

PENGLASIFIKASIAN EMOSI BERDASARKAN LIRIK

NATASYA IZZATY KHOO

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Lagu-lagu patriotik telah diiktiraf sebagai cara untuk mempromosikan rasa perpaduan dan kebanggaan negara di kalangan warganya. Lirik yang digunakan dalam lagu-lagu patriotik bertujuan untuk membangkitkan emosi tertentu yang membolehkan orang ramai merasakan semangat kekitaan. Matlamat projek ini adalah untuk mengkaji kepentingan perkataan dan lirik yang terkandung dalam lagu-lagu patriotik yang dinyanyikan dan dimainkan di Malaysia dari tahun 1988 hingga 2021 dengan menggunakan teknik pembelajaran mesin seperti Mesin Vektor Sokongan (SVM), Naive Bayes sebagai model garis dasar, Regresi Logistik dan Pengelas Hutan Rawak serta mengukur pengekstrakan ciri semantik. Projek ini berminat untuk memahami bagaimana persepsi orang ramai telah dibentuk dan diubah secara ketat melalui cara hiburan, dengan tujuan untuk menyatukan dan mendidik orang ramai ke arah visi dan misi negara. Kajian awal yang dijalankan terhadap lirik lagu menunjukkan analisis emosi lirik telah dijalankan menggunakan dataset Moody Lyrics, tetapi set data yang terdiri daripada lirik yang ditulis dalam bahasa Melayu masih belum dijalankan sehingga kini. Oleh itu, set data yang dijana dalam projek ini juga boleh digunakan sebagai garis asas untuk kajian akan datang. Kesimpulannya, hasil kajian ini boleh dijadikan rujukan kepada pencipta lagu untuk menghasilkan lagu-lagu patriotik yang lebih baik dan lebih baik yang dapat memupuk perpaduan dalam kalangan rakyat Malaysia yang terdiri daripada pelbagai etnik.

1 PENGENALAN

Lagu-lagu patriotik telah memainkan peranan penting dalam memotivasikan patriotisme dan nasionalisme di banyak negeri dan sepanjang sejarah. Terutamanya, muzik dan lagu mempunyai kesan yang mendalam terhadap tamadun. Muzik adalah medium utama untuk menyatakan nilai budaya dan tradisional di banyak negara. Muzik rakyat dan desa sangat tertanam dalam budaya dan tradisi Amerika Syarikat. Sebagai contoh, muzik Amerika menggabungkan aspek tradisi rakyat Anglo-Amerika, Afro -Amerika, asli-Amerika, Sepanyol dan Perancis, serta muzik rakyat Amerika kontemporari. Muzik rakyat Afrika Amerika mempunyai kesan yang ketara, meletakkan asas untuk muzik asli Amerika, bersama-sama dengan muzik rakyat Ireland dan Inggeris. Dalam kes lain, muzik dan lagu-lagu patriotik seperti muzik rakyat dan negara mempunyai kesan yang mendalam pada setiap kelas, kaum dan masyarakat, seperti yang ditunjukkan oleh George McCue dalam bukunya "Muzik dan Amerika," yang menganalisis muzik dari segi arus sosial yang lebih luas yang wujud di sekeliling mereka menunjukkan bagaimana muzik boleh menggerakkan sentimen nasional semasa konflik.(McCue,2019).

2 PENYATAAN MASALAH

Terdapat kajian yang dibuat oleh Awang AA et al ., di mana mereka menggunakan dataset lagu patriotik untuk analisis semantik (Awang AA,2020) tetapi dataset yang digunakan hanya dalam lingkungan 2008 hingga 2013. Dalam konteks ini , tiada kajian yang menggunakan lagu melayu patriotik daripada 1988 hingga 2021 , mana set data terdiri daripada tempoh yang lebih lama dan meliputi konteks lirik atau teks yang lebih besar dalam lagu .

Tambahan pula, kajian hanya melakukan analisis semantik pada lirik tetapi tidak menganalisis emosi . Perkara ini disebabkan lagu-lagu menyentuh kita secara emosi tetapi tidak secara sentimen . Oleh itu , kajian dalam mengklasifikasikan klasifikasi emosi dengan bahasa ibunda kita diperlukan untuk memahami emosi manusia atau emosi yang dibawa dalam bentuk lirik .Industri hiburan , khususnya dalam penghasilan lagu-lagu patriotik , mempunyai keupayaan untuk mempengaruhi hati dan fikiran orang ramai untuk menerima atau menolak agenda yang disebarkan , serta memotivasikan orang ramai melalui mesej halus dalam lirik dan menimbulkan emosi pendengar . Namun , kesan lagu patriotik kepada pendengar jarang diperhatikan . Kajian ini juga menunjukkan bahawa beberapa kajian telah dilakukan untuk mengenal pasti kelemahan dan isu dengan perbendaharaan kata yang digunakan dalam lagu-lagu patriotik , serta bagaimana muzik memberi kesan kepada emosi pendengar melalui lirik lagu .

3 OBJEKTIF KAJIAN

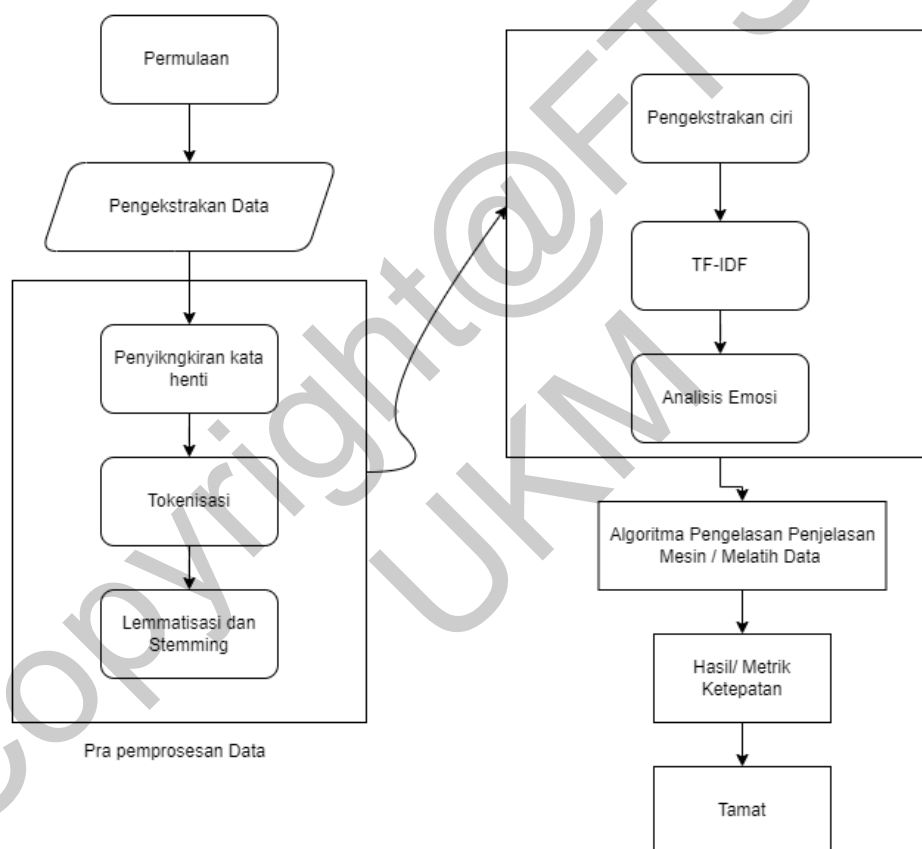
Objektif kertas kerja ini dibahagikan kepada tiga bahagian iaitu:

