# ANALISIS SENTIMEN PILIHAN RAYA UMUM MALAYSIA KE-15 PADA MEDIA SOSIAL

TAN JING XUAN

MOHD RIDZWAN YAAKUB

*Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia*

**ABSTRAK**

Pilihan raya merupakan proses mengundi wakil rakyat yang amat penting bagi sesebuah negara demokrasi. Setiap warganegara yang berlayak mengundi berhak untuk mengundi calon-calon yang berkemampuan untuk memimpin negara ke arah cemerlang pada masa depan. Analisis Sentimen adalah sejenis teknik penganalisisan yang membolehkan pengkaji menganalisis sentimen-sentimen dari segi ayat perkataaan. Ia boleh diaplikasikan melalui algorithma seperti kaedah *lexicon-based*, atau diaplikasikan dalam bidang pembelajaran mesin dan rangkaian neural buatan. Dalam kajian ini, pengkaji akan menekankan pengaplikasian analisis sentimen berasasakan kaedah *lexicon-based* dan pembelajaran mesin. Pengkaji akan melaksanakan teknik analisis sentimen ini bagi mengkaji emosi pengguna pelantar media sosial seperti *Twitter* yang dikumpulkan dalam bentuk *tweets* data. *Tweets* data ini akan dianalisis sentimen untuk memahami perasaan rakyat terhadap parti-politik di negara Malaysia dengan hasil keputusan analisis sentimen yang diperoleh.



## 1. PENGENALAN

Pilihan raya merupakan proses mengundi wakil rakyat yang berlayak dalam sesebuah negara demokrasi. Bagi sesebauh negara demokrasi, setiap rakyat berhak untuk mengundi calon-calon yang berpengaruhi dan berkemampuan untuk memimpin negara. Sebagai negara demokrasi berparlimen, Malaysia juga akan melaksanakan proses Pilihan Raya Umum (PRU) selepas pembubaran Parlimen atau Dewan Undangan Negari(DUN) selepas tempoh 5 tahun. Pilihan Raya Umum Malaysia yang terkini merupakan Pilihan Raya

Umum ke-14 (PRU-14) yang dilaksanakan pada tahun 2018. Dalam pilihanraya tersebut, gabungan parti politik Pakatan Harapan yang diketuai oleh Tun Dr. Mahathir bin Mohamad dari parti politik *Malaysian United Indigenous Party* (BERSATU) pada masa itu. Sekarang, kerajaan Malaysia adalah diperintah oleh gabungan parti politik Barisan Nasional yang diketuai oleh parti *United Malays National Organisation* (UMNO).

Analisis sentimen merupakan teknik untuk menganalisis sentimen pengguna-pengguna dalam ayat. Teknik ini membolehkan pengkaji untuk mengkaji emosi penggunanya, contohnya perasaan gembira, marah dan netral. Teknik ini juga boleh dikategorikan sebagai salah satu bidang daripada *Natural-Language-processing* (*NLP*) yang merupakan salah satu bidang pintar sains data.

## 2. PENYATAAN MASALAH

Kini, masyarakat cenderung untuk menyampirkan pandangan mereka melalui media sosial. Menurut Berita Harian, Malaysia mempunyai sebanyak 28 juta pengguna media sosial yang telah direkodkan. Ini mempamerkan sebanyak 86 peratus pengguna daripada rakyat negara kita yang memiliki akaun media sosial sendiri. (Berita Harian 2021). Terdapat pelbagai media sosial yang dipilih, antaranya termasuk *Facebook, Twitter, Instagram* dan sebagainya. Semua pelantar media sosial ini telah merekodkan sumber data yang kaya bagi penyelidik untuk menjalankan penganalisisan data dalam pelbagai bidang.

Antaranya satu bidang yang amat bermakna bagi anggota masyarakat membincang dalam media sosial merupakan isu politik. Rakyat sering mempamerkan pandangan-pandangan mereka dalam pelantar media sosial terhadap sesuatu topik. Hal ini demikian kerana, peranti pintar yang canggih dan perisian media sosial yang senang digunakan telah membawa kemudahan yang banyak kepada pengguna internet. Bagi isu politik, mereka akan meninggalkan komen terhadap parti-parti politik atau calon mengikut kegemaran atau keutamanya mereka sendiri. Mereka juga gemar untuk *like, follow* atau *retweet tweets* daripada parti politik atau calon pengundi yang mereka suka untuk menyampaikan

sokongan mereka terhadap parti politik tersebut. Bagi parti politik yang mereka benci, mereka juga boleh meninggalkan komen buruk terhadap parti politik tersebut.

Hasil pilihan raya di negara kita masih tidak pernah dianalisis oleh pengkaji negara kita melalui teknik algorithma analisis sentimen secara mendalam. Semua respon dari media sosial ini merupakan sumber data yang amat penting bagi kita untuk mengkaji emosi rakyat dan pengaruhnya terhadap keputusan pilihan raya di Malaysia.

## 3. OBJEKTIF KAJIAN

Projek ini bertujuan merekodkan komen dan sentimen rakyat pada media sosial berasaskan pilihan raya di Malaysia, mengkaji model yang sesuai bagi menganalisis sentimen rakyat terhadap pilihan raya dan menganalisis sentimen rakyat yang terwujud di media sosial mengenai pilihan raya di Malaysia.

## 4. METOD KAJIAN

Kajian ini akan menggunakan teknik analisis sentimen untuk menguasai emosi dalam kadangan pengguna media sosial. Bagi mendapat informasi daripada media sosial, alat pengikisan data boleh dipakai. Terdapat banyak alat pengikisan data yang tersedia di internet secara percuma. Misalnya, alat *Tweepy*, *Twint*, *Twitter Scraper* dan lain-lain. Pengkaji boleh menperoleh *tweets* yang khusus dengan menggunakan kata kunci tertentu dan tanda pagar atau *hashtag* mengenai parti politik di Malaysia.

### 4.1. Fasa Pengumpulan Data

Fasa pengumpulan data merupakan langkah pertama bagi kajian analisis sentimen ini. Pengkaji memerlukan satu *Twitter Developer Account* untuk memulakan pengambilan data dari laman web Twitter. *Twitter Developer Account* adalah sejenis akaun *Twitter* khas yang

berhak mengakseskan data pengguna Twitter dengan *API*-nya. Selepas akaun *Twitter Developer* sudah disediakan, proses pengumpulan data juga boleh bermula. Pengkaji akan menggunakan akaun *Twitter* dan kunci API Twitter untuk mendapat akses bagi mengumpulkan Tweets data.

Pengkaji akan menetapkan spesifikasi *Tweets* semasa proses pengumpulan data. Hanya Tweets yang berkaitkan dengan topik pilihanraya di Malaysia akan dipilih sebagai sumber data untuk analisis. Selain itu, pengkaji juga akan menspesifikasikan tempoh tarikh yang tertentu bagi memastikan Tweets yang diperoleh adalah dipamerkan semasa tempoh tarikh sempana pilihanraya di negara kita. Seluruh *Tweets* data yang dikumpul akan dimasukkan dalam fail CSV.

