PENGELASAN SENTIMEN ULASAN TWIT PRODUK BERSAMA EMOJI MENGGUNAKAN TEKNIK PEMBELAJARAN MESIN TERSELIA

NG YU TAO
SHAHRUL AZMAN MOHD NOAH

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Analisis sentimen menggunakan teknik pembelajaran mesin dengan pemprosesan bahasa tabii untuk membaca dan mengklasifikasikan teks sebagai positif, negatif, atau neutral secara automatik. Ia dapat menganalisis emosi penulis dan memberikannya kepada nilai tertentu secara automatik mengikut polariti. Pengguna menggunakan emoji dalam ulasan untuk meluahkan perasaan mereka. Syarikat menggunakan teknik analisis sentimen untuk mendapatkan pendapat dari ulasan pelanggan. Lazimnya, langkah pra-pemprosesan akan menyaringkan emoji. Walau bagaimanapun, ulasan mengenai nama produk sahaja dengan emoji akan menghasilkan sentimen yang lebih jelas. Penyelidikan ini memilih *Twitter* sebagai sumber maklumat dalam analisis sentimen kerana boleh memberikan akses langsung kepada pendapat umum berdasarkan teks. Algoritma mengumpulkan twit produk mudah alih *Apple* iaitu iPhone, iPad, dan Macbook dalam bahasa Inggeris. Penyelidikan ini akan membangunkan set data yang mewakili emoji dan tanpa emoji. Set data yang sama harus digunakan. Satu menjalani proses menukar emoji menjadi emosi manusia dan lain menyaring emoji mengikut Python Unicode dalam langkah pra-pemprosesan. Teks yang telah menghilangkan kehingaran akan diproses dengan Studio Label untuk mendapatkan sentimen label. Penyelidikan ini menggunakan klasifikasi dalam teknik pembelajaran mesin terselia dan menggunakan Naive Bayes and Mesin Vektor Sokongan (SVM) sebagai pengelas. Hasil analisis mengesahkan bahawa emoji akan mempengaruhi sentimen. Penggunaan emoji dalam analisis sentimen menghasilkan skor sentimen yang lebih tinggi dan mempengaruhi sentimen keseluruhan. Twit dengan emoji yang menjalankan proses penukaran emoji ke perasaan manusia mempunyai ketepatan keseluruhan yang lebih tinggi iaitu 83.55% untuk iPhone, 84.55% untuk iPad, dan 83.99% untuk Macbook.

1 PENGENALAN

Analisis sentimen adalah teknik untuk mengelaskan sesuatu pandangan sama ada positif, negatif, neutral. Ini adalah cara untuk menilai tulisan untuk menentukan apakah ungkapan itu disukai (positif), tidak disukai (negatif), atau neutral. (Clarabridge, 2019). Pendapat dan pilihan yang dinyatakan dalam rangkaian sosial dan perkhidmatan blog mikro sangat penting dalam analisis sentimen dan pertimbangan pendapat. Analisis sentimen media sosial bertujuan

melihat ketidakpuasan pelanggan atau masalah produk, menganggarkan harga pasaran saham. Twitter telah mendapat populariti tertinggi berbanding dengan semua platform blog mikro lain dalam beberapa tahun kebelakangan. Terdapat 6000 jumlah twit yang disiarkan setiap saat oleh pengguna Twitter, dan 500 juta twit dihantar setiap hari (Kit Smith, 2020).

Penggunaan emoji di Internet meningkat dengan pesat kerana membolehkan pengguna mengekspresikan emosi mereka dengan lebih mudah. Laporan oleh Novak etl al. (2015) menunjukkan bahawa 92% populasi dalam talian menggunakan emoji (Emoji Research Team of Emogi, 2016). Emoji 😂 adalah emoji paling popular di Twitter dan telah digunakan lebih dari dua bilion kali (Emojipedia, 2020).

Analisis sentimen dapat dilakukan dengan menggunakan teknik pembelajaran mesin terselia. Ia menggunakan pengelas untuk mengembangkan model ramalan. Antara algoritma pengelas dalam analisis sentimen ialah Naive Bayes dan Mesin Vektor Sokongan (SVM). Naive Bayes digambarkan sebagai sekumpulan algoritma probabilistik yang menggunakan Teorem Bayes untuk meramalkan kategori teks. Mesin Vektor Sokongan adalah model bukan probabilistik yang menggunakan perwakilan contoh teks sebagai titik dalam ruang multidimensi. Matriks konfusi ialah kaedah untuk memvisualisasikan prestasi pengelas dan menyampaikan maklumat yang lebih jelas daripada skor ketepatan. Penyelidikan ini meneliti kemunculan emoji pada ulasan produk mudah alih Apple di Twitter dan bagaimana ia mempengaruhi label analisis sentimen.