1. Mengklasifikasi dan mengkaji kepentingan perkataan atau lirik dalam lagu patriotik menggunakan data set lagu patriotik bahasa Melayu untuk mengklasifikasikan emosi manusia.
2. Untuk memahami dan mengenal pasti istilah utama yang boleh digunakan untuk klasifikasi sentimen dan emosi menggunakan lirik lagu patriotik sebagai set data
3. Untuk mencadangkan teknik klasifikasi untuk sentimen dan emosi menggunakan lirik lagu sebagai set data menilai kaedah yang dicadangkan melalui metrik penilaian. seperti skor F1, ketepatan, ketepatan dan metrik ingat semula.

4 METOD KAJIAN

Kajian ini dibangunkan menggunakan Model Air Terjun yang mudah untuk difahami dan digunakan. Dengan menggunakan kaedah ini, setiap fasa harus dilengkapkan sebelum fasa seterusnya dimulakan. Di penghujung setiap fasa, penilaian dibuat bagi memastikan projek berjalan seperti yang dirancang.

4.1 Sistem Model



Rajah 4.1 Carta Alir Keseluruhan Projek

Rajah 4.1 menunjukkan carta alir bagi keseluruhan algoritma projek pengelasan berasaskan pembelajaran mesin ini. Secara keseluruhannya, aliran projek ini terbahagi kepada lima bahagian iaitu pengekstrakan data, pra pemrosesan data, fasa pengekstrakan ciri, pembinaan algoritma klasifikasi pembelajaran mesin dan yang terakhir fasa pengujian untuk mendapatkan hasil menggunakan metrik penilaian.

Fasa pertama menitikberatkan pengekstrakan data dalam Bahasa Melayu oleh itu lagu-lagu yang dipilih hanya dalam bahasa melayu dan fokusnya ke atas lagu-lagu patriotik sahaja. Teknik pemprosesan Bahasa Tabii akan dilaksanakan pada fasa ke dua untuk memastikan tiada nilai nol dan data set adalah bersih dari elemen yang kurang penting seperti kata henti , token yang mempunyai kurang dari 3 perkataan dan kata imbuhan.

Fasa ketiga menggunakan Emosi Leksikon NRC dan TF-IDF untuk menganalisa emosi kepada 8 kelas dan mengekstrak ciri ciri penting dalam dataset yang telah diproses dalam fasa kedua. Fasa keempat merangkumi latihan data ke dalam empat model pilihan pembelajaran mesin iaitu “Naive Bayes” , “Random Forest Classifier”, “Logistic Regression” dan “SVM”.Akhir sekali, fasa kelima akan menghasilkan keputusan menggunakan metrik penilaian seperti F1-skor, ketepatan, “recall” dan “precision”.

4.2 Reka Bentuk Pra pemprosesan Data

Prapemprosesan data adalah sejenis pemprosesan yang dilakukan pada set data mentah untuk menyediakannya untuk prosedur pemprosesan lain seperti latihan data , pengekstrakan ciri dan agak penting untuk proses perlombongan data supaya latihan data dapat belajar dari ciri dan maklumat penting untuk menghasilkan hasil yang berkualiti dan optimum.

Terdapat beberapa langkah utama yang terlibat dalam prapemprosesan set data. Salah satunya ialah pembersihan data yang akan digunakan dalam projek ini. Dalam proses ini, data yang hilang dikenal pasti dan disusun dari set data. Penyelidik perlu memutuskan dan mengenal pasti sama ada lebih baik untuk membuang medan yang hilang, mengabaikannya atau mengisinya dengan median atau bermaksud nilai untuk set data berangka. Seterusnya, membuang data yang berulang perlu dilakukan kerana pengulangan data akan mempunyai ukuran yang sama yang direkodkan dua kali yang akan menjejaskan pemberat perkataan dalam fasa pengekstrakan ciri.

Selain itu, dalam projek ini, penyingkiran kata-kata berhenti, token yang kurang daripada 3 perkataan, simbol akan dikeluarkan kerana maklumat seperti ini tidak akan membantu dalam mempelajari ciri-ciri penting untuk latihan set data. Terjemahan teks juga

dikategorikan sebagai prapemrosesan data kerana ia dilakukan untuk menyeragamkan input untuk pengekstrakan ciri atau proses lain dalam saluran paip pembelajaran mesin dan lebih mudah untuk membandingkan antara set data kerana ia adalah dari bahasa yang sama dan mengeluarkan ciri yang sama untuk kedua-dua set data.

Kesimpulannya, prapemrosesan data menjadikannya lebih mudah untuk mentafsir dan menggunakan. Proses ini menghapuskan ketidakkonsistenan dan pertindihan data yang sebaliknya boleh menjejaskan ketepatan model secara negatif. Prapemrosesan data juga akan memastikan bahawa tiada nilai yang salah, nilai yang hilang dan mengurangkan nilai yang tidak relevan dalam set data.

4.3 Reka Bentuk Algoritma

Model yang dicadangkan untuk digunakan dalam kertas ini ialah Mesin Vektor Sokongan(SVM), Regresi Logistik, Naive Bayes dan Pengelas Hutan Rawak.

Naive Bayes akan menjadi model asas dalam projek ini. Ini disebabkan oleh Naive Bayes adalah algoritma pengelasan yang sangat pantas dan mudah yang sesuai untuk set data dimensi tinggi. Mereka menyediakan ramalan kebarangkalian yang sangat mudah menjadikannya mudah ditafsirkan sebagai model klasifikasi dasar awal.

Regresi logistik adalah lebih mudah untuk dilaksanakan, ditafsirkan, dan sangat cekap untuk dilatih. Ia sangat pantas dalam mengklasifikasikan rekod yang tidak diketahui. Ia berfungsi dengan baik apabila data set boleh dipisahkan secara linear. Ia boleh mentafsirkan model pekali sebagai penunjuk kepentingan ciri. Dalam kertas ini, kita akan menggunakan “Logistic Regression Multinomial” kerana kita mempunyai 8 kelas emosi dan 3 kelas sentimen.

Mesin Vektor Sokongan digunakan secara meluas untuk tugas klasifikasi kerana ia menggunakan kurang pengiraan dan memberikan ketepatan yang ketara. Vektor sokongan ialah titik data yang lebih dekat dengan hyperplane dan mempengaruhi kedudukan dan orientasi hyperplane . Menggunakan vektor sokongan ini, kami memaksimumkan margin pengelas . SVM akan mempunyai hyperplanes yang membantu dengan sempadan keputusan terutamanya jika tugas pengelasan mempunyai banyak ciri atau kelas untuk dikelaskan.

Pengelas Hutan Rawak ialah salah satu pengelas terbaik yang digunakan secara meluas untuk tugas pengelasan. (Jalal, N. et al., 2022). Kesederhanaan algoritma mengubah teks pilihan yang menarik untuk pengelasan dan keupayaannya untuk mengendalikan data berdimensi tinggi malah dan prestasi tinggi di bawah set data tidak seimbang kelebihan ketara berbanding model Pembelajaran Mesin yang lain. (Khoshgoftaar, T. M., 2007).

4.4 Reka Bentuk Eksperimen

Metrik penilaian yang akan digunakan ialah skor F1, Ketepatan (Precision) dan Ingat (Recall).

Skor F1 akan digunakan kerana ia menyampaikan keseimbangan antara ketepatan dan penarikan balik. Seperti yang dinyatakan dalam artikel Penilaian: daripada ketepatan, ingat semula dan ukuran F kepada ROC, kemakluman, penandaan dan korelasi. (Powers DM, 2020), formula F-Measure ditakrifkan dalam (1):

$$\text{skor} - F1 = \frac{(2 \times \text{Ketepatan}(\text{Precision}) \times \text{Ingat}(\text{Recall}))}{(\text{Ketepatan} + \text{Ingat})} \dots \dots (1)$$

Ketepatan (Precision) digunakan dalam kes ini apabila objektif kami adalah untuk meminimumkan positif palsu, dan mengingat semula digunakan apabila objektif adalah untuk meminimumkan negatif palsu. Secara umum, kami mengoptimumkan prestasi model kami dengan menggunakan metrik yang dipilih. Formula yang dinyatakan dalam artikel Penilaian: daripada ketepatan, ingat semula dan ukuran F kepada ROC, kemakluman, penandaan dan korelasi. (Powers DM, 2020), untuk ketepatan ditakrifkan dalam (2) dan formula untuk mengingat semula yang dinyatakan dalam kajian yang bertajuk “pembelajaran tidak seimbang” (He et al., 2013) ditakrifkan dalam (3):

$$\text{Ketepatan}(\text{Precision}) = \frac{(\text{Kadar Positif Benar})}{(\text{Kadar Positif Benar} + \text{kecerunan} \times \text{Kadar Positif Palsu})} \dots \dots (2)$$

$$\text{Ingat}(\text{Recall}) = \frac{(\text{Positif Benar})}{(\text{Positif Benar} + \text{Negatif Palsu})} \dots \dots (3)$$

Dalam kes ini, ketepatan juga akan digunakan sebagai metrik penilaian untuk melihat peratusan ramalan yang betul untuk data ujian seperti yang dinyatakan dalam kajian Sensitiviti, kekhususan, ketepatan, selang keyakinan yang berkaitan dan analisis ROC dengan pelaksanaan SAS praktikal (Zhu, W. et al., 2010) formula ditakrifkan dalam (4)

$$\text{Ketepatan} = \frac{(\text{Negatif Benar} + \text{Positif Benar})}{(\text{Negatif Benar} + \text{Positif Benar} + \text{Negatif Palsu} + \text{Positif Palsu})} \dots (4)$$

Matriks kekeliruan yang ditunjukkan di Rajah 4.2 biasanya digunakan untuk menggambarkan analitik kepentingan seperti ingat kembali, kekhususan, ketepatan dan ketepatan. Matriks kekeliruan dikatakan sangat berguna kerana ia memberikan perbandingan nilai-nilai seperti Positif Benar, Positif palsu, Negatif benar dan Negatif Palsu. Ia ialah matriks NxN yang digunakan untuk menilai prestasi model klasifikasi, memandangkan N ialah bilangan kelas sasaran.

	Benar	
Diramal	Positif Benar	Positif Palsu
	Negatif Palsu	Negatif Benar

Rajah 4.2 Matriks Kekeliruan

5 HASIL KAJIAN

5.1 Pengekstrakan Data

Berdasarkan rajah 5.1, Set data telah disediakan dengan 83 lagu secara keseluruhan khususnya dari 1988 hingga 2020 yang diekstrak daripada YouTube. Dalam proses penyediaan dataset, redundansi lirik yang ketara berlaku dari tahun 2000 hingga 2006, ini disebabkan lagu “Keranamu Malaysia” digunakan tanpa sebarang perubahan pada lirik, tetapi perubahan dibuat pada melodi sahaja. Bagi mengatasi isu ini, lagu Negaraku dan lagu 14 negeri ditambah serta 35 lagu patriotik diperoleh daripada web merangkak. Set data disediakan dalam bahasa Melayu dan set data yang digunakan dalam keseluruhan proses latihan akan digunakan dalam bahasa Melayu dan kemungkinan besar akan diterjemahkan ke bahasa lain. Prapemprosesan data seperti membuang tanda baca, kata henti, tokenisasi dan lemmatisasi telah dilakukan.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Title	Lyrics												
2	1988 Bersatu	BERSATU.BERSATU BERSATU. BERSATU MENENTANG MUSUH NEGARA PERTAHANKAN MARUAH BANGSA ERATKAN PERPADUAN MALAYSIA T												
3	1989 Bersatu	BERSATU.BERSATU BERSATU. BERSATU MENENTANG MUSUH NEGARA PERTAHANKAN MARUAH BANGSA ERATKAN PERPADUAN MALAYSIA T												
4	1990 Berjaya	MALAYSIA KITA SUDAH BERJAYAAMAN MAKMUR BAHAGIAMALAYSIA ABADI SELAMANYABERJAYA DAN BERJAYABERBAGAI KAUM SUDAH BER												
5	1991 Wawasan 2020	BENDERA BERKIBAR DI ANGKASALAMBANG NEGARA JAYARAKYAT SEPAKAT SEHATI SEJIWADENGAN BERBAKTI DENGAN MEGAHHADAPI CABA												
6	1993 Bersatu Menuju Wawasan	MURNINYA SEBUAH WAWASAN MEMBINA BUDAYA YANG GEMILANG BANGUNKAN NUSA MAJUKANLAH BANGSA MENUJU PUNCAK JAYA MA												
7	1994 Nilai Murni Jayakan Wawasan	NILAI MURNI JAYAKAN WAWASAN BUKTINYA BANGSA BERHEMAM MULIA BERHATI NURANI TUNAIKAN URUSAN RAJIN BERUSAHAKESETIAAN												
8	1995 Jatidiri Penggerak Wawasan	VISI WAWASAN SUMBER KESEJAHTERAAN MAJULAH KE HADAPAN MASA DEPAN CEMERLANG TEGUH KITA BINA TAMADUN ABADI INDUSTRI I												
9	1996 Budaya Penentu Kecapaian	KE ARAH BUDAYA MURNI ITULAH SATU WAWASAN KITA WUJUDKAN SATU BANGSA BEBAS DAN MERDEKA LANGKAH SEHALA SEIRINGAN GEM												
10	1997 Akhlak Mulia Masyarakat Berjaya	BERBEKALKAN ILTIZAM YANG PADU KITA BANGUN DENGAN SEMANGAT BARU MENJUJUNG TINGGI MATLAMAT BANGSA SEIRING JALAN SA												
11	1998 Negara Kita Tanggungjawab Kita	MENDUKUNG CITA-CITA NEGARA ITULAH TEKAD KITA BERSATU BERKERJASAMA SEHALA SEIRINGAN GERAK MESRA KITA UTAMAKAN MENYAI												
12	1999 Bersatu Ke Alaf Baru	KIBARKAN JALUR GEMILANG LAMBANG KECEMERLANGAN KITA RAKYAT PENGGERAK WAWASAN BAKTI JASA DICURAHKAN MELANGKAH DENI												
13	2000 Keranamu Malaysia	BURUH NELAYAN DAN JUGA PETANI GAYA HIDUP KINI DAH BERUBAH ANAK-ANAK TERASUH MINDANYA LAHIR GENERASI BIJAK PANDAI PAKA												
14	2001 Keranamu Malaysia	BURUH NELAYAN DAN JUGA PETANI GAYA HIDUP KINI DAH BERUBAH ANAK-ANAK TERASUH MINDANYA LAHIR GENERASI BIJAK PANDAI PAKA												
15	2002 Keranamu Malaysia	BURUH NELAYAN DAN JUGA PETANI GAYA HIDUP KINI DAH BERUBAH ANAK-ANAK TERASUH MINDANYA LAHIR GENERASI BIJAK PANDAI PAKA												
16	2003 Keranamu Malaysia	BURUH NELAYAN DAN JUGA PETANI GAYA HIDUP KINI DAH BERUBAH ANAK-ANAK TERASUH MINDANYA LAHIR GENERASI BIJAK PANDAI PAKA												
17	2004 Keranamu Malaysia	BURUH NELAYAN DAN JUGA PETANI GAYA HIDUP KINI DAH BERUBAH ANAK-ANAK TERASUH MINDANYA LAHIR GENERASI BIJAK PANDAI PAKA												
18	2005 Keranamu Malaysia	BURUH NELAYAN DAN JUGA PETANI GAYA HIDUP KINI DAH BERUBAH ANAK-ANAK TERASUH MINDANYA LAHIR GENERASI BIJAK PANDAI PAKA												
19	2006 Keranamu Malaysia	BURUH NELAYAN DAN JUGA PETANI GAYA HIDUP KINI DAH BERUBAH ANAK-ANAK TERASUH MINDANYA LAHIR GENERASI BIJAK PANDAI PAKA												
20	2006 Keranamu Malaysia	BURUH NELAYAN DAN JUGA PETANI GAYA HIDUP KINI DAH BERUBAH ANAK-ANAK TERASUH MINDANYA LAHIR GENERASI BIJAK PANDAI PAKA												
21	2007 Malaysiaku Gemilang	MARILAH KITA SEMUA ATAS NAMA NEGARA BANGSA DENGAN TEKAD MULIA MAJU BERWAWASAN MENCIPTA KEAGUNGAN BERPADULAH KIT												
22	2008 Perpaduan Teras Kejayaan	KINI SEMANGAT JALAN SEIRING SATU AMALAN KERJASAMA JADI TERAS BANGS MULIA SEHATI SEJIWA BANGUN MEMBANTU AGAR SEPAKAT I												
23	2009 1Malaysia, Rakyat Didahulukan, Pencapaian C	BERDIRI TEGUH DI BUMI NYATA PUPUK SEMANGAT 1MALAYSIA SEPAKAT RAKYAT DIDAHULUKAN PENCAPAIAN DIUTAMAKAN PERPADUAN M												
24	2010 1Malaysia, Menjana Tranformasi	BERBEKALKAN ILTIZAM YANG PADU BANGUN DENGAN SEMANGAT BARU WALAU RANJAU PENUH BERLIKU BERUSAHA TANPA RASA JEMU ME												
25	2011 1Malaysia, Transformasi Berjaya, Rakyat Seja	KEMAKMURAN KEDAMAIAN TERUS DINIKMATI KEJAYAAN KEMASYHURAN SAMA DIRASAI AYUHLAH KITA MEREALISASIKANNYA KITALAH PEMU												
26	2012 1Malaysia, Janji Ditepati	MACAM MALAM IKUT SIANG MACAM BULAN IKUT MATAHARI 1MALAYSIA TERUS BERJUANG JANJI KITA DITEPATI KATA KITA DIKOTA JANJI KI												
27	2013 Malaysiaku Berdaulat: Tanah Tumpahnya Dar	MARA KITA HALILINTAR TIADA GENTAR PUNCAK DUNIA MALAYSIA SUDAH BERGEMA KITA HALILINTAR TIADA GENTAR BENDERA KITA TERUS												
28	2014 Malaysia... Di Sini Lahirnya Sebuah Cinta	CINTA CINTA KEPADA NUSA KINI DAN SELAMANYA TAK BERBELAH BAGI CINTA BERGEMA SEJARAH KEBANGKITAN SEMANGAT BAGAI PAHLAW												

Rajah 5.1 Set data Lagu –Lagu patriotik

5.2 Penterjemahan Data Set

```

headers = ['Title', 'Lyrics', 'Title_en', 'Lyrics_en']
data = pd.read_csv('patriotic_dataset.csv')
translator = Translator()

df = pd.DataFrame(index=range(0, len(data)), columns=headers)

def translate_row(row):
    ''' Translate elements A and B within `row`. '''
    a = translator.translate(row[0], dest='en')
    b = translator.translate(row[1], dest='en')
    return pd.Series([a.origin, b.origin, a.text, b.text], headers)

for i, row in enumerate(data.values):
    # Fill empty dataframe with given serie.
    df.loc[i] = translate_row(row)

print(df)

```

Rajah 5.2 Kod penterjemahan data set

Penterjemahan set data daripada bahasa Melayu ke bahasa Inggeris dilakukan menggunakan googletrans . Sel pertama pada rajah 4.6 sedang menamakan 2 lajur baharu yang akan terdiri daripada tajuk terjemahan lagu dan lirik terjemahan. Sel kedua ialah menterjemah unsur A

dan B yang masing-masing terdiri daripada tajuk dan lirik lagu dalam bahasa Melayu . Ia dikembalikan sebagai data satu dimensi untuk dimuatkan dalam setiap baris set data.

5.3 Reka Bentuk Pra Pemprosesan Data

```
def remove_punct(text):
    return"".join([ch for ch in text if ch not in st.punctuation])
```

Python

```
df['removed_punct'] = df['Lyrics_en'].apply(lambda x: remove_punct(x))
df.head()
```

Python

	Title_en	Lyrics_en	clean_text	removed_punct
0	1988 united	United. Be united.Unite against the enemy of t...	united unitedunite enemy country defend dignit...	United Be unitedUnite against the enemy of the...

Rajah 5.3 Mengeluarkan tanda baca

Rajah 5.3 di atas menunjukkan fungsi mengalih keluar tanda baca seperti titik , tanda seru dan koma yang tidak membantu dalam projek ini.

```
def tokenize(text):
    text = re.split('\s+', text)
    return[x.lower() for x in text]
```

Python

```
df['tokens'] = df['removed_punct'].apply(lambda msg : tokenize(msg))
df.head()
```

Python

	Title_en	Lyrics_en	clean_text	removed_punct	tokens
0	1988 united	United. Be united.Unite against the enemy of t...	united unitedunite enemy country defend dignit...	United Be unitedUnite against the enemy of the...	[united, be, unitedunite, against, the, enemy,...

Rajah 5.4 Tokenisasi Data set

Dalam rajah 5.4, tokenisasi telah dilakukan untuk memecahkan data tidak berstruktur dan teks bahasa semula jadi kepada cebisan maklumat untuk lebih memahami konteks dalam membangunkan model untuk NLP.

```
#remove tokens of less than 3
def remove_small_words(text):
    return [ x for x in text if len(x) > 3]
```

Python

```
df['larger_tokens'] = df['tokens'].apply(lambda x : remove_small_words(x))
df.head()
```

Python

	Title_en	Lyrics_en	clean_text	removed_punct	tokens	larger_tokens
0	1988 united	United. Be united.Unite against the enemy of t...	united unitedunite enemy country defend dignit...	United Be unitedUnite against the enemy of the...	[united, be, unitedunite, against, the, enemy,...	[united, unitedunite, against, enemy, country,...

Rajah 5.5 Mengeluarkan token yang mempunyai kurang daripada tiga perkataan

Rajah 5.5 menunjukkan dalam kajian ini ayat yang mengandungi kurang daripada 3 perkataan dikeluarkan seperti “ an ”, “ a ”, “ be ” kerana perkataan tersebut tidak memainkan peranan penting kerana model tidak benar-benar belajar daripadanya.

```
#Removing stopwords
def remove_stopwords(text):
    return [word for word in text if word not in nltk.corpus.stopwords.words('eng
```

Python

```
df['clean_tokens'] = df['larger_tokens'].apply(lambda x: remove_stopwords(x))
df.head()
```

Python

	Lyrics_en	clean_text	removed_punct	tokens	larger_tokens	clean_tokens
88 ed	United. Be united.Unite against the enemy of t...	united unitedunite enemy country defend dignit...	United Be unitedUnite against the enemy of the...	[united, be, unitedunite, against, the, enemy,...	[united, unitedunite, against, enemy, country,...	[united, unitedunite, enemy, country, defend, ...

Rajah 5.6 Mengeluarkan perkataan berhenti

Berdasarkan rajah 5.6 kata henti seperti “a”, “ the ”, “ is ” , “dan” dialih keluar dan dihapuskan kerana ini biasanya digunakan yang membawa sedikit maklumat berguna.

```
# Apply lemmetization on tokens
def lemmatize (text):
    word_net = WordNetLemmatizer()
    return [word_net.lemmatize(word) for word in text]

df['lemma_words'] = df['clean_tokens'].apply(lambda x : lemmatize(x))
df.head()
```

clean_text	removed_punct	tokens	larger_tokens	clean_tokens	lemma_words
united unitedunite enemy country defend dignit...	United Be unitedUnite against the enemy of the...	[united, be, unitedunite, against, the, enemy,...	[united, unitedunite, against, enemy, country,...	[united, unitedunite, enemy, country, defend, ...	[united, unitedunite, enemy, country, defend, ...

Rajah 5.7 Menggunakan lemmatisasi pada token yang telah dibersihkan

Rajah 5.7 memnunjukkan kajian ini memerlukan makna perkataan dalam bentuk akar kerana ia lematisasi dipilih berbanding dengan berpunca. Lemmatisasi boleh membantu mengekalkan makna perkataan. Contohnya, kajian akan ditukar kepada kajian selepas lemmatisasi .

Lyrics_en	removed_punct	tokens	larger_tokens	clean_tokens	lemma_words	clean_text
BERSATU BERSATU BERSATU BERSATU AGAINST THE EN...	BERSATU BERSATU BERSATU BERSATU AGAINST THE EN...	[bersatu, bersatu, bersatu, bersatu, against, ...	[bersatu, bersatu, bersatu, bersatu, against, ...	[bersatu, bersatu, bersatu, bersatu, enemies, ...	[bersatu, bersatu, bersatu, bersatu, enemy, na...	bersatu bersatu bersatu bersatu enemy nation d...
BERSATU BERSATU BERSATU BERSATU AGAINST THE EN...	BERSATU BERSATU BERSATU BERSATU AGAINST THE EN...	[bersatu, bersatu, bersatu, bersatu, against, ...	[bersatu, bersatu, bersatu, bersatu, against, ...	[bersatu, bersatu, bersatu, bersatu, enemies, ...	[bersatu, bersatu, bersatu, bersatu, enemy, na...	bersatu bersatu bersatu bersatu enemy nation d...
Malaysia we have succeeded in prosperity and d...	Malaysia we have succeeded in prosperity and d...	[malaysia, we, have, succeeded, in, prosperity...	[malaysia, have, succeeded, prosperity, divers...	[malaysia, succeeded, prosperity, diverse, div...	[malaysia, succeeded, prosperity, diverse, div...	malaysia succeeded prosperity diverse diversif...
Flag flags in the country of the country of th...	Flag flags in the country of the country of th...	[flag, flags, in, the, country, of, the, count...	[flag, flags, country, country, people, world,...	[flag, flags, country, country, people, world,...	[flag, flag, country, country, people, world, ...	flag flag country country people world result ...
It is a vision of building a glorious culture ...	It is a vision of building a glorious culture ...	[it, is, a, vision, of, building, a, glorious,...	[vision, building, glorious, culture, develop,...	[vision, building, glorious, culture, develop,...	[vision, building, glorious, culture, develop,...	vision building glorious culture develop natio...

Rajah 5.8 Akhir hasil daripada pra pemprosesan daripada data

Rajah 5.8 menunjukkan yang muktamad hasil data di mana yang set data ialah sekarang diterjemahkan , dibersihkan dan tiada redundansi daripada data.

5.4 Pengekstrakan Ciri

	TF-IDF
unite	0.665784
strengthen	0.454624
nation	0.276314
malaysia	0.254145
successful	0.197201
love	0.193132
enemy	0.171799
defend	0.151541
dignity	0.151541
village	0.144311
city	0.132903

Rajah 5.9 10 Teratas perkataan yang paling kerap dalam set data

Rajah 5.9 menunjukkan 10 perkataan paling kerap dalam set data patriotik Melayu yang "bersatu" mengambil tempat pertama, "kekuatan" sebagai yang kedua paling kerap dan "negara" di tempat ketiga. Ia dapat dilihat bahawa 10 perkataan teratas yang muncul paling kerap membawa rasa positif dalam pendengar kerana itu tidak syak lagi bahawa emosi seperti "kepercayaan", "kegembiraan" dan "jangkaan" akan mengambil sebagai kekerapan emosi tertinggi selepas menggunakan kaedah frekuensi NRC.

5.5 Pengaplikasian Emosi Leksikon NRC

```
#Function of retrieving nrc affect frequencies
def emotion_freq(lyrics):
    emotions = {'anger':0.0, 'anticipation':0.0, 'disgust':0.0, 'fear':0.0, 'jo
    #creating object
    lyrics = NRCLex(lyrics)
    #using affect frequencies method
    freq = lyrics.affect_frequencies
    #Iterates through list
    for k, fq in freq.items():
        emotions[k] = emotions.get(k,0.0) + fq
    return emotions

#calculate word count in each headline
def word_count(row):
    row = nltk.word_tokenize(row)
    count = len(row)
    return count
```

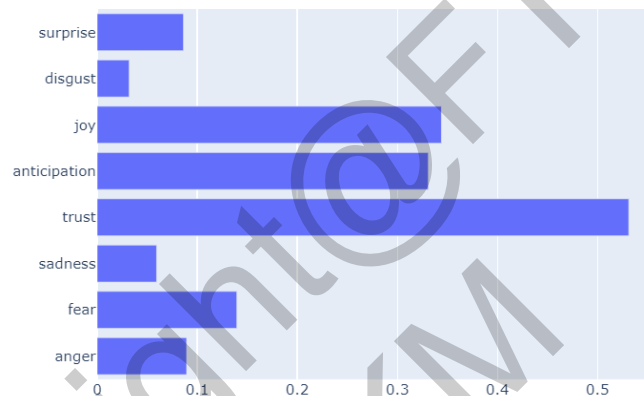
Python

```
#retrieving affect frequencies in each lyrics(row)
df['emotions_frequency'] = df['clean_text'].apply(emotion_freq)
```

Python

Rajah 5.10 Pengaplikasian Emosi Leksikon NRC

Berdasarkan Rajah 5.10 , kod tersebut menunjukkan metod “affect frequency” untuk menghasilkan frekuensi lapan emosi bagi setiap lagu atau setiap baris. Ianya akan berulang sehingga senarai dalam data set mencapai baris terakhir. Leksikon NRC Emosi merupakan senarai perkataan Bahasa Inggeris dan perkaitannya dengan lapan emosi iaitu kemarahan, ketakutan, jangkauan, kepercayaan, terkejut, kesedihan, kegembiraan dan jijik serta dua kelas sentimen iaitu positif dan negatif tetapi fokus utama projek ini adalah hanya kepada analisis emosi. Fungsi seterusnya pada rajah 4.4 adalah untuk mengira jumlah perkataan bagi setiap baris yang untuk tujuan normalisasi.

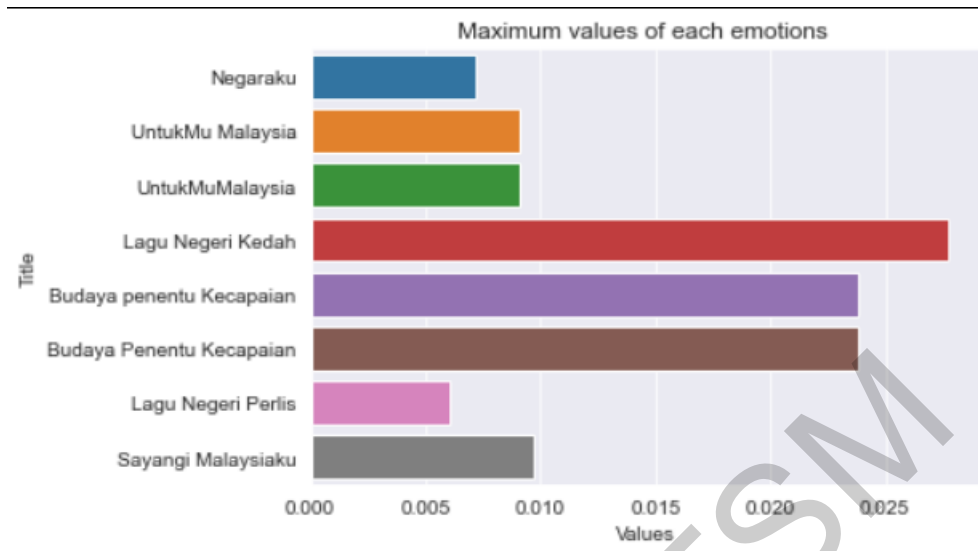


Rajah 5.11 Frekuensi maksimum lapan kelas emosi dalam graf bar

Berdasarkan rajah 5.10 dan 5.11, emosi kepercayaan dapat dilihat mempunyai jumlah kekerapan tertinggi dan kekerapan meluat terendah. Hal ini kerana, dalam lagu-lagu patriotik, perkataan yang melibatkan pemerintah negara sering muncul dan dalam konteks ini rakyat mempercayai pemimpin dan mengamankan mereka untuk memimpin negara.

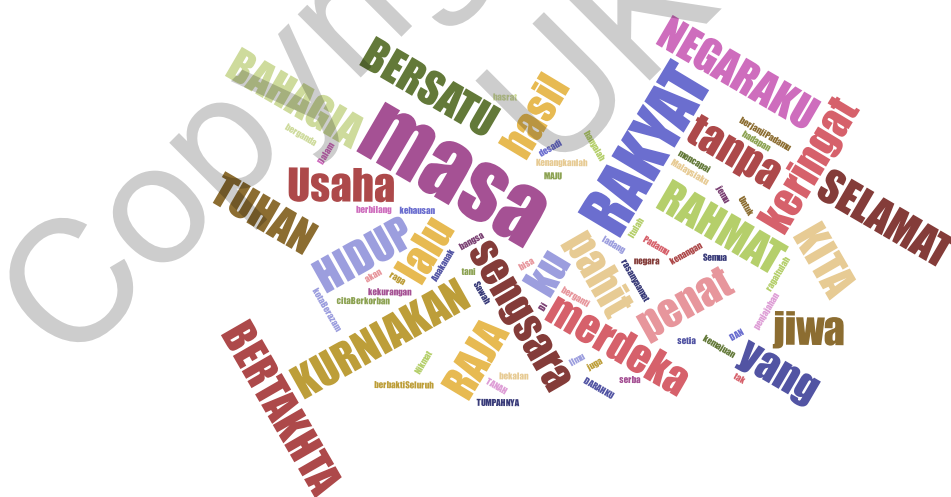
EMOTIONS	TITLE	MAX VALUES
anger	Negaraku	0.007142857
fear	Untukmu Malaysia	0.009090909
sadness	Untukmu Malaysia	0.009090909
trust	Lagu Negeri Kedah	0.027777778
anticipation	Budaya penentu Kecapaian	0.023809524
joy	Budaya penentu Kecapaian	0.023809524
disgust	Lagu Negeri Perlis	0.006060606
surprise	Sayangi Malaysiaku	0.00974026

Rajah 5.12 Lagu yang memperoleh nilai maksimum bagi setiap emosi



Rajah 5.13 Lagu yang memperoleh nilai maksimum bagi setiap emosi dalam graf bar

Berdasarkan Rajah 5.12 dan 5.13, Lagu Negeri Kedah paling banyak mengeluarkan frekuensi untuk emosi kepercayaan, UntukMuMalaysia mempunyai frekuensi ketakutan dan kesedihan yang paling banyak, Budaya penentu Kecapaian nampaknya mempunyai emosi yang bertentangan dengan UntukMu Malaysia iaitu kegembiraan. dan jangkauan.



Rajah 5.14 Wordcloud lirik lagu yang mempunyai frekuensi tinggi untuk emosi kesedihan, kemarahan, kejutan, ketakutan

Berdasarkan rajah 5.14 , perkataan perkataan seperti pahit , penat , sengsara sering muncul di dalam lirik lagu-lagu patriotik menyebabkan hasil frekuensi emosi terdorong kepada kelas kesedihan. Lagu negaraku nampaknya dikategorikan kepada kelas kemarahan

kerana berdasarkan sejarah lagu ini dicipta selepas merdeka di mana ianya membawa semangat berkobar kobar untuk berbakti kepada negara dan emosinya boleh dikaitkan dengan emosi kemarahan.



Rajah 5.15 Wordcloud lirik lagu yang mempunyai frekuensi tinggi untuk emosi kegembiraan, jangkaan dan kepercayaan

Berdasarkan rajah 5.15, Analisa dapat dilakukan bahawa lirik yang mengandungi perkataan seperti bertuah, sempurna, sejahtera, gembira dan rahmat akan membawa frekuensi yang lebih tinggi dalam kegembiraan dan penantian kerana ia membawa maksud positif. Lagu Negeri Kedah dalam konteks ini mempunyai perkataan seperti Allah yang bermaksud tuhan, Sultan Mahkota, dan nabi-nabi yang membawa maksud pemerintah yang rakyat amanahkan kepada mereka. Budaya Penentu Kecapaian yang membawa kekerapan tertinggi untuk kegembiraan dan penantian terdapat perkataan seperti “ bina semangat ”, “ bebas dan merdeka ” “ beretika ” yang membawa maksud kebebasan yang dinantikan.

5.6 Pembinaan Model Klasifikasi

Jadual 5.1 Ketepatan dan perbandingan kedua-dua set data

Model Pembelajaran Mesin	Ketepatan Set Data Patriotik Melayu	Ketepatan data set Lagu Kebangsaan Dunia
Naif Bayes	0.571	0.667
Regresi Logistik	0.762	0.604
Mesin Vektor Sokongan	0.619	0.604
Pengelas Hutan Rawak	0.762	0.624

Berdasarkan jadual di atas, Pengelas Rawak Hutan untuk Set Data Patriotik Melayu memberikan ketepatan tertinggi sebanyak 0.762 dan ketepatan SVM dan Regresi Logistik Lagu Kebangsaan kumpulan data dunia berkongsi nilai yang sama iaitu 0.604 . Ketepatan Naif Bayi bagi data set patriotik Melayu meletakkan nilai terendah 0.571.

5.7 Keputusan Ujian

Jadual 5.2 F1-skor dan perbandingan antara kedua-dua set data

Emosi	F1-skor Set Data Patriotik Melayu	F1-skor data set Lagu Kebangsaan Dunia
Kemarahan	0.00	0.00
Jangkaan	0.00	0.00
Kegembiraan	0.00	0.00
Kepercayaan	0.83	0.74

Jadual 5.3 “Precision” dan perbandingan antara kedua-dua set data

Emosi	“Precision” Set Data Patriotik Melayu	“Precision” data set Lagu Kebangsaan Dunia
Kemarahan	0.00	0.00
Jangkaan	0.00	0.00
Kegembiraan	0.00	0.00
Kepercayaan	0.71	0.58

Jadual 5.4 “Recall” dan perbandingan antara kedua-dua set data

Emosi	“Recall” Set Data Patriotik Melayu	“Recall” data set Lagu Kebangsaan Dunia
Kemarahan	0.00	0.00
Jangkaan	0.00	0.00
Kegembiraan	0.00	0.00
Kepercayaan	1.00	1.00

Berdasarkan jadual 5.2, 5.3 dan 5.4, kepercayaan emosi memperoleh kekerapan tertinggi di kalangan semua emosi di mana ia mendapat skor 0.83, 0.71 dan nilai optimum 1 untuk F1-skor, ketepatan, dan ingat masing-masing. Set data lagu Patriotik Melayu boleh dilihat untuk mengatasi nilai Set DataSet Lagu Kebangsaan dalam penilaian metrik.

Sebab-sebab di mana terdapat nilai 0 dalam jadual di atas mungkin kerana data tidak mempunyai maklumat yang mencukupi untuk meramalkan label dan pengelas membenarkan

kelas yang paling kerap iaitu emosi kepercayaan mendominasi oleh itu emosi kepercayaan memperoleh nilai frekuensi tertinggi dalam semua kaedah penilaian metrik.

6 KESIMPULAN

Lagu-lagu patriotik seharusnya mampu memotivasikan pendengar tentang cintakan negara dengan mempertingkatkan mesej yang disampaikan melalui lirik yang digubah kerana terbukti bahawa lirik juga boleh mempengaruhi dan menyemarakkan emosi di hati manusia. Oleh itu, adalah penting untuk menyusun lirik dengan berhati-hati untuk menyampaikan emosi positif kepada pendengar.

Kajian ini menunjukkan lapan emosi atau lapan kelas untuk mengklasifikasikan lirik lagu patriotik iaitu kemarahan, jangkitan, kegembiraan, kepercayaan, ketakutan, terkejut, sedih dan meluat dan bukannya mengklasifikasikannya kepada tiga nilai sentimen asas iaitu positif, negatif dan neutral sahaja. Dalam konteks ini, kertas kerja ini menunjukkan emosi yang lebih spesifik dan membuktikan melalui lirik juga dapat menyemarakkan perasaan patriotisme di hati rakyat.

Kajian ini menyimpulkan bahawa lirik lagu-lagu patriotik semasa membawakan emosi “kepercayaan” yang paling banyak kepada pendengar dengan kekerapan 0.278 sebagai maksimum antara kesemua lapan emosi yang dirujuk kepada roda emosi Plutchik (Plutchik, 1980). Komposer boleh mengambil perhatian pada perkataan yang digunakan untuk menunjukkan pemerintah negara, perpaduan rakyat untuk digunakan dalam lagu-lagu patriotik akan datang.

Kajian ini boleh ditambah baik dengan menambah lebih banyak set data dan menambah lebih banyak ciri pada set data untuk mendapatkan lebih banyak maklumat mengenai lirik yang digunakan oleh komposer untuk membentuk pendapat umum yang positif terhadap negara melalui hiburan.

Kesimpulannya, dapat kita lihat lirik yang dibuat bukan sahaja melahirkan kesan positif seperti kegembiraan, penantian dan kepercayaan tetapi pada masa yang sama, ia juga akan menunjukkan emosi negatif seperti meluat dan takut. Ini disebabkan pemilihan perkataan untuk lagu dalam garis masa tertentu menyerupai isu dan kesan agenda tertentu terhadap pendengar.

7 RUJUKAN

- Abbasi, MM, & Beltiukov, AP (2019). Meringkaskan emosi daripada teks menggunakan roda emosi Plutchik. Dalam Persidangan Saintifik mengenai Teknologi Maklumat untuk Sokongan Membuat Keputusan Pintar.
- Abdillah , J., Asror , I., & Wibowo, YFA (2020). Klasifikasi emosi lirik lagu menggunakan kaedah lstm dwiarah dengan pemberat perwakilan perkataan sarung tangan. Jurnal RESTI (Kejuruteraan Sistem dan Teknologi Maklumat), 4 (4), 723-729.
- Ahuja, M., & Sangal , AL (2018, Disember). Perlombongan pendapat dan klasifikasi lirik muzik menggunakan algoritma pembelajaran yang diselia. Pada 2018 Persidangan Antarabangsa Pertama mengenai Pengkomputeran dan Komunikasi Siber Selamat (ICSCCC) (ms. 223-227). IEEE.
- Awang, AA, Lokman , AM, Fathir , MFM, & Salleh, SM (2016). Analisis Semantik Lagu Patriotik Malaysia. Surat Sains Lanjutan, 22 (5-6), 1373-1379.
- Durahim , AO, SETİREK, AC, ÖZEL, BB, & Kebapci , H. (2018). Klasifikasi emosi muzik untuk lagu Turki menggunakan lirik. Universiti Pamukkale Mühendislik Bilimleri Dergisi , 24 (2), 292-301.
- He, H., & Ma, Y. (Eds.). (2013). Pembelajaran tidak seimbang: asas, algoritma dan aplikasi.
- Mohammad, SM, & Turney, PD (2013). Crowdsourcing perkataan – leksikon persatuan emosi. Kecerdasan pengiraan, 29 (3), 436-465.
- McCue, G. (2019). Muzik dalam Masyarakat Amerika. Routledge.
- Mohammad, SM, & Turney, PD (2013). Leksikon emosi NRC. Majlis Penyelidikan Kebangsaan, Kanada.
- Powers, DM (2020). Penilaian: daripada ketepatan, ingat semula dan ukuran F kepada ROC , kemakluman , penandaan dan korelasi. pracetak arXiv arXiv : 2010.16061.
- Joachims, T. (1998, April). Pengkategorian teks dengan mesin vektor sokongan: Belajar dengan banyak ciri yang berkaitan. Dalam persidangan Eropah mengenai pembelajaran mesin (ms. 137-142). Springer, Berlin, Heidelberg.
- Jalal, N., Mehmood, A., Choi, GS, & Ashraf, I. (2022). Sebuah novel menambah baik hutan rawak untuk klasifikasi teks menggunakan kedudukan ciri dan bilangan pokok yang optimum. Jurnal Universiti King Saud-Sains Komputer dan Maklumat.
- Khoshgoftaar, TM, Golawala, M., & Van Hulse, J. (2007, Oktober). Kajian empirikal pembelajaran daripada data tidak seimbang menggunakan hutan rawak. Dalam 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007) (Jilid 2, ms. 310-317). IEEE.
- Ohana , B., & Tierney, B. (2009). Klasifikasi sentimen ulasan menggunakan SentiWordNet . Prosiding IT&T.
- Wang, J., & Yang, Y. (2019). Penandaan mood berasaskan pembelajaran mendalam untuk lirik lagu Cina. pracetak arXiv arXiv : 1906.02135.
- Xiang, Z., Dong, X., Li, Y., Yu, F., Xu, X., & Wu, H. (2020). Model Pengecaman Emosi Bimodal untuk Lagu Minnan . Maklumat, 11 (3), 145.
- Zhu, W., Zeng, N., & Wang, N. (2010). Sensitiviti, kekhususan, ketepatan, selang keyakinan yang berkaitan dan analisis ROC dengan pelaksanaan SAS praktikal. Prosiding NESUG: penjagaan kesihatan dan sains hayat, Baltimore, Maryland, 19, 67.
- Data tidak seimbang menggunakan hutan rawak. Dalam Perisikan IEEE ke-19 (ICTAI 2007) (Jilid 2, ms. 310-317). IEEE.

Natasya Izzaty Khoo (A177696)
 Fakulti Teknologi & Sains Maklumat,
 Universiti Kebangsaan Malaysia