihan	Peratus Ketepatan	Ukuran F
1.	Bayes Naif	60.58	0.61
2.	Bayes Naif Multinomial	68.09	0.68
3.	Regresi Logistik	60.17	0.65
4.	Mesin Sokongan Vektor	50.08	0.54
5.	Pemaksimuman Jangkaan	54.73	0.57

Berdasarkan Jadual 6, pengelas data latihan yang menggunakan Bayes Naif Multinomial memperolehi peratus ketepatan tertinggi iaitu sebanyak 68.09% dan diikuti oleh Bayes Naif yang memperolehi peratus ketepatan sebanyak 60.58%. Seterusnya adalah Regresi Logistik yang memperolehi peratus ketepatan sebanyak 60.17% dan diikuti oleh Pemaksimuman Jangkaan yang memperolehi peratus ketepatan sebanyak 54.73%. Terakhir sekali adalah Mesin Sokongan Vektor yang memperolehi peratus ketepatan sebanyak 50.08%.

Jadual 6 juga menunjukkan keputusan dalam ukuran F. Keputusan ukuran F yang tertinggi adalah yang menggunakan pengelas Bayes Naif Multinomial yang memperolehi keputusan sebanyak 0.68 dan diikuti oleh Regresi Logistik yang memperolehi keputusan sebanyak 0.65. Seterusnya adalah Bayes Naif yang memperolehi keputusan sebanyak 0.61 dan

diikuti oleh Pemaksimuman Jangkaan yang memperolehi keputusan sebanyak 0.57. Terakhir sekali adalah Mesin Sokongan Vektor yang memperolehi keputusan sebanyak 0.54.

Seterusnya pula adalah penggunaan algoritma Bayes Naif Multinomial sebagai pengelas dalam proses pengujian. Jadual 7 di bawah menunjukkan keputusan pengujian set data 1 yang menggunakan pengelas Bayes Naif Multinomial.

JADUAL 7. Keputusan pengujian set data 1 menggunakan pengelas Bayes Naif Multinomial

Bil	Pengelas Data Latihan	Peratus Ketepatan	Ukuran F
1.	Bayes Naif	68.43	0.68
2.	Bayes Naif Multinomial	72.34	0.72
3.	Regresi Logistik	64.94	0.69
4.	Mesin Sokongan Vektor	62.62	0.66
5.	Pemaksimuman Jangkaan	68.93	0.70

Merujuk kepada Jadual 7, pengelas data latihan yang menggunakan Bayes Naif Multinomial memperolehi peratus ketepatan tertinggi iaitu sebanyak 72.34% dan diikuti oleh Pemaksimuman Jangkaan yang memperolehi peratus ketepatan sebanyak 68.93%. Seterusnya adalah Bayes Naif yang memperolehi peratus ketepatan sebanyak 68.43% dan diikuti oleh Regresi Logistik yang memperolehi peratus ketepatan sebanyak 64.94%. Terakhir sekali adalah Mesin Sokongan Vektor yang memperolehi peratus ketepatan sebanyak 62.62%.

Jadual 7 juga menunjukkan keputusan dalam ukuran F. Keputusan ukuran F tertinggi adalah yang mengunakan pengelas Bayes Naif Multinomial yang memperolehi keputusan sebanyak 0.72 dan diikuti oleh Pemaksimuman Jangkaan yang memperolehi keputusan sebanyak 0.70. Seterusnya adalah Regresi Logistik yang memperolehi keputusan sebanyak 0.69 dan diikuti oleh Bayes Naif yang memperolehi keputusan sebanyak 0.68. Terakhir sekali adalah Mesin Sokongan Vektor yang memperolehi keputusan sebanyak 0.66.

Manakala Jadual 8 menunjukkan penggunaan algoritma Mesin Sokongan Vektor sebagai pengelas dalam proses pengujian. Jadual 7 di bawah turut menunjukkan keputusan pengujian set data 1 yang menggunakan pengelas Mesin Sokongan Vektor.