### 4.2. Fasa Pra-Pemprosesan Data

Fasa pra-pemprosesan data merupakan fasa yang dijalankan selepas fasa pengumpulan data. Lazimnya, fasa pra-pemprosesan data merupakan proses yang paling mengambil masa dalam projek data sains. *Tweets* data yang telah dikumpul merupakan data yang belum diproseskan. Data begini mengandungi banyak perkataan yang tidak bermakna dan tidak bantu pengkaji menganalisis sentimen yang terkandung dalam ayat tersebut. Data yang belum diproses juga akan mengandungi segelintir data yang berulang. Oleh demikian, fasa pra-pemprosesan data merupakan fasa yang amat penting bagi setiap tugasan analisis data.

### 4.3. Fasa Penganalisisan Data

Fasa penganalisian data merupakan fasa yang bertujuan untuk menganalisis data Tweets dengan pelbagai pakej perpustakaan berasaskan *lexicon*. Pakej perpuskataan yang digunakan adalah *VADER* dan *TextBlob*.

Selepas fasa pra-pemprosesan data, pengkaji akan mengimportksn Fail CSV yang dipra-pemprosesan ke dalam kod Fasa penganalisisan data ini. Fail CSV tersebut akan dianalisiskan dengan pakej perpustakaan *VADER* dahulu sebelum diproseskan ke pakej

perpustakaan TextBlob. Keputusan analisis sentiment untuk kedua-dua pakej perpustakaan ini adalah disimpan sebagai Fail CSV asing untuk mengelakkan kekeliruan pengkaji.

### 4.4. Fasa Pembelajaran Mesin

Untuk fasa pembelajaran mesin, pengkaji akan membina model pembelajaran mesin untuk mengkaji set data yang dianalisis pada fasa penganalisisan data tadi. Sebelum permulaan fasa ini, pengkaji sepatutnya menyediakan *Tweets* yang telah dianalisis melalui pakej *VADER* atau *TextBlob* dengan sentimen positif, negatif atau neutral. Daripada keputusan kedua pakej librari ini, pengkaji akan memilih hasil analisis yang mengandungi sentiment neutral yang lebih rendah sebagai set data untuk meneruskan kajian projek ini. Pada fasa pembelajaran mesin, terdapat tiga algorithma pembelajran mesin yang diadakan untuk menganalisis emosi dalam ayat Tweets. Algorithmanya merupakan *Naïve Bayes*, *Support Vector Machine* dan *Random Forest*.

### 4.5. Fasa Visualisasi Data

Fasa visualisasi akan menghasilkan keputusan yang didapati dari fasa penganalisian data. Antaranya hasil ini termasuk skor polaritas dari setiap jenis model analisis. Semua hasil keputusan akan dipamerkan dalam bentuk angka dan alat visualisasi seperti graf, carta dan lain-lain. Pengkaji juga akan membandingkan keputusan yang diperoleh melalui setiap jenis model yang berbeza untuk berbanding kesesuaian setiap jenis model. Akhirnya, pengkaji juga akan menerangkan kesimpulan keputusan analisis pilihan raya.

## 5. HASIL KAJIAN

Bab ini bertujuan untuk menerangkan proses implementasi untuk kajian ini. Untuk setiap fasa yang terlibat dalam kajian ini, hasil keputusan yang tertentu akan dihasilkan. Keputusan yang dihasil dari setiap fasa juga akan ditunjukkan dalam bab ini.

Pilihan raya merupakan proses mengundi wakil rakyat yang berlayak dalam sesebuah negara demokrasi. Bagi sesebauh negara demokrasi, setiap rakyat berhak untuk mengundi calon-calon yang berpengaruhi dan berkemampuan untuk memimpin negara. Sebagai negara demokrasi berparlimen, Malaysia juga akan melaksanakan proses Pilihan Raya Umum (PRU) selepas pembubaran Parlimen atau Dewan Undangan Negari(DUN) selepas tempoh 5 tahun. Pilihan Raya Umum Malaysia yang terkini merupakan Pilihan Raya Umum ke-14 (PRU-14) yang dilaksanakan pada tahun 2018.

Analisis sentimen merupakan teknik untuk menganalisis sentimen pengguna-pengguna dalam ayat. Teknik ini membolehkan pengkaji untuk mengkaji emosi penggunanya, contohnya perasaan gembira, sedih, marah dan neutral. Teknik ini juga boleh dikategorikan sebagai salah satu bidang daripada *Natural-Language-preocessing*(NLP) yang merupakan salah satu bidang pintar sains data.

Untuk fasa pengumpulan data, hasil yang didapati adalah set data yang mengandungi *Tweets* data berbincangkan pelbagai parti politik di negara Malaysia. Untuk fasa pra-pemprosesan data, semua set data yang dikumpul dalam fasa sebelum ini akan diggunakan dan menjalani proses pra-pemprosesan data. Set data yang dipra-pemproseskan juga akan dieksport bagi persediaan fasa yang seterusnya.

Pengkaji menggunakan *Twitter* API yang bertahap *Elevated* sahaja. Twitter API telah diwujudkan dalam pelbagai tahap, seperti *Essential*, *Elevated* dan *Academic Research*. Setiap tahap mempunyai limitasinya sendiri. Bagi *Twitter* API bertahap *Elevated* yang pengkaji guna ini, ia hanya dapat mengakses data yang dipaparkan di plakform *Twitter* selama seminggu sebelum masa pengkaji menggunakannya sahaja. Oleh itu, pengkaji hanya dapat mengumpulkan data secara mingguan. Dengan demikian, set data yang telah pengkaji kumpulkan juga akan diwujudkan secara banyak dan tabur.

Pengkaji bermula proses pengumpulan data dari 22 Mei 2022 sampai 25 June 2022. Oleh itu, set data yang dikumpul adalah amat tabur kerana set data adalah terdiri daripada Tweets data setiap minggu. Oleh yang demikian, selepas pengkaji mengumpul semua set data yang berkaitan, pengkaji juga telah mencipta kod fungsi yang tertentu untuk menggabungkan seluruh set data berkaitkan sesuatu parti politik.

Sebagai contohnya, pengkaji menyimpan *tweets* data berkiatan parti politik UMNO yang dikumpul pada minggu pertama dengan namanya : GE15 BN - UMNO (22.5.22 1528).csv. Kemudian, pengkaji menyimpan *tweets* data berkiatan parti politik UMNO untuk minggu kedua proses pengumpulan data dengan namanya : GE15 BN - UMNO (31.5.22 1159).csv. Angka-angka yang direkodkan dalam kolum merupakan tarikh dan masa pengkaji menggumpul data tersebut. Sekiranya semua data disimpan lepas beberapa minggu, pengkaji akan menggabungkan semua set datanya dan menghasilkan sesuatu set data yang baru bernama: GE15 BN - UMNO. Final Dataset.csv.

Dalam fasa penganalisisan kajian ini yang seterusnya, pengkaji akan menggunakan librari *Vader* dan *TextBlob* untuk menganalisiskan set data yang dipra-pemproseskan. Bagi set data yang dianalisiskan menggunakan pakej *VADER*, lajur 'Positive', 'Negative', 'Neutral', 'Compound', dan 'Sentiment' akan direkodkan. Bagi set data yang dianalisiskan menggunakan pakej Textblob, lajur yang direkodkan adalah 'Subjectivity', 'Polarity', dan 'Sentiment'.

Tidak mengira pakej librari yang mana telah diaplikasikan, keputusan yang paling penting adalah elemen *Sentiment* untuk setiap tweets data. *Sentiment* ini juga merupakan elemen penting yang membolehkan pengkaji memahami emosi-emosi yang disampaikan dalam ayat Tweets pengguna media sosial.

Set data yang dikategorikan mengikut parti politik akan menjalankan proses penganalisisan ini. Oleh yang demikian, set data mengandungi keputusan analisis dari Vader dan Textblob akan dihasilkan. Semua set data yang dianalisiskan kemudian akan diexport sebagai set data dalam bentuk fail CSV yang asing,

## 5.1. PROSES PENGUMPULAN DATA

Dalam proses pembangunan projek ini, tweets yang mewakili sentimen pelbagai parti politik telah dikumpulkan pada proses awal pembangunan kajian.

Terdapat 3 gabungan parti politik yang utama di negara Malaysia, iaitu Barisan Nasional(BN), Pakatan Harapan(PH) dan Perikatan Nasional(PN). Selain 3 gabungan

parti politik ini, terdapat juga parti politik yang tidak masuk dalam ketiga-tiga gabungan parti politik ini. Antaranya seperti parti MUDA, parti Pejuang, parti PBM dan lain-lain.

*Tweets* data dari setiap parti politik dari gabungan BN, PH dan PN telah dikumpulkan pada fasa ini. Tweets data dari parti politik yang tidak termasuk dalam 3 gabungan parti politik ini juga dikumpul. Menurut Bab III kajian, pengkaji telah menyediakan pelbagai kata kunci yang berkiatan dengan setiap parti politik. Semua kata kunci yang disediakan telah disimpan sebagai fail csv bagi memudahkan proses pembacaan oleh kod *Python* dalam *Google Colab*.

Jadual-jadual yang berikut menunjukkan jumlah kata kunci dan contoh-contoh kata kunci yang digunakan untuk mengumpul Tweets dari pelbagai gabungan parti politik.

| **Barisan Nasional (BN)** | | |
|---|---|---|
| **Parti Politik** | **Contoh Kata Kunci** | **Jumlah Kata Kunci** |
| **UMNO** | #UMNOBangkit, #UMNOOnline, NajibRazak, Bossku, #ZahidHamidi, Khairy Jamaluddin, … | **33** |
| **MCA** | Malaysian Chinese Association, Persatuan Cina Malaysia, Wee Ka Siong, Mah Hang Soon, … | **5** |
| **MIC** | Malaysian Indian Congress, Kongres India Malaysia, #MICMemimpin, #MICPrihatin, … | **7** |
| **PBRS** | Parti Bersatu Rakyat Sabah, Joseph Kurup | **2** |
| **Kata kunci secara umum** | Undi BN, vote BN Malaysia, Barisan Nasional, Kestabilan Untuk Kemakmuran, … | **7** |

Jadual 5.1: Barisan Nasional (BN): *United Malays National Organisation*(UMNO), *Malaysian Chinese Association*(MCA), *Malaysian Indian Congress*(MIC) & *Parti Bersatu Rakyat Sabah*(PBRS).

| **Pakatan Harapan (PH)** | | |
|---|---|---|
| **Parti Politik** | **Contoh Kata Kunci** | **Jumlah Kata Kunci** |
| **PKR** | PKR GE15, Undi PKR, Anwar Ibrahim, Parti Keadilan Rakyat, Wan Azizah, … | **28** |

| | | |
|---|---|---|
| **DAP** | #DAPMalaysia, Democratic Action Party, Lim Guan Eng, Lim Kit Siang, Loke Siew Fook, Gobind Singh Deo, … | **30** |
| **Amanah** | Parti Amanah Negara, National Trust Party, Mohamad Sabu, Salahuddin Ayub, … | **16** |
| **UPKO** | #UPKO, United Progressive Kinabalu Organisation, #UPKOMalaysia, #UPKOPenampang, #DonMojuntin, Wilfred Madius Tangau, … | **8** |
| **Kata kunci secara umum** | Pakatan Harapan, #MajuBersamaHarapan, Undi PH, PakatanHarapan | **4** |

Jadual 5.2: Pakatan Harapan (PH): *People's Justice Party* (PKR), *Democratic Action Party* (DAP), *Parti Amanah Negara* (Amanah) & *United Progressive Kinabalu Organisation* (UPKO).

| **Perikatan Nasional (PN)** | | |
|---|---|---|
| **Parti Politik** | **Contoh Kata Kunci** | **Jumlah Kata Kunci** |
| **BERSATU** | BERSATU, PPBM, Parti Pribumi Bersatu Malaysia, Muhyiddin Yassin, TSMY, Ahmad Faizal, … | **24** |
| **PAS** | PAS, Parti Islam Se Malaysia, Malaysian Islamic Party, Abdul Hadi Awang, TGHH, Tuan Ibrahim Tuan Man, Idris Ahmad, … | **18** |
| **GERAKAN** | Parti Gerakan Rakyat Malaysia, PGRM, Lau Hoe Chai, Tan Kok Hong, Yong Fui Ling, … | **7** |
| **STAR** | Homeland Solidarity Party, Parti Solidariti Tanah Airku, Jeffrey Kitingan | **3** |
| **SAPP** | Sabah Progressive Party, Parti Progresif Sabah, Yong Teck Lee, Liew Teck Chan, … | **6** |
| **Kata kunci secara umum** | Perikatan Nasional, Undi PN, PN Malaysia, PerikatanNasional | **4** |

Jadual 5.3: Perikatan Nasional (PN): *Parti Pribumi Bersatu Malaysia* (BERSATU), *Parti Islam Se-Malaysia* (PAS), *Parti Gerakan Rakyat Malaysia* (GERAKAN), *Parti Solidariti Tanah Airku* (STAR) & *Sabah Progressive Party*(SAPP).

| **Parti-parti politik yang lain - lain** |
|---|

| Parti Politik | Contoh Kata Kunci | Jumlah Kata Kunci |
|---|---|---|
| **MUDA** | undi muda, #MUDAsudahMULA, Syed Saddiq, #MUDASelangor, Dr. Thanussha, … | **13** |
| **PEJUANG** | Parti Pejuang Tanah Air, undi PEJUANG, Mahathir, Marzuki Yahya, Amiruddin Hamzah, … | **12** |
| **WARISAN** | Parti Warisan Sabah, Parti Warisan, Shafie Apdal, Ignatius Darell Leiking, … | **8** |
| **PBM** | Parti Bangsa Malaysia, Malaysian Nation Party, Zuraida Kamaruddin, Larry Sng, … | **8** |
| **PSB** | Parti Sarawak Bersatu, #partisarawakbersatu, Wong Soon Koh, … | **4** |
| **PBS** | United Sabah Party, Parti Bersatu Sabah, Undi PBS, Maximus Ongkili, Joniston Bangkuai, … | **6** |

Jadual 5.4: Parti politik yang lain - lain: *Malaysian United Democratic Alliance* (MUDA), *Parti Pejuang Tanah Air* (PEJUANG), *Parti Warisan Sabah* (WARISAN), *Parti Bangsa Malaysia* (PBM), *Parti Bersatu Sabah* (PBS), *Parti Sarawak Bersatu* (PSB).

Pengaplikasian kata kunci yang banyak melancarkan proses pengumpulan data ini. Ia memastikan pengkaji memperoleh data yang cukup dan seimbang untuk proses analisis sentimen yang seterusnya. Di samping itu, pengkaji juga menggunakan kata-kunci secara umum dalam proses penggumpulan data. Kata kunci secara umum merupakan kata kunci yang tidak terlibat dengan mana-mana parti politik, tetapi berkiatan dengan gabungan parti politik tersebut sahaja. Contohnya kata kunci secara umum untuk gabungan Barisan Nasional(BN) adalah slogan BN, perkataan '*Vote* BN', 'undi BN' dan lain-lain. Manakala *tweets* yang dikumpul dengan kata kunci secara umum ini tidak mempunyai kaitan secara langsung kepada anggota parti politik gabungannya seperti UMNO dan MCA.

Selepas pengaplikasian semua kata kunci yang disediakan, pengkaji mengumpulkan *tweets* data dari setiap gabungan parti politik. Secara keseluruhannya, terdapat banyak *tweets* dikumpul untuk kajian ini. Terdapat beberapa parti politik yang kurang popular didapati bilangan *tweets* yang amat rendah, terdapat juga parti politik yang popular mendapati bilangan *tweets* yang banyak. Jadual 5.5 yang berikut

menunjukkan jumlah *tweets* yang dikumpul untuk setiap parti politik dengan penggunaan kata kunci dalam proses pengumpulan data.

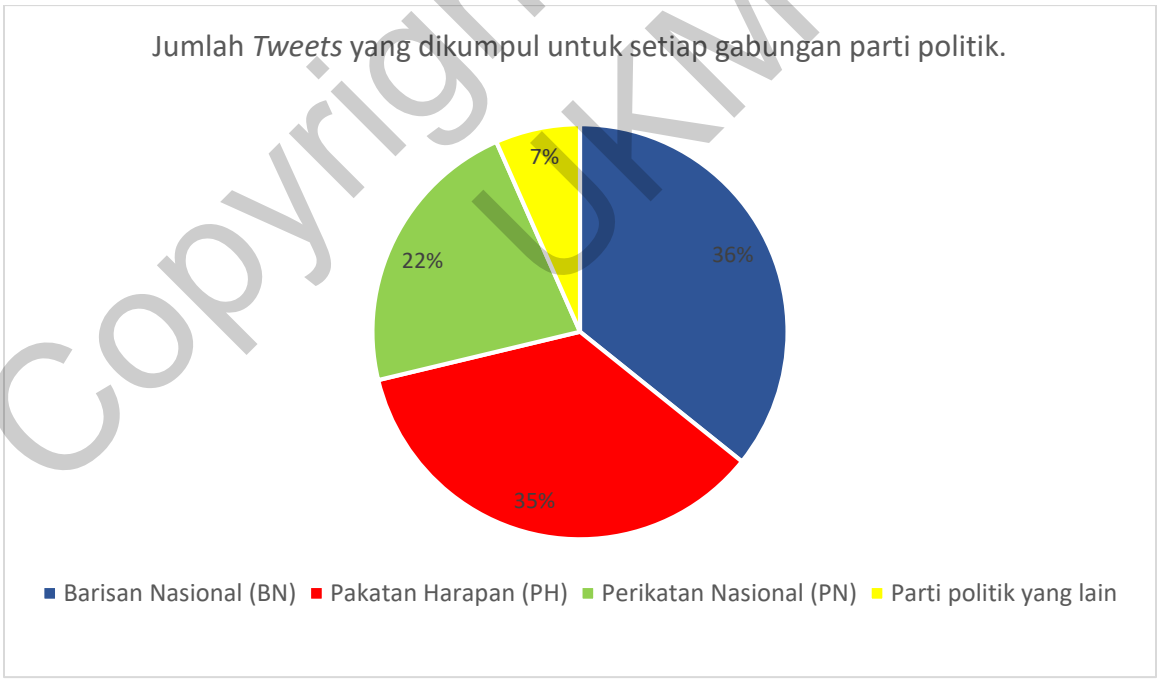
| *Parti Politik/ Gabungan Parti Politik* | *Tweets yang dikumpul* |
|---|---|
| ***Barisan Nasional (BN)*** | |
| *United Malays National Organisation (UMNO)* | **7681** |
| *Malaysian Chinese Association (MCA)* | **104** |
| *Malaysian Indian Congress (MIC)* | **18** |
| *Parti Bersatu Rakyat Sabah (PBRS)* | **2** |
| *Kata kunci secara umum* | **2122** |
| ***Pakatan Harapan (PH)*** | |
| People's Justice Party (PKR) | **4394** |
| Democratic Action Party (DAP) | **4347** |
| Parti Amanah Negara (Amanah) | **636** |
| United Progressive Kinabalu Organisation (UPKO) | **8** |
| Kata kunci secara umum | **473** |
| ***Perikatan Nasional (PN)*** | |
| *Parti Pribumi Bersatu Malaysia (BERSATU)* | **1729** |
| *Parti Islam Se-Malaysia (PAS)* | **4294** |
| *Parti Gerakan Rakyat Malaysia (GERAKAN)* | **29** |
| *Sabah Progressive Party (SAPP)* | **15** |
| *Parti Solidariti Tanah Airku (STAR)* | **4** |
| *Kata kunci secara umum* | **76** |
| ***Parti-parti Politik yang lain*** | |
| *Malaysian United Democratic Alliance (MUDA)* | **473** |
| *Parti Pejuang Tanah Air (Pejuang)* | **1122** |
| *Parti Bangsa Malaysia (PBM)* | **206** |
| *Parti Warisan Sabah (WARISAN)* | **27** |
| *Parti Bersatu Sabah (PSB)* | **4** |
| *Parti Sarawak Bersatu (PBS)* | **2** |

Jadual 5.5: Tweets yang dikumpul untuk setiap gabungan parti politik dan parti-parti politik yang lain.

| Gabungan Parti Politik | Jumlah Tweets yang dikumpul |
|---|---|
| *Barisan Nasional (BN)* | **9927** |
| *Pakatan Harapan (PH)* | **9858** |
| *Perikatan Nasional (PN)* | **6147** |
| *Parti politik yang lain* | **1834** |

Jadual 5.6: Jumlah *Tweets* yang dikumpul untuk setiap gabungan parti politik.

Berdasarkan Jadual di atas, Tweets data yang paling banyak dikumpulkan adalah dari Barisan Nasional(BN), iaitu sebanyak 9927 *tweets*. Jumlah tweets data yang dikumpulkan untuk parti-parti politik yang lain adalah paling kurang, iaitu sebanyak 1834 *tweets* sahaja. Selain dari Barisan Nasional, Pakatan Harapan juga merupakan gabungan parti politik yang amat dibincangkan dan popular di kalangan masyarakat. Ia telah mendapat sebanyak 9858 *tweets*, iaitu 69 *tweets* lebih rendah daripada gabungan Barisan Nasional sahaja.




Rajah 5.1: Carta Pai terhadap jumlah *Tweets* yang dikumpul untuk setiap gabungan parti politik

Secara keseluruhannya, sebanyak 27766 *tweets* data telah dikumpul dalam kajian ini. Bilangan *Tweets* yang dikumpul ini boleh dianggap sebagai satu jumlah yang besar untuk menjalankan tugas analisis sentimen ini.

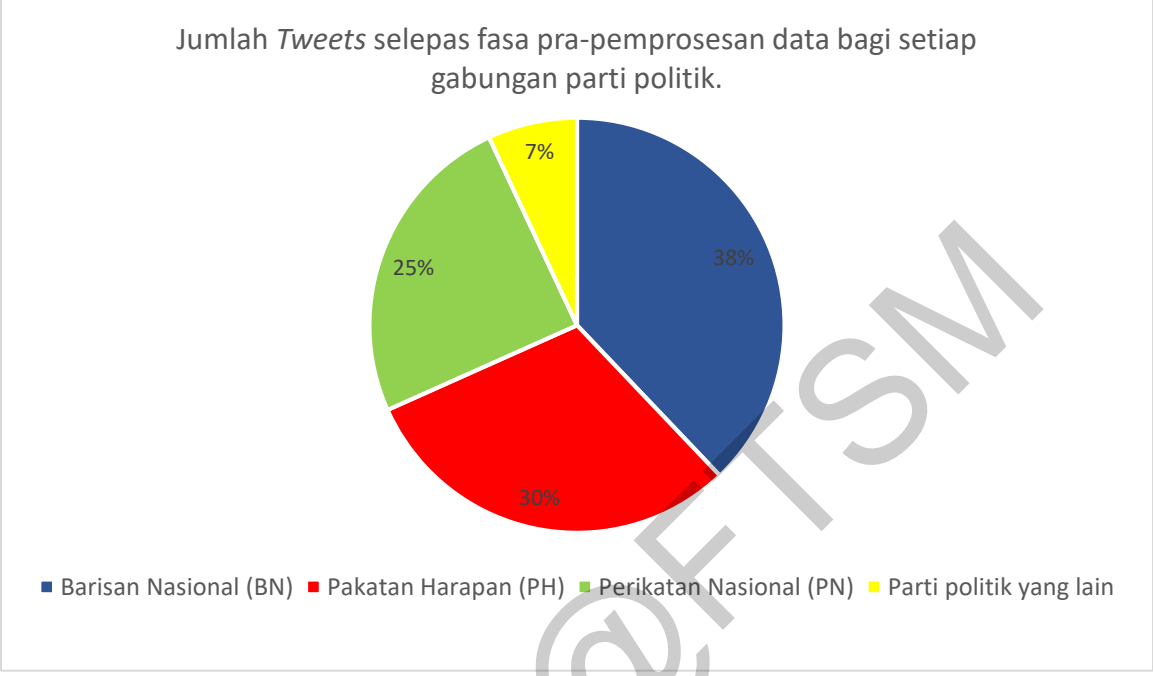
## 5.2. PROSES PRA-PEMPROSESAN DATA

Selepas semua dataset dikumpul, pengkaji juga meneruskan kajian dengan mengjalankan tugas pra-pemprosesan data pada set data yang dikumpul. Jumlah *tweets* kajian akan dikurangkan selepas fasa pra-pemprosesan. Hal ini demikian kerana, fasa pra-pemprosesan ini terlibat dengan prosedur penghapusan *tweets* data yang salinan dalam set data.

Jadual dan raja di berikut menunjukkan gabungan parti politik dan *Tweets* yang tertinggal selepas fasa pra-pemprosesan. Jelasnya bahawa, bilangan *tweets* data selepas pra-prmprosesan adalah lebih rendah berbanding dengan *tweets* data yang baru dikumpul. Bilangan *tweets* data untuk setiap gabungan parti telah dikurangkan kerana prosedur penghapusan *tweets* data yang salinan dalam fasa pra-pemprosesan data tersebut.

| Gabungan Parti Politik | Tweets data (Selepas pra-pemprosesan) |
|---|---|
| *Barisan Nasional (BN)* | **9170** |
| *Pakatan Harapan (PH)* | **7356** |
| *Perikatan Nasional (PN)* | **5980** |
| *Parti politik yang lain* | **1686** |

Jadual 5.7: Jumlah Tweets data selepas fasa pra-pemprosesan untuk setiap gabungan parti politik.




Rajah 5.2: Carta Pai terhadap jumlah *Tweets* selepas fasa pra-pemprosesan data bagi setiap gabungan parti politik

Secara keseluruhannya, terdapat 24192 *tweets* data yang kekal selepas fasa pra-pemprosesan. Semua *tweets* data yang salianan telah dihapuskan untuk menjamin kredibiliti keputusan analisis sentimen kajian ini. Jelasnya, terdapat sebanyak 3574 *tweets* data yang dihapuskan. Bilangan tweets yang dihapuskan adalah ramai, ia merupakan sebanyak 12.87% daripada *tweets* data yang asal. Antaranya punca yang menyebabkan keadaan ini adalah kewujudan pengguna *Twitter* yang suka menyampaikan mesej-mesej secara spam dan berulang dalam pelantar *Twitter*. Semua tweets yang dipapar secara bentuk spam ini juga telah dikumpulkan pada fasa pengumpulan data, tetapi akan dihapuskan semasa fasa pra-pemprosesan data.

Dalam fasa ini, semua tweets data yang sudah menjalankan proses pra-prmprosesan akan dieksport sebagai set data yang baru yang akan digunakan dalam fasa selepas ini. Pengkaji menggunakan fungsi *.to_csv* untuk mengeksportkan set data tersebut sebagai hasil keputusan fasa ini dalam bentuk fail CSV.

## 5.3. PROSES PENGANALISISAN SENTIMEN

Selepas ini, pengkaji akan mengadakan proses penganalisisan data. Terdapat dua pakej perpustakaan akan diaplikasikan untuk setiap set data yang dipra-pemprosesan, iaitu *VADER* dan *TextBlob*. Tujuan bagi kedua-dua pakej perpustakaan ini adalah untuk mengira *Sentiment* atau emosi yang terkandung dalam *tweets* data. Namun demikian, *VADER* dan *Textblob* memakai konsep API yang tidak sama. *VADER* menggunakan markah kompound manakala *TextBlob* mengaplikasikan markah *subjectivity* dan *polarity*. Jadi, kedua-dua pakej perpustakaan ini juga mungkin akan menghasilkan hasil analisis yang berbeza.

Selepas proses penganalisisan data ini, semua set data yang dianalisis dengan pakej *Vader* atau *Textblob* juga telah dieksportkan. Keputusan analisis yang didapati akan dieksportkan dengan fungsi *to_csv* sama dengan fasa sebelum ini bagi kegunaan fasa pembelajaran mesin. Keputusan analisis untuk setiap gabungan parti politik dalam kajian ini adalah ditunjukkan seperti berikut.

**Analisis Sentimen melalui pakej librari VADER**

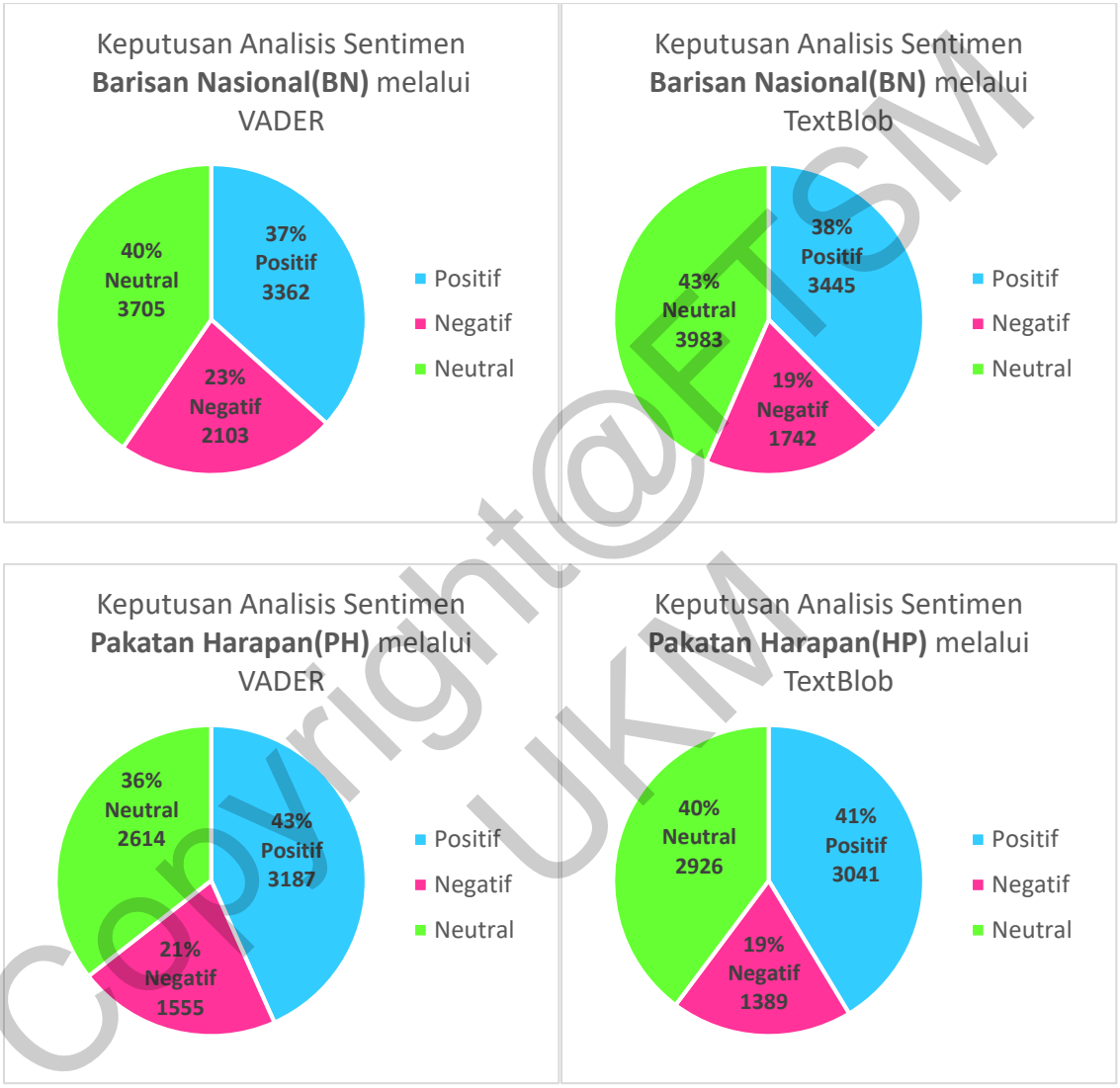
| **Gabungan Parti Politik** | Sentimen / Peratus | | | | | | Jumlah |
|---|---|---|---|---|---|---|---|
| | Positif | % | Negatif | % | Neutral | % | |
| Barisan Nasional (BN) | 3362 | 36.6 | 2103 | 22.9 | 3705 | 40.4 | 9170 |
| Pakatan Harapan (PH) | 3187 | 43.3 | 1555 | 21.1 | 2614 | 35.5 | 7356 |
| Perikatan Nasional (PN) | 1909 | 31.9 | 1096 | 18.3 | 2975 | 49.7 | 5980 |
| Parti Politik yang lain | 563 | 33.4 | 422 | 25.0 | 701 | 41.6 | 1686 |

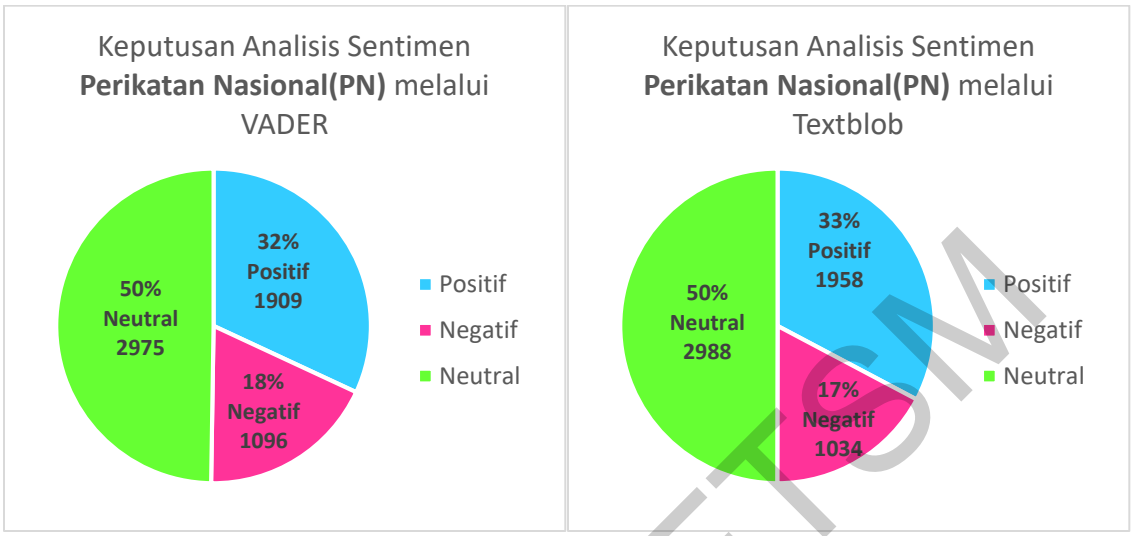
Jadual 5.8: Keputusan analisis sentimen terhadap setiap gabungan parti politik melalui pakej perpustakaan *VADER*.

**Analisis Sentimen melalui pakej librari TextBlob**

| **Gabungan Parti Politik** | Sentimen / Peratus | | | | | | Jumlah |
|---|---|---|---|---|---|---|---|
| | Positif | % | Negatif | % | Neutral | % | |
| Barisan Nasional (BN) | 3445 | 37.6 | 1742 | 19.0 | 3983 | 43.4 | 9170 |
| Pakatan Harapan (PH) | 3041 | 41.3 | 1389 | 18.9 | 2926 | 39.8 | 7356 |
| Perikatan Nasional (PN) | 1958 | 32.7 | 1034 | 17.3 | 2988 | 50.0 | 5980 |
| Parti Politik yang lain | 581 | 34.5 | 319 | 18.9 | 786 | 46.6 | 1686 |

Jadual 5.9: Keputusan analisis sentimen terhadap setiap gabungan parti politik melalui pakej perpustakaan *TextBlob*.






Rajah 5.3, Rajah 5.4, Rajah 5.5, Rajah 5.6, Rajah 5.7 & Rajah 5.8: Carta pai yang menunjukkan keputusan analisis sentimen terhadap setiap gabungan parti politik melalui *VADER* atau *TextBlob*.

Untuk kajian analisis sentimen, sentimen positif dan negatif adalah unsur elemen yang paling penting untuk memahami perasaan dan emosi pengguna sosia media. Sentimen neutral merupakan elemen yang kurang berharga untuk mendapati perasaan pengguna media sosial. Oleh itu, pengkaji akan memilih keputusan yang mengandungi lebih kurang sentimen neutral sahaja, daripada keputusan analisis pakej librari *VADER* atau *Textblob* yang diperoleh. Sebagai contohnya, keputusan analisis bagi BN dari pakej *Textblob* mempunyai 3983 tweets data (43% neutral), manakala keputusannya dari pakej *VADER* mempunyai sebanyak 3705 tweets data sahaja (40% neutral). Oleh yang demikian, pengkaji akan menggunakan hasil keputusan dari pakej *VADER* untuk gabungan parti BN tersebut.

Dengan demikian, pengkaji haruslah membuat pemilihan keputusan dari pakej perpustakaan, sama ada *VADER* atau *TextBlob*, yang mana mempunyai bilangan sentimen neutral yang paling rendah. Pengkaji boleh membuatkan konklusi keputusan analisis sentimen seperti jadual di bawah.

| Gabungan Parti Politik | Sentimen | | | Pakej librari yang dipilh (VADER/ TextBlob) |
|---|---|---|---|---|
| | Positif | Negatif | Neutral | |
| Barisan Nasional (BN) | 3362 | 2103 | 3705 | *VADER* |
| Pakatan Harapan (PH) | 3187 | 1555 | 2614 | *VADER* |
| Perikatan Nasional (PN) | 1909 | 1096 | 2975 | *VADER* |
| Parti Politik yang lain | 563 | 422 | 701 | *VADER* |

Jadual 5.9: Keputusan analisis sentimen terhadap setiap gabungan parti politik yang dipilih (pemilihan mengikut bilangan sentimen neutral yang paling rendah).

## 5.4. PROSES PEMBELAJARAN MESIN

Terdapat 3 algorithma pembelajaran mesin yang diaplikasikan dalam kajian ini, iaitu *Naïve Bayes*, *Support Vector Machine(SVM),* dan *Random Forest*. Tujuan pengaplikasian pembelajaran mesin ini adalah untuk mendapati prestasi terhadap keputusan penganalisisan sentimen. Jadual 5.10 di berikut menunjukkan ketepatan model pembelajaran mesin terhadap keputusan analisis sentimen dari Barisan Nasional, Perikatan Harapan dan Perikatan Nasional.

| Gabungan Parti Politik | Pembelajaran Mesin | Ketepatan(%) |
|---|---|---|
| **Barisan Nasional (BN)** | *Naïve Bayes* | 64.96 |
| | *Support Vector Machine* | 78.00 |
| | *Random Forest* | 79.86 |
| **Pakatan Harapan (PH)** | *Naïve Bayes* | 56.23 |
| | *Support Vector Machine* | 68.83 |
| | *Random Forest* | 66.24 |
| **Perikatan Nasional (PN)** | *Naïve Bayes* | 62.99 |
| | *Support Vector Machine* | 71.18 |
| | *Random Forest* | 70.29 |

Jadual 5.10: Ketepatan model pembelajaran mesin terhadap keputusan analisis sentimen dari gabungan parti politik.

Dari jadual di atas, jelasnya bahawa keputusan analisis dari ketiga-tiga gabungan parti politik ini mendapat ketepatan yang tinggi untuk tiga modal pembelajaran mesin ini.

Untuk Barisan Nasional, ketepatan dari algorithma *Random Forest* adalah yang paling tinggi, iaitu ketepatan sebantak 79.86%. Untuk Pakatan Harapan, ketepatan yang tertinggi adalah diwujudkan dari algorithma *Support Vector Machine(SVM)*, iaitu 68.83%. Bagi Perikatan Nasional, ketepatan yang paling tinggi juga ditunjukkan oleh *Support Vector Machine(SVM)*, iaitu sebanyak 71.18%.

Dalam bahagian ini, ketepatan bagi parti-parti politik yang lain sudah tidak dipertimbangkan. Hal ini demikian kerana, bilangan *Tweets* bagi parti-parti politik yang lain adalah amat kurang. Hanya parti politik MUDA dan PEJUANG mempunyai jumlah tweets yang agar tinggi. Namun, jumlah tweets dari kedua parti politik ini juga amat rendah dan tidak sesuai untuk dianalisiskan menggunakan algorithma pembelajaran mesin. Pengkaji juga tidak boleh menggabungkan jumlah kedua-dua parti politik ini sebab mereka tidak merupakan gabungan parti politik dan tidak memakai konsep pemerintahan yang sama.

Kewujudnya pembezaannya dalam ketepatan setiap algorithma disebabkan oleh gabungan parti politik ini diwakili dengan set data yang berbeza. Dengan demikian, keputusannya juga biasa jika wujudnya perbezaan dari segi ketepatan.

Selepas ini, laporan klasifikasi telah digunakan untuk menganalisis keputusan ketiga-tiga pembelajaran mesin. Terdapat beberapa unsur yang diwujudkan dalam hasil laporan klasifikasi ini. Antaranya adalah kejituan, dapatan dan markah F1. Contoh laporan klasifikasi untuk gabungan Barisan Nasional untuk algorithma pembelajaran mesin *Naive Bayes* adalah ditunjukkan seperti di rajah berikut.

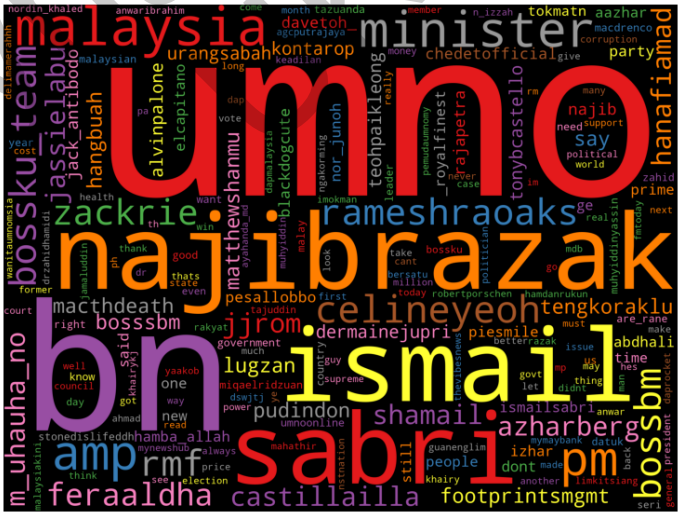
```
[17]  1 from sklearn.metrics import classification_report, confusion_matrix
      2
      3 nb_prediction = nb_model.predict(x_test)
      4 print(classification_report(y_test, nb_prediction))
```

```
                precision    recall  f1-score   support

      Negative       0.95      0.08      0.15       526
       Neutral       0.70      0.77      0.73      1192
      Positive       0.59      0.81      0.68      1033

      accuracy                           0.65      2751
     macro avg       0.75      0.55      0.52      2751
  weighted avg       0.71      0.65      0.60      2751
```



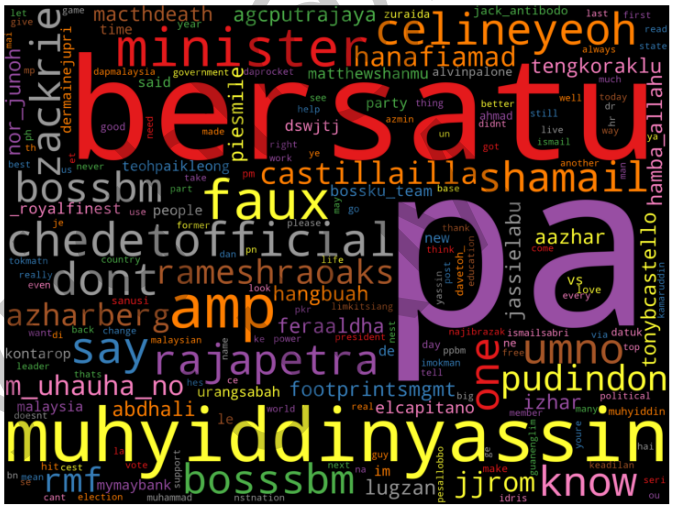
## 5.5. PROSES VISUALISASI DATA

Seterusnya, *tweets* data akan divisualisasikan. Untuk kajian ini, *tweets* data akan divisualisasikan dengan dua cara. Yang pertama, pengkaji akan menggunakan kod Google Colab untuk proses penganalisisan. *Wordcloud* akan diwujudkan untuk mempamerkan perkataan yang kerap wujud dalam set data selepas fasa pra-pemprosesan data. Ketiga-tiga rajah di bawah menunjukkan *wordcloud* berkaitan tiga gabungan parti politik yang dihasilkan dengan kod *Google Colab*.



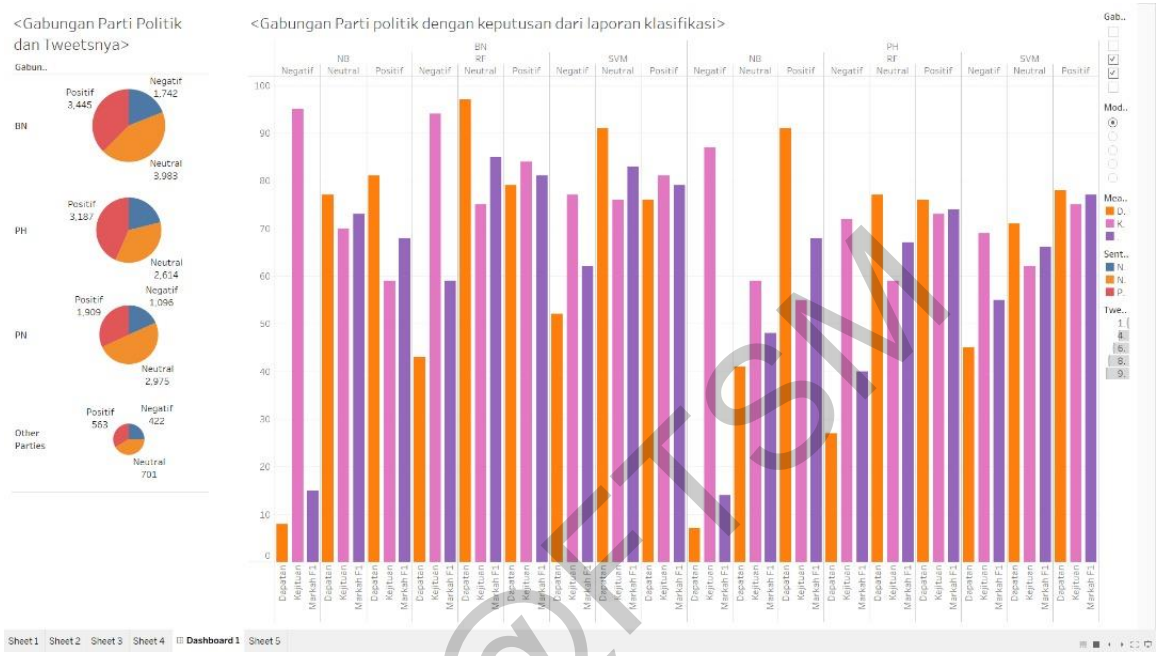

Rajah 5.9: *WordCloud* untuk gabungan parti Barisan Nasional (BN).


Rajah 5.10: *WordCloud* untuk gabungan parti Pakatan Harapan (PH).




Rajah 5.11: *WordCloud* untuk gabungan parti Perikatan Nasional (PN).

Selain itu, teknik visualisasi