2 PENYATAAN MASALAH

Kebanyakan pengguna memberikan ulasan produk melalui media sosial seperti *Twitter* untuk berkongsi pengalaman mereka menggunakan produk tersebut. Mereka menggunakan emoji untuk mengemukakan pendapat dengan lebih mudah. Jumlah komen di rangkaian sosial dan laman tinjauan secara dalam talian sangat besar. Dalam data yang tidak kemas ini, muncul kesempatan untuk memahami pendapat subjektif mengenai teks. Dengan analisis sentimen, ia dapat menganalisis ulasan produk secara automatik dan mengelaskan kepada positif, negatif, atau neutral sesuai dengan polariti dokumen yang diberikan. Walaupun penggunaan emoji dalam ulasan dapat lebih menjelaskan emosi dan perasaan dari pengguna, ia masih mengelirukan kerana is bersifat subjektif bergantung kepada persepsi. Oleh itu, penyelidikan

ini bertujuan mengkaji sama ada penggunaan emoji dapat mempengaruhi sentimen positif, negatif, dan neutral korpus dalam analisis sentimen. Nama produk tidak mempunyai nilai sentimen apabila disiarkan sendiri. Namun, apabila emoji wajah tersenyum digunakan bersama dengan nama produk, teks mungkin memiliki nilai sentimen yang lebih jelas. Perwatakan emoji seperti wajah yang tersenyum dapat menunjukkan perasaan positif terhadap produk. Sebaliknya, emoji wajah yang marah bersama dengan beberapa nama produk dapat menunjukkan perasaan negatif terhadap produk tersebut. Perwatakan emoji yang sederhana dapat memberikan makna yang lebih mendalam kepada pendapat yang diberikan.

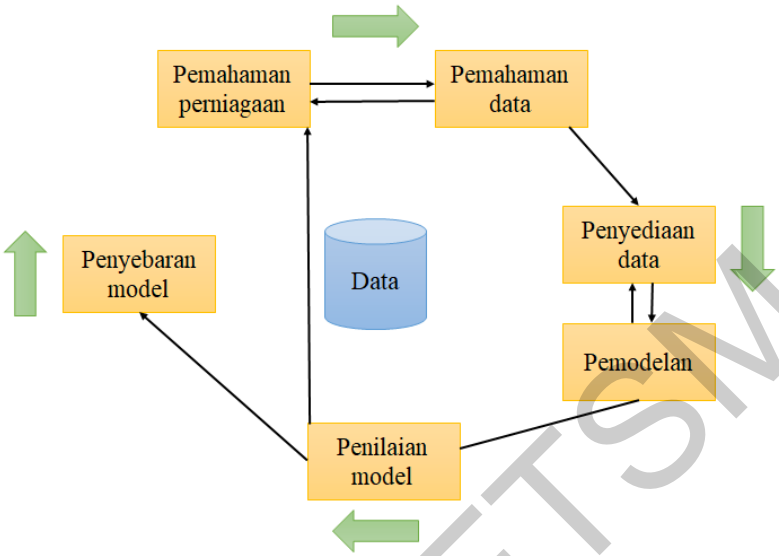
3 OBJEKTIF KAJIAN

Matlamat penyelidikan ini adalah untuk mengkaji sejauh mana emoji mempengaruhi nilai sentimen sesuatu ulasan twit berdasarkan teknik pengelas terselia.

Kertas ini menggunakan set data yang mewakili ulasan dengan emoji dan ulasan tanpa emoji dengan menggunakan Naive Bayes dan Mesin Vektor Sokongan (SVM) sebagai pengelas untuk meramalkan nilai ketepatan.

4 METOD KAJIAN

Penggunaan model pembangunan yang sesuai penting untuk memasti perjalanan penyeldikan berjalan dengan lancar dan menjamin hasil kerja yang berkualiti. Rajah 4.1 menunjukkan proses pemodelan perlombongan data. Fasa pembangunan termasuk pemahaman perniagaan, pemahaman data, penyediaan data, pemodelan, penilaian model, dan penyebaran model. Proses perlombongan data adalah penemuan melalui data kumpulan besar, hubungan, dan pandangan yang membantu perusahaan dalam mengukur dan mengurus di mana mereka berada dan meramalkan di mana mereka akan berada di masa depan (Rob Peterson, 2018).



Rajah 4.1 Proses Pemodelan Perlombongan Data

4.1 Fasa Pemahaman Perniagaan

Fasa ini melibatkan pemahaman objektif perniagaan dengan jelas dan mengetahui keperluan perniagaan. Seterusnya, menilai keadaan semasa dengan mencari sumber, andaian, kekangan, dan faktor penting lain yang harus dipertimbangkan. Dari objektif perniagaan dan situasi semasa, menghasilkan tujuan perlombongan data untuk mencapai objektif perniagaan dalam keadaan semasa. Perancangan perlombongan data yang baik dan terperinci dapat mencapai tujuan perniagaan dan perlombongan data.

4.2 Fasa Pemahaman Data

Fasa ini bermula dengan pengumpulan data awal, pengumpulan data dari sumber data yang tersedia, untuk membantu membiasakan diri dengan data tersebut. Kemudian, data perlu dieksplorasi dengan menangani pertanyaan-pertanyaan perlombongan data iaitu menggunakan pertanyaan, pelaporan, atau visualisasi. Contohnya, penyelidikan ini menganalisis sama ada sentimen label dari ulasan pengguna *Twitter* akan dipengaruhi oleh penggunaan emoji.

4.3 Fasa Penyediaan Data

Fasa ini biasanya akan menggunakan 90% masa projek. Hasil dari fasa ini adalah set data terakhir. Setelah sumber data yang tersedia dikenal pasti, sumber tersebut harus dipilih, dihilangkan kehingaran, dibina, dan diformat ke dalam bentuk yang diinginkan. Penyelidikan ini mengekstrak data dalam bahasa Inggeris dari Twitter dengan menggunakan Twitter API.

Dengan menggunakan data yang sama, data perlu menjalankan penukaran emoji ke perasaan pengguna berdasarkan penjelasan dari Emojipedia dan EmojiAll sebelum pra-pemprosesan dan dalam fail Python yang menjalankan analisis sentimen tanpa emoji, algoritma yang menyaringkan emoji diperlukan. Data perlu menjalankan langkah pra-pemprosesan untuk menapis kehingaran. Lemmatisasi adalah untuk mengurangkan bentuk infleksi setiap kata menjadi dasar atau akar yang sama. Seterusnya, data perlu dilabelkan dengan menggunakan alat pelabelan data yang memiliki sumber terbuka iaitu Label Studio. Jadual 4.1 menunjukkan contoh emoji yang mempunyai sentimen.

Jadual 4.1 Contoh Emoji yang Mempunyai Sentimen

Sentimen	Emoji	Emosi dan Perasaan
Positif		Gembira Cinta Setuju
Neutral		Tiada ekspresi
Negatif		Marah
		Sedih
		Kecewa
		Cemas dan Takut

4.4 Fasa Pemodelan

Fasa ini adalah mengenal pasti tugas perlombongan data seperti menggunakan klasifikasi, ramalan, dan pengelompokan. Teknik pemodelan harus dipilih untuk digunakan dalam set data yang disediakan. Senario ujian mesti dihasilkan untuk mengesahkan kualiti model. Model perlu dinilai dengan teliti dengan melibatkan pihak berkepentingan untuk memastikan bahawa model memenuhi inisiatif perniagaan. Penyelidikan ini menggunakan klasifikasi sebagai tugas

perlombongan data. Selepas mengenal pasti tugas perlombongan data, penyelidkan ini menggunakan teknik pembelajaran mesin terselia sebagai sebagai teknik untuk melombong data. Pengelas Naive Bayes dan Mesin Vektor Sokongan (SVM) akan digunakan untuk ramalan.

4.5 Fasa Penilaian Model

Fasa ini bertujuan untuk menghasilkan keputusan penilaian tinggi supaya boleh terus menjalankan ke fasa terakhir. Penilaian yang kurang tepat dicadangkan untuk mengulangkan fasa sebelumnya iaitu fasa penyediaan data dengan menambahbaik langkah pra-pemprosesan yang menghilangkan kehingaran. Penilaian model menggunakan *cross validation* berstrata untuk membahagikan data. Sisihan piwaian yang lebih rendah bermaksud model berprestasi dengan baik. Matriks konfusi digunakan untuk menunjukkan prestasi model dari data ujian. aporan klasifikasi dapat menghasilkan nilai dapatan (*precision*), kejituan (*recall*), dan ukuran-f1 (*f1-score*) yang memberikan maklumat yang lebih terperinci mengenai prestasi model.

4.6 Fasa Peyebaran Model

Fasa ini merupakan fasa untuk membuat laporan seperti proses perlombongan data berulang, rancangan penyebaran, pemeliharaan, dan pemantauan harus dibuat untuk pelaksanaan dan sokongan masa depan. Laporan akhir penyelidikan perlu meringkaskan pengalaman projek dan mengkaji semula projek untuk mengenal pastikan limitasi yang perlu diperbaiki.

5 HASIL KAJIAN

a) Kata kunci: iphone

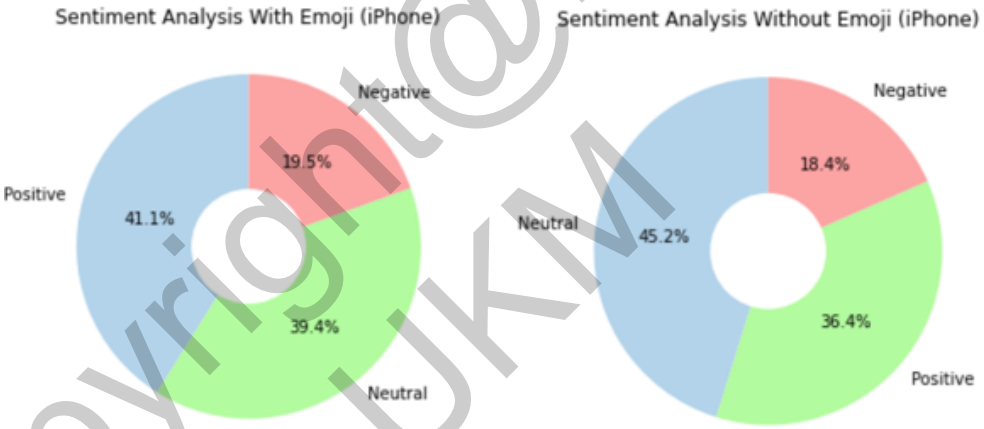
   Data mempunyai jumlah 30017 twit tentang iphone.

   Jadual 5.1 menunjukkan jumlah kelas twit iphone yang mengandungi dan tidak mengandungi emoji.

Jadual 5.1 Jumlah Kelas Twit Iphone yang Mengandungi dan Tidak Mengandungi Emoji

Label	Mengandungi Emoji	Tiada Emoji
Positif : 1	12332	10925
Neutral : 0	11837	13567
Negatif : -1	5848	5525

Rajah 5.1 menunjukkan perbandingan carta pie twit iphone yang mengandungi dan tidak mengandungi emoji.



Rajah 5.1 Perbandingan Carta Pie Twit Iphone yang Mengandungi dan Tidak Mengandungi Emoji

Dari Jadual 5.1 dan Rajah 5.1, dengan menggunakan set data yang sama, jumlah twit positif dan negatif yang mengandungi emoji ke perasaan pengguna menjadi lebih berbanding dengan twit yang tiada emoji. Jumlah twit neutral yang mengandungi emoji menjadi menjadi kurang berbanding dengan twit yang tiada emoji.

Jadual 5.2 Penilaian Ketepatan Pengelas Model (iPhone)

	Model / Ciri	Naive Bayes	Mesin Vektor Sokongan (SVM)
Emoji	BoW (%)	61.31	63.59
	TF-IDF (%)	77.61	**83.55**
Tiada Emoji	BoW (%)	61.20	64.09
	TF-IDF (%)	71.77	**75.03**

Jadual 5.2 menunjukkan penilaian ketepatan pengelas model (iPhone). Berdasarkan Jadual 5.2, bagi set data iphone, twit yang mengandungi emoji ke perasaan pengguna mempunyai ketepatan yang lebih tinggi dengan menggunakan TF-IDF dan SVM iaitu 83.55% berbanding dengan twit yang tidak mempunyai emoji iaitu 75.03%.

Rajah 5.2 menunjukkan matriks konfusi twit iphone yang mengandungi emoji. Rajah 5.3 menunjukkan matriks konfusi twit iphone yang tidak mengandungi emoji.

```
Accuracy:  83.5459787887715 %

Confusion matrix:
 [[ 323  171   91]
 [   6 1144   33]
 [  45  168 1020]]


Sentiment Label
 -1: Negative, 0: Neutral, 1: Positive

              precision    recall  f1-score   support

          -1       0.86      0.55      0.67       585
           0       0.77      0.97      0.86      1183
           1       0.89      0.83      0.86      1233

    accuracy                           0.83      3001
   macro avg       0.84      0.78      0.80      3001
weighted avg       0.84      0.83      0.82      3001
```

Rajah 5.2 Matriks Konfusi Twit Iphone yang Mengandungi Emoji

```
Accuracy:  75.02744921135549 %

Confusion matrix:
 [[ 229  241   82]
 [  11 1319   26]
 [  51  338  704]]


Sentiment Label
 -1: Negative, 0: Neutral, 1: Positive

              precision    recall  f1-score   support

          -1       0.79      0.41      0.54       552
           0       0.69      0.97      0.81      1356
           1       0.87      0.64      0.74      1093

    accuracy                           0.75      3001
   macro avg       0.78      0.68      0.70      3001
weighted avg       0.77      0.75      0.74      3001
```

Rajah 5.3 Matriks Konfusi Twit Iphone yang Mengandungi Emoji

Rajah 5.2 dan Rajah 5.3 menggunakan jumlah 3001 twit sebagai set data latihan. Dari Rajah 5.2, model yang mempunyai nilai ketepatan tertinggi ialah SVM dengan pengekstrak ciri TF-IDF iaitu 83.55%. Jumlah twit yang betul negatif ialah 323, betul neutral ialah 1144, dan betul positif ialah 1020. Nilai dapatan bermaksud berapa ramalan label yang diramalkan dengan betul daripada jumlah ramalan untuk kelas tersebut. Nilai dapatan 86% bagi label negatif (-1), 77% bagi label neutral (0), dan 89% bagi label positif (1). Dari matriks di Rajah 5.2, model berprestasi baik kerana dapat meramalkan label sentimen yang tepat untuk kelas 0 dan 1 iaitu neutral dan positif dengan lebih baik sebab mempunyai ukuran-f1 yang paling tinggi iaitu 0.86.

Dari Rajah 5.3, model yang mempunyai nilai ketepatan tertinggi ialah SVM dengan pengekstrak ciri TF-IDF iaitu 75.03%. Jumlah twit yang betul negatif ialah 229, betul neutral ialah 1319, dan betul positif ialah 704. Nilai dapatan 79% bagi label negatif (-1), 69% bagi label neutral (0), dan 87% bagi label positif (1). Dari matriks di Rajah 5.3, pengelas SVM dapat meramalkan label sentimen yang tepat untuk kelas 0 iaitu neutral dengan lebih baik sebab mempunyai ukuran-f1 yang paling tinggi iaitu 0.81.

b) Kata kunci: ipad

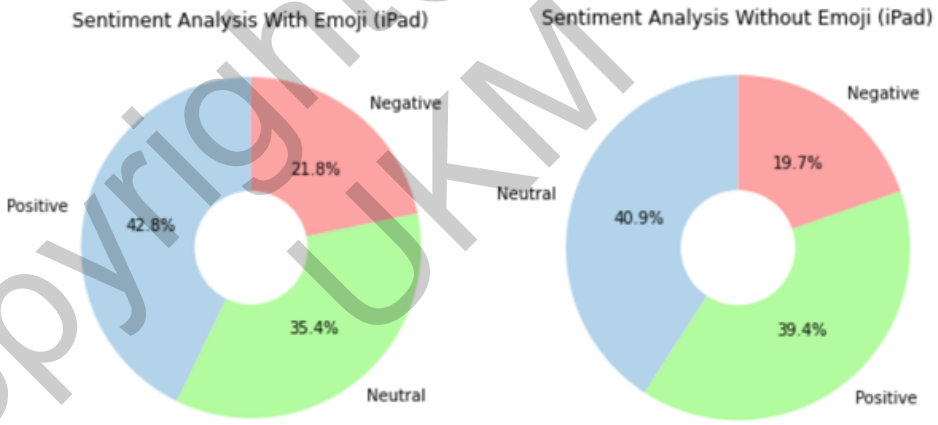
Data mempunyai jumlah 25092 twit tentang ipad.

Jadual 5.3 menunjukkan jumlah kelas twit ipad yang mengandungi dan tidak mengandungi emoji.

Jadual 5.3 Jumlah Kelas Twit Ipad yang Mengandungi dan Tidak Mengandungi Emoji

Label	Mengandungi Emoji	Tiada Emoji
Positif : 1	10744	9898
Neutral : 0	8890	10258
Negatif : -1	5458	4936

Rajah 5.4 menunjukkan perbandingan carta pie twit ipad yang mengandungi dan tidak mengandungi emoji.



Rajah 5.4 Perbandingan Carta Pie Twit Ipad yang Mengandungi dan Tidak Mengandungi Emoji

Dari Jadual 5.3 dan Rajah 5.4, dengan menggunakan set data yang sama, jumlah twit positif dan negatif yang mengandungi emoji ke perasaan pengguna menjadi lebih berbanding dengan twit yang tiada emoji. Jumlah twit neutral yang mengandungi emoji menjadi menjadi kurang berbanding dengan twit yang tiada emoji.

Jadual 5.4 Penilaian Ketepatan Pengelas Model (iPad)

	Model / Ciri	Naive Bayes	Mesin Vektor Sokongan (SVM)
Emoji	BoW (%)	61.65	62.94
	TF-IDF (%)	76.92	**84.54**
Tiada Emoji	BoW (%)	60.80	61.73
	TF-IDF (%)	73.00	**77.18**

Jadual 5.4 menunjukkan penilaian ketepatan pengelas model (iPad). Berdasarkan Jadual 5.4, bagi set data ipad, twit yang mengandungi emoji menukar ke perasaan pengguna mempunyai ketepatan yang lebih tinggi dengan meggunakan TF-IDF dan SVM iaitu 84.54% berbanding dengan twit yang tiada emoji iaitu 77.18%.

Rajah 5.5 menunjukkan matriks konfusi twit ipad yang mengandungi emoji. Rajah 5.6 menunjukkan matriks konfusi twit ipad yang tidak mengandungi emoji.

```
Accuracy:  84.54485128438023 %

Confusion matrix:
 [[340 137  69]
 [ 13 860  16]
 [ 41 134 899]]


Sentiment Label
 -1: Negative, 0: Neutral, 1: Positive

              precision    recall  f1-score   support

          -1       0.86      0.62      0.72       546
           0       0.76      0.97      0.85       889
           1       0.91      0.84      0.87      1074

    accuracy                           0.84      2509
   macro avg       0.85      0.81      0.82      2509
weighted avg       0.85      0.84      0.83      2509
```

Rajah 5.5 Matriks Konfusi Twit Ipad yang Mengandungi Emoji

```
Accuracy:  77.17601653966041 %

Confusion matrix:
 [[225 192  77]
 [ 12 997  16]
 [ 51 272 667]]


Sentiment Label
 -1: Negative, 0: Neutral, 1: Positive

              precision    recall  f1-score   support

          -1       0.78      0.46      0.58       494
           0       0.68      0.97      0.80      1025
           1       0.88      0.67      0.76       990

    accuracy                           0.75      2509
   macro avg       0.78      0.70      0.71      2509
weighted avg       0.78      0.75      0.74      2509
```

Rajah 5.6 Matriks Konfusi Twit Ipad yang Tidak Mengandungi Emoji

Rajah 5.5 dan Rajah 5.6 menggunakan jumlah 2509 twit sebagai set data latihan. Dari Rajah 5.5, model yang mempunyai nilai ketepatan tertinggi ialah SVM dengan pengekstrak ciri TF-IDF iaitu 84.55%. Jumlah twit yang betul negatif ialah 340, betul neutral ialah 860, dan betul positif ialah 899. Nilai dapatan 86% bagi label negatif (-1), 76% bagi label neutral (0), dan 91% bagi label positif (1). Model berprestasi baik meramalkan label sentimen yang tepat untuk kelas 1 iaitu positif dengan lebih baik sebab mempunyai ukuran-f1 yang paling tinggi iaitu 0.87 diikuti oleh kelas neutral (0) iaitu 0.85.

Dari Rajah 5.6, model yang mempunyai nilai ketepatan tertinggi ialah SVM dengan pengekstrak ciri TF-IDF iaitu 77.18%. Jumlah twit yang betul negatif ialah 225, betul neutral ialah 997, dan betul positif ialah 667. Nilai dapatan 78% bagi label negatif (-1), 68% bagi label neutral (0), dan 88% bagi label positif (1). Model mempunyai ukuran-f1 yang paling tinggi ialah 0.80 bagi 0 iaitu neutral.

c) Kata kunci: macbook

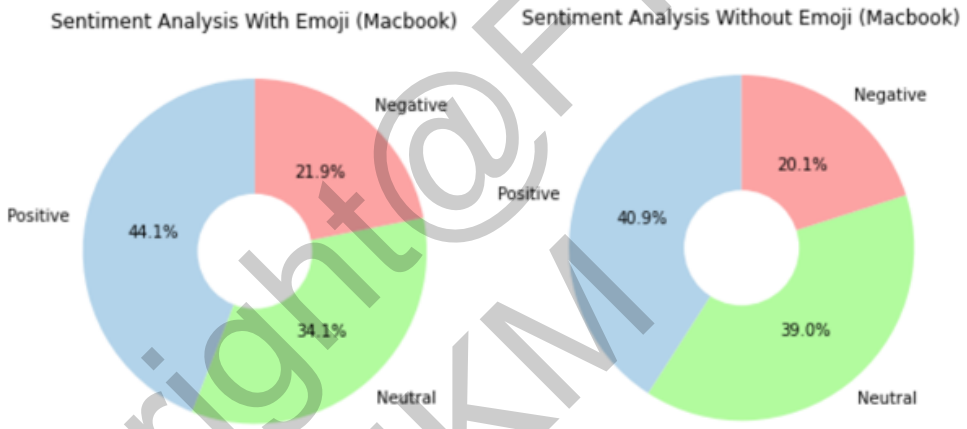
Data mempunyai jumlah 16546 twit tentang macbook.

Jadual 5.5 menunjukkan jumlah kelas twit macbook yang mengandungi dan tidak mengandungi emoji.

Jadual 5.5 Jumlah Kelas Twit Macbook yang Mengandungi dan Tidak Mengandungi Emoji

Label	Mengandungi Emoji	Tiada Emoji
Positif : 1	7289	6770
Neutral : 0	5636	6456
Negatif : -1	3621	3320

Rajah 5.7 menunjukkan perbandingan carta pie twit macbook yang mengandungi dan tidak mengandungi emoji.



Rajah 5.7 Matriks Konfusi Twit Ipad yang Tidak Mengandungi Emoji

Dari Jadual 5.5 dan Rajah 5.7, dengan menggunakan set data yang sama, jumlah twit positif dan negatif yang mengandungi emoji ke perasaan pengguna menjadi lebih berbanding dengan twit yang tiada emoji. Jumlah twit neutral yang mengandungi emoji menjadi menjadi kurang berbanding dengan twit yang tiada emoji.

Jadual 5.6 Penilaian Ketepatan Pengelas Model (Macbook)

	Model / Ciri	Naive Bayes	Mesin Vektor Sokongan (SVM)
Emoji	BoW (%)	61.78	66.71
	TF-IDF (%)	76.03	**83.99**
Tiada Emoji	BoW (%)	61.13	65.20
	TF-IDF (%)	71.85	**76.36**

Jadual 5.6 menunjukkan penilaian ketepatan pengelas model (Macbook). Berdasarkan Jadual 5.6, bagi set data macbook, twit yang mempunyai emoji menukar ke perasaan pengguna mempunyai ketepatan yang lebih tinggi dengan meggunakan TF-IDF dan SVM iaitu 83.99% berbanding dengan twit yang tiada emoji iaitu 76.36%.

Rajah 5.8 menunjukkan matriks konfusi twit macbook yang mengandungi emoji. Rajah 5.9 menunjukkan matriks konfusi twit macbook yang tidak mengandungi emoji.

```
Accuracy:  83.98998673909628 %

Confusion matrix:
 [[221  97  44]
 [  4 550   9]
 [ 25  98 606]]


Sentiment Label
 -1: Negative, 0: Neutral, 1: Positive

              precision    recall  f1-score   support

          -1       0.88      0.61      0.72       362
           0       0.74      0.98      0.84       563
           1       0.92      0.83      0.87       729

    accuracy                           0.83      1654
   macro avg       0.85      0.81      0.81      1654
weighted avg       0.85      0.83      0.83      1654
```

Rajah 5.8 Matriks Konfusi Twit Macbook yang Mengandungi Emoji

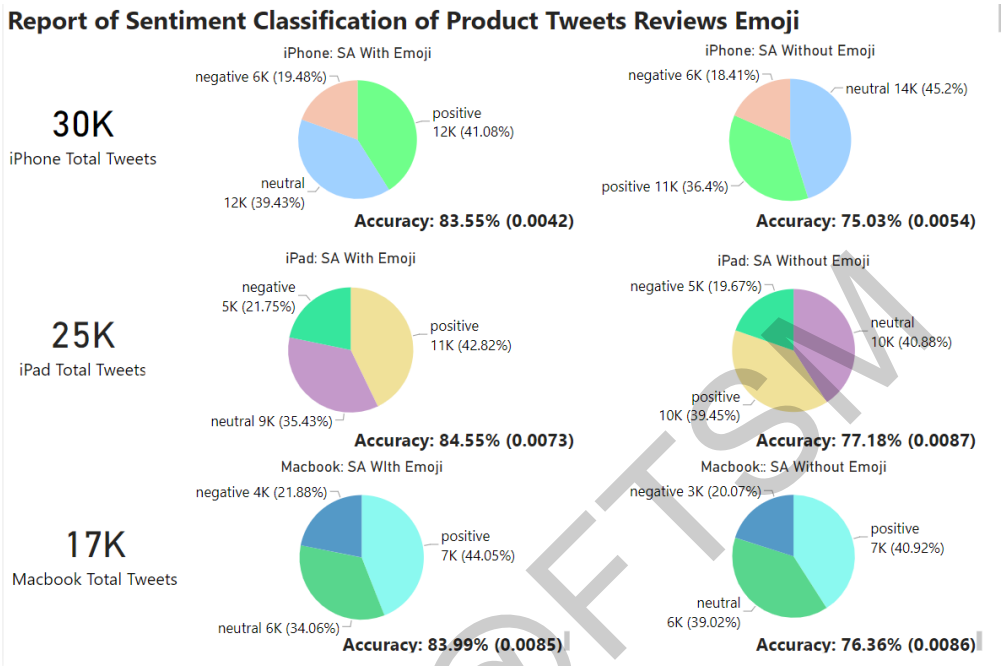
```
Accuracy:  76.3568717418544 %

Confusion matrix:
 [[157 132  43]
 [  3 635   7]
 [ 28 155 494]]


Sentiment Label
 -1: Negative, 0: Neutral, 1: Positive

              precision    recall  f1-score   support

          -1       0.84      0.47      0.60       332
           0       0.69      0.98      0.81       645
           1       0.91      0.73      0.81       677

    accuracy                           0.78      1654
   macro avg       0.81      0.73      0.74      1654
weighted avg       0.81      0.78      0.77      1654
```

Rajah 5.9 Matriks Konfusi Twit Macbook yang Tidak Mengandungi Emoji

Rajah 5.8 dan Rajah 5.9 menggunakan jumlah 2509 twit sebagai set data latihan. Dari Rajah 5.8, model yang mempunyai nilai ketepatan tertinggi ialah SVM dengan pengekstrak ciri TF-IDF iaitu 83.99%. Jumlah twit yang betul negatif ialah 221, betul neutral ialah 550, dan betul positif ialah 606. Nilai dapatan 88% bagi label negatif (-1), 74% bagi label neutral (0), dan 92% bagi label positif (1). Model berprestasi baik meramalkan label sentimen yang tepat untuk kelas 1 iaitu positif dengan lebih baik sebab mempunyai ukuran-f1 yang paling tinggi iaitu 0.87 diikuti oleh kelas neutral (0) iaitu 0.84.

Dari Rajah 5.9, model yang mempunyai nilai ketepatan tertinggi ialah SVM dengan pengekstrak ciri TF-IDF iaitu 76.36%. Jumlah twit yang betul negatif ialah 157, betul neutral ialah 635, dan betul positif ialah 494. Nilai dapatan 47% bagi label negatif (-1), 98% bagi label neutral (0), dan 73% bagi label positif (1). Model mempunyai ukuran-f1 yang paling tinggi ialah 0.98 bagi 0 iaitu neutral.

Untuk menunjukkan visualisasi yang jelas dan kemas untuk semua kata kunci, mengumpulkan semua data dan model dengan nilai ketepatan tertinggi dan menggambarkan dengan menggunakan *Microsoft Power BI*. Rajah 5.10 menunjukkan perbandingan visualisasi nilai ketepatan.



Rajah 5.10 Perbandingan Visualisasi Nilai Ketepatan

6 KESIMPULAN

Penyelidikan ini menggunakan set data yang sama dari *Twitter* untuk menjalankan analisis sentimen bertujuan untuk mengenalpastikan sama ada emoji akan mempengaruhi sentimen label. Analisis sentimen yang mengandungi emoji ke perasaan manusia mempunyai nilai ketepatan yang lebih tinggi berbanding dengan yang tiada emoji.

Sistem algoritma dalam penyeldikan ini mempunyai beberapa limitasi iaitu tidak dapat menyaringkan twit yang mempunyai bahasa negara lain. Sesetengah twit tentang produk *Apple* mengandungi bahasa Inggeris bercampur dengan bahasa negara lain seperti Spanish. Jumlah twit yang diektrak tidak mempunyai label positif, negatif, dan neutral. Proses pelabelan melalui Label Studio secara atas talian menggunakan masa yang panjang dan memerlukan penolongan dari pentadbir atas talian.

Peningkatan masa depan adalah menambahkan algoritma yang boleh menyaringkan bahasa negara lain dalam langkah pra-pemprosesan. Cadangan seterusnya ialah menggunakan pengekstrak ciri yang berbeza seperti *Word2Vec* atau *Glove* dan pengelas lain seperti *Decision Tree* untuk mengenalpastikan nilai ketepatan. Cadangan lain ialah menggunakan teknik pembelajaran mesin separa terselia dan tanpa terselia untuk meramalkan twit yang tidak mempunyai label dan menggunakan *textblob* atau *Vader* untuk memberikan label.

7 RUJUKAN

Rob Petersen. 2018. 6 Essential Steps to the Data Mining Process. https://barnraisersllc.com/2018/10/01/data-mining-process-essential-steps/

Barbieri, F., Ronzano, F., & Saggion, H. 2016. What does this emoji mean? A vector space skip-gram model for twitter emojis. *Proceedings of Language Resources and Evaluation Conference, LREC, Slovenia: Portoroz* 5(2): 3968-3969. https://www.researchgate.net/publication/303520355_What_does_this_Emoji_Mean_A_Vector_Space_Skip-Gram_Model_for_Twitter_Emojis

EmojiAll. 2021. https://www.emojiall.com/en/all-emojis

Emojipedia. 2021. https://emojipedia.org/twitter/

Gaël Guibon, Magalie Ochs, Patrice Bellot. 2016. From Emojis to Sentiment Analysis. *HAL, WACAI 2016, Lab-STICC; ENIB; LITIS* 12(10):152-198. https://hal-amu.archives-ouvertes.fr/hal-01529708/document

GOOGLE COLAB. 2021. Python Version 3.6.9. Google Inc.

Kit Smith. 2020. 60 Incredible and Interesting Twitter Stats and Statistics. https://www.brandwatch.com/blog/twitter-stats-and-statistics/

Kralj Novak, P., Smailović, J., Sluban, B., & Mozetič, I. 2015. Sentiment of Emojis. *International Journal of Computer and Electrical Engineering* 10 (12): 360-369. https://www.researchgate.net/publication/320446679_The_Effects_of_Emoji_in_Sentiment_Analysis

LABEL STUDIO. 2021. Version 1.0.0. Stok Kangri: Heartex.

MICROSOFT POWER BI. 2021. Version 2.93.981.0. Microsoft.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, Igor Mozetič. 2015. Sentiment of Emojis. *University of Maribor* 10: 1371-1390. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0144296

Serkan Ayvaz, Mohammed O.Shiha. 2017. The Effects of Emoji in Sentiment Analysis. *International Journal of Computer and Electrical Engineering* 1770: 360-369. https://www.researchgate.net/publication/320446679_The_Effects_of_Emoji_in_Sentiment_Analysis

Wieslaw Wolny. 2016. Twitter Sentiment Analysis Using Emoticons and Emoji Ideograms. *Information Systems Development: Complexity in Information Systems Development* 11 (9): 108-120. https://www.researchgate.net/publication/308413240_TWITTER_SENTIMENT_ANALYSIS_USING_EMOTICONS_AND_EMOJI_IDEOGRAMS