JADUAL 8 Keputusan pengujian set data 1 menggunakan pengelas Mesin Sokongan Vektor

Bil	Pengelas Data Latihan	Peratus Ketepatan	Ukuran F
1.	Bayes Naif	59.46	0.59
2.	Bayes Naif Multinomial	70.21	0.70
3.	Regresi Logistik	61.89	0.67
4.	Mesin Sokongan Vektor	64.46	0.67
5.	Pemaksimuman Jangkaan	66.57	0.68

Berdasarkan Jadual 8 pula, pengelas data latihan yang menggunakan Bayes Naif Multinomial memperolehi peratus ketepatan tertinggi iaitu sebanyak 70.21% dan diikuti oleh Pemaksimuman Jangkaan yang memperolehi peratus ketepatan sebanyak 66.57%. Seterusnya adalah Mesin Sokongan Vektor yang memperolehi peratus ketepatan sebanyak 64.46% dan diikuti oleh Regresi Logistik yang memperolehi peratus ketepatan sebanyak 61.89%. Terakhir sekali adalah Bayes Naif yang memperolehi peratus ketepatan sebanyak 59.46%.

Jadual 8 juga menunjukkan keputusan dalam ukuran F. Keputusan ukuran F tertinggi adalah pengelas Bayes Naif Multinomial yang memperolehi keputusan sebanyak 0.70 dan diikuti oleh Pemaksimuman Jangkaan yang memperolehi ukuran F sebanyak 0.68. Seterusnya adalah Regresi Logistik dan Mesin Sokongan Vektor yang memperolehi keputusan yang sama

iaitu sebanyak 0.67. Terakhir sekali adalah Bayes Naif yang memperolehi keputusansebanyak 0.66.

Jadual 9 di bawah menunjukkan penggunaan algoritma Hutan Rawak sebagai pengelas dalam proses pengujian. Jadual ini turut menunjukkan keputusan pengujian ke atas set data 1 yang menggunakan pengelas Hutan Rawak.

JADUAL 9 Keputusan set data 1 menggunakan pengelas Hutan Rawak

Bil	Pengelas Data Latihan	Peratus Ketepatan	Ukuran F
1.	Bayes Naif	63.30	0.63
2.	Bayes Naif Multinomial	69.36	0.69
3.	Regresi Logistik	62.56	0.67
4.	Mesin Sokongan Vektor	64.15	0.67
5.	Pemaksimuman Jangkaan	69.43	0.70

Merujuk kepada keputusan pengujian di Jadual 9, pengelas data latihan yang menggunakan Pemaksimuman Jangkaan memperolehi peratus ketepatan tertinggi iaitu sebanyak 69.43% dan diikuti oleh Bayes Naif Multinomial yang memperolehi peratus ketepatan sebanyak 69.36%. Seterusnya adalah Mesin Sokongan Vektor yang memperolehi peratus ketepatan sebanyak 64.15% dan diikuti oleh Bayes Naif yang memperolehi peratus ketepatan sebanyak 63.30%. Terakhir sekali adalah Regresi Logistik yang memperolehi peratus ketepatan sebanyak 62.56%.

Selain itu, Jadual 9 juga menunjukkan keputusan dalam ukuran F. Keputusan ukuran F tertinggi adalah yang menggunakan pengelas Pemaksimuman Jangkaan dengan memperolehi keputusan sebanyak 0.70 dan diikuti oleh Bayes Naif Multinomial yang memperolehi keputusan sebanyak 0.69. Seterusnya adalah Regresi Logistik dan Mesin Sokongan Vektor yang memperolehi keputusan yang sama iaitu sebanyak 0.67. Terakhir sekali adalah Bayes Naif yang memperolehi keputusan sebanyak 0.63.

PENGUJIAN ANTARA LEKSIKON SENTIMEN DAN EMOTIKON SENTIMEN

Pengujian ini akan memberi fokus kepada keputusan peratus ketepatan dan ukuran F yang diperolehi daripada proses pengujian data latihan baru yang telah dibangunkan dengan menggunakan leksikon sentimen dan emotikon sentimen. Proses pengujian ini dimulakan dengan pengelasan data menggunakan emotikon sentimen seperti yang ditunjukkan di JADUAL 10.

JADUAL 10. Bilangan benih data dan data latihan daripada emotikon sentimen

Bil.	Data	Positif	Negatif	Jumlah
1.	Set data 2	14	19	33
2.	Set data 1	2916	2084	5000
3.	Set data 1 (selepas tapisan)	757	175	932

Daripada Jadual 10 ini, sebanyak 33 benih data daripada set data 2 telah dihasilkan dengan menggunakan emotikon sentimen. Daripada keseluruhan benih data ini, sebanyak 14 daripadanya adalah data yang berpolariti positif dan 19 lagi adalah berpolariti negatif. Kemudian, pengelasan untuk penambahan data latihan baru telah dijalankan dengan menggunakan benih data. Hasil daripada pengelasan tersebut, sebanyak 2916 polariti positif dan 2084 polariti negatif telah dihasilkan.

Seterusnya, tapisan berdasarkan polarisasi dibuat bagi memastikan kualiti data latihan yang mana akhirnya menghasilkan jumlah keseluruhan data adalah 932 dengan pecahan polariti positif adalah sebanyak 757 dan polariti negatif adalah sebanyak 175. Pecahan polariti ini dilihat tidak seimbang. Oleh itu, polariti yang mempunyai bilangan yang tinggi dikurangkan mengikut jumlah polariti yang sedikit.

PENGGABUNGAN BENIH DATA DAN DATA LATIHAN BARU

Selepas proses pengelasan set data 1 yang menghasilkan data latihan baru telah selesai sehingga ke peringkat pengujian, set data 1 ini digabungkan dengan set data 2 yang merupakan benih data. Penggabungan ini adalah untuk mendapatkan set data latihan yang lebih besar dalam korpus analisis sentimen dalam Bahasa Melayu ini. Pembelajaran daripada data latihan yang besar memainkan peranan penting dalam tugas mengklasifikasikan polariti (Wicaksono et al., 2014).

Setelah penggabungan dijalankan, satu lagi pengujian akan dijalankan ke atas gabungan data latihan ini. Tujuan pengujian ini adalah untuk menguji peratus ketepatan dan ukuran F bagi pengelasan yang akan dijalankan dengan menggunakan gabungan data latihan. Oleh itu, set data 4 yang merupakan set data baru yang belum dianotasi digunakan dalam proses ini untuk dibuat pengelasan dengan menggunakan gabungan data latihan. JADUAL 11 di bawah menunjukkan hasil keputusan pengujian tersebut.

JADUAL 11. Keputusan pengujian set data 4 menggunakan data latihan gabungan

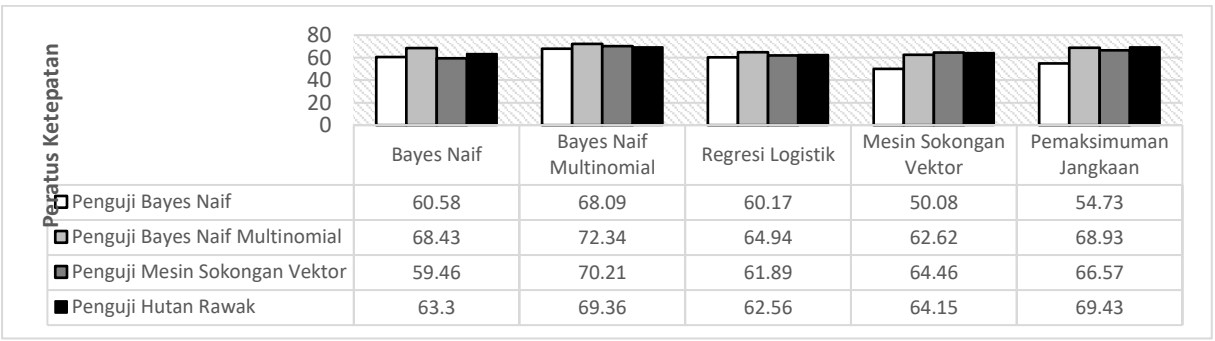
Peratus Ketepatan	Ukuran -F	Ukuran-F Positif	Ukuran-F Negatif
91.905	0.92	0.69	0.95

KEPUTUSAN DAN PERBINCANGAN

Penilaian dijalankan ke atas pengujian yang telah dibahagikan kepada dua tugasan pengujian bagi pengelas untuk set data 1 serta pengujian antara leksikon sentimen dan emotikon sentimen. Selain itu, penilaian juga dibuat ke atas pengujian penggabungan data latihan.

PENILAIAN PENGUJIAN BAGI PENGELAS UNTUK SET DATA 1

Penilaian yang dijalankan adalah untuk mendapatkan model yang mempunyai keputusan peratusan ketepatan dan nilai ukuran F yang paling tinggi. Berdasarkan pengujian yang telah dijalankan dengan menggunakan empat jenis algoritma, terdapat satu sahaja algoritma pengujian yang memperolehi peratusan ketepatan yang melebihi 70% dan nilai ukuran F yang melebihi 0.7. Rajah 3 menunjukkan graf keputusan peratus ketepatan bagi pengujian untuk pengelasan set data 1.

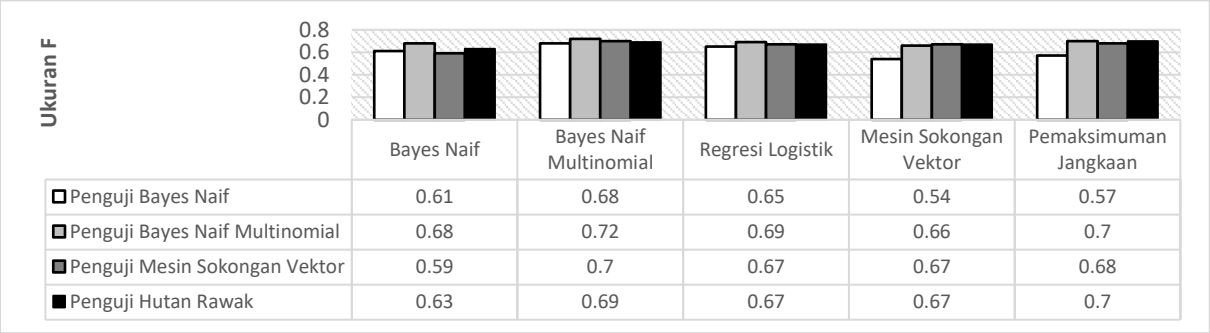


	Bayes Naif	Bayes Naif Multinomial	Regresi Logistik	Mesin Sokongan Vektor	Pemaksimuman Jangkaan
Penguji Bayes Naif	60.58	68.09	60.17	50.08	54.73
Penguji Bayes Naif Multinomial	68.43	72.34	64.94	62.62	68.93
Penguji Mesin Sokongan Vektor	59.46	70.21	61.89	64.46	66.57
Penguji Hutan Rawak	63.3	69.36	62.56	64.15	69.43

RAJAH 3. Keputusan peratus ketepatan bagi pengelasan set data 1

Graf pada Rajah 3 turut menunjukkan lima jenis algoritma yang digunakan untuk pengelasan set data 1 ketika proses penambahan data baru. Berdasarkan keputusan pada graf tersebut, prestasi keputusan yang ditunjukkan oleh Bayes Naif Multinomial adalah lebih konsisten berbanding pengelas yang lain. Selain itu, Bayes Naif Multinomial juga menunjukkan hasil keputusan tertinggi iaitu dengan memperolehi peratus ketepatan sebanyak 72.21%.

Keputusan yang dihasilkan ini menunjukkan Naif Bayes Multinomial mempunyai ketepatan yang lebih baik daripada pengelas yang lain. Ini kerana Naif Bayes Multinomial menganggap bahawa dokumen adalah beg perkataan dan mengambil kira kekerapan dan maklumat perkataan. Multinomial sering menjadi kaedah yang digemari untuk sebarang jenis klasifikasi teks seperti pengesanan spam, pengkategorian topik dan analisis sentimen sebagai pertimbangan kekerapan perkataan, dan untuk mendapatkan ketepatan yang lebih baik daripada hanya memeriksa kejadian kata (Muhammad Abbas et al., 2019).



	Bayes Naif	Bayes Naif Multinomial	Regresi Logistik	Mesin Sokongan Vektor	Pemaksimuman Jangkaan
Penguji Bayes Naif	0.61	0.68	0.65	0.54	0.57
Penguji Bayes Naif Multinomial	0.68	0.72	0.69	0.66	0.7
Penguji Mesin Sokongan Vektor	0.59	0.7	0.67	0.67	0.68
Penguji Hutan Rawak	0.63	0.69	0.67	0.67	0.7

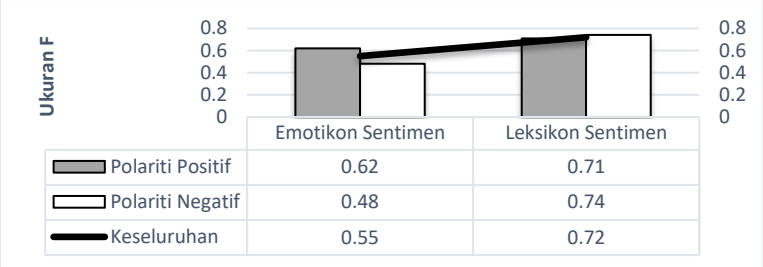
RAJAH 4. Keputusan ukuran F bagi pengelasan set data 1

Rajah 4 pula menunjukkan graf keputusan ukuran F bagi pengujian pengelasan yang dijalankan ke atas set data 1 dengan menggunakan empat jenis algoritma pengujian. Berdasarkan graf di Rajah 3 ini juga, prestasi keputusan ukuran F menunjukkan Bayes Naif Multinomial adalah lebih konsisten berbanding algoritma yang lain. Selain itu, Bayes Naif Multinomial juga telah menunjukkan bahawa keputusan yang diperolehi adalah yang tertinggi sekali iaitu dengan memperolehi peratus ketepatan sebanyak 0.72.

Ukuran F sering digunakan untuk mendapatkan keputusan ketepatan yang tidak cenderung kepada sesuatu kelas tertentu. Oleh itu keputusan Bayes Naif Multinomial merupakan keputusan ketepatan paling tinggi daripada yang lain yang tidak dipengaruhi oleh sesuatu kelas. Secara asasnya, pengelas Bayes Naif Multinomial ini adalah pengelas Bayes Naif tetapi ia menggunakan pengagihan multinomial untuk setiap fitur. Fitur-fitur tersebut diagihkan secara multinomial untuk mengira kebarangkalian dokumen bagi setiap label dan mengekalkan pemaksimuman kebarangkalian label (Lohar et al., 2017).

PENILAIAN PENGUJIAN ANTARA LEKSIKON SENTIMEN DAN EMOTIKON SENTIMEN

Penilaian antara leksikon sentimen dan emotikon sentimen adalah dengan menggunakan ukuran F. RAJAH 5 menunjukkan perbandingan keputusan ukuran F bagi emotikon sentimen dan leksikon sentimen. Nilai ukuran F bagi polariti positif yang tertinggi adalah leksikon sentimen yang menghasilkan nilai keputusan sebanyak 0.711 dan diikuti oleh emotikon sentimen yang menghasilkan nilai keputusan sebanyak 0.616. Oleh itu, hasil daripada keputusan ini menunjukkan ketepatan penggunaan leksikon sentimen untuk menghasilkan benih data adalah lebih tinggi berbanding penggunaan emotikon sentimen.



RAJAH 5. Perbandingan ukuran F bagi leksikon sentimen dan emotikon sentimen

PENGUJIAN PENGGABUNGAN BENIH DATA DAN DATA LATIHAN BARU

Penilaian pengujian ini adalah untuk membuat perbandingan antara keputusan pengelasan yang dibuat dengan menggunakan data latihan yang menggunakan benih data sahaja dan data latihan yang menggunakan gabungan benih data dan data latihan baru. JADUAL 12 menunjukkan keputusan yang diperolehi bagi kedua-dua pengujian.

JADUAL 12. Perbandingan antara benih data serta benih data dan data latihan baru

Bil.	Data Latihan	Bilangan Data	Peratus Ketepatan	Ukuran-F
1.	Benih Data Latihan	1102	72.34	0.72
2.	Gabungan Data Latihan	2354	91.91	0.92

Merujuk kepada perbandingan di Jadual 12, gabungan data latihan mempunyai saiz bilangan data yang lebih besar daripada saiz bilangan data bagi benih data latihan. Keputusan daripada kedua-dua data latihan yang ditunjukkan di Jadual 12 juga menunjukkan bahawa peratus ketepatan dan ukuran F bagi gabungan data latihan adalah lebih tinggi berbanding benih data latihan iaitu sebanyak 91.91% dan 92%. Berdasarkan keputusan ini, didapati ketepatan keputusan semakin meningkat apabila saiz bilangan data latihan yang digunakan adalah lebih besar. Oleh itu, bilangan data latihan memberikan pengaruh kepada keputusan ketepatan untuk pengelasan.

KESIMPULAN

Kajian yang dijalankan ini adalah untuk mencapai objektif utamanya iaitu membangunkan benih data menggunakan pendekatan leksikon sentimen bagi pembangunan korpus analisis sentimen Bahasa Melayu serta membangunkan korpus analisis sentimen Bahasa Melayu secara separa selia yang menggunakan gabungan kaedah pengelasan leksikon dan pembelajaran mesin. Oleh itu, tiga pengujian telah berjaya dijalankan untuk menentukan kaedah yang terbaik yang boleh digunakan untuk pembangunan korpus analisis sentimen dalam Bahasa Melayu ini. Berdasarkan penilaian pengujian yang telah dijalankan ini juga, beberapa sumbangan telah diberikan iaitu dengan membuktikan bahawa penggunaan leksikon sentimen adalah lebih tepat berbanding penggunaan emotikon leksikon. Selain itu, kaedah pembangunan benih data telah dihasilkan sebagai salah satu kaedah untuk membuat penganotasian tanpa menggunakan bantuan manusia. Kaedah penambahan data latihan untuk tujuan pengembangan data latihan supaya dapat menghasilkan korpus yang mempunyai data latihan yang lebih besar juga telah dihasilkan. Pengujian yang dijalankan telah menunjukkan Bayes Naif Multinomial sebagai model pengelas yang terbaik yang dapat menghasilkan ketepatan yang lebih tinggi yang mempunyai prestasi yang lebih baik sesuai untuk digunakan dalam pembangunan korpus

analisis sentimen. Akhir sekali, sebuah korpus Bahasa Melayu telah dapat dibina secara automatik dengan menggunakan gabungan kaedah leksikon sentimen dan pembelajaran mesin.

RUJUKAN

Adebisi, A., Olusayo, O., & Olatunde, O. 2012. An Exploratory Study of K-Means and Expectation Maximization Algorithms. *British Journal of Mathematics & Computer Science, 2*(2), 62-71.

Adel Assiri, Ahmed Emam, & Hmood al-Dossari. 2016. Saudi Twitter Corpus for Sentiment Analysis. *International Journal of Computer and Information Engineering*, 272-275.

Ankit, & Saleena, N. 2018. An Ensemble Classification System for Twitter Sentiment Analysis. *International Conference on Computational Intelligence and Data Science (ICCIDS 2018)*, hlm. 937-946.

Brum, H. B., & Volpe Nunes, M. d. 2018. Building a Sentiment Corpus of Tweets in Brazilian Portuguese. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* hlm. 4167-4172. Miyazaki: European Language Resources Association (ELRA).

Chumwatana, T. 2018. Comment Analysis for product and service satisfaction from Thai customers' review in social network. *Journal of Information and Communication Technology, 17*(2), 271-289.

Flender, M., & Gips, C. 2017. Sentiment Analysis of a German Twitter-Corpus. *Lernen, Wissen, Daten, Analysen (LWDA) Conference Proceedings (LWDA 2017)* (hlm. 25). Rostock: http://ceur-ws.org/Vol-1917/.

Gohil, S., Vuik, S., & Darzi, A. 2018. Sentiment Analysis of Health Care Tweets: Review of the Methods Used. (G. Eysenbach, Ed.) *JMIR Public Health Surveill, 4*(2). doi:10.2196/publichealth.5789

Lohar, P., Chowdhury, K., Afli, H., Hasanuzzaman, M., & Way, A. 2017. A Multinomial Naive Bayes Classification Approach for Customer Feedback Analysis. *Proceedings of the 8th International Joint Conference on Natural Language Processing*, (hlm. 161–169). Taipei.

Lourdusamy, R., & Abraham, S. 2018. A Survey on Text Pre-processing Techniques and Tools. *International Journal of Computer Sciences and Engineering, 6*(3), 148-157.

Modak, S., & Mondal, A. 2014. A Study on Sentiment Analysis. *International Journal of Advanced Research in Computer Science & Technology, 2*(2), 284-288. Didapatkan dari http://ijarcst.com/doc/vol2-issue2/ver.2/santanu_modak.pdf

Muhammad Abbas, Kamran Ali Memon, Abdul Aleem Jamali, Saleemullah Memon, & Anees Ahmed. 2019. Multinomial Naive Bayes Classification Model for Sentiment Analysis. *International Journal of Computer Science and Network Security, 19(3),* 62-67.

Nur Sharmini Alexander, & Nazlia Omar. 2017. Generating A Malay Sentiment Lexicon Based on Wordnet. *Asia-Pacific Journal of Information Technology and Multimedia,* 126 - 140.

Pak, A., & Paroubek, P. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (hlm. 1320-1326). Valletta, Malta: European Languages Resources Association (ELRA. Didapatkan dari http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

S. Jagdale, R., S. Shirsat, V., & N. Deshmukh, S. 2016. Sentiment Analysis of Events from Twitter Using Open Source Tool. *International Journal of Computer Science and Mobile Computing (IJCSMC)*, 475-485.

Sharma, R., Gupta, M., Agarwal, A., & Bhattacharyya, P. 2015. Adjective Intensity and Sentiment Analysis. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (hlm. 2520-2526). Lisbon: Association for Computational Linguistics.

Sharma, S., & P. Shetty, N. 2018. Determining the Popularity of Political Parties Using Twitter Sentiment Analysis. *Advances in Intelligent Systems and Computing. 701*, hlm. 21-29. Bhubaneswar: Springer Verlag. doi:https://doi.org/10.1007/978-981-10-7563-6_3

Wicaksono, A., Vania, C., Distiawan T., B., & Adriani, M. 2014. Automatically Building a Corpus for Sentiment Analysis on Indonesian Tweets. *Proceedings of the 28th Pacific Asia Conference on Language, Information, and Computation* (hlm. 185-194). Phuket: Department of Linguistics, Chulalongkorn University. Didapatkan dari https://www.aclweb.org/anthology/Y14-1024

Ezuana Sukawai
Nazlia Omar
Fakulti Teknologi & Sains Maklumat
Universiti Kebangsaan Malaysia
ezuana@gmail.com, nazlia@ukm.edu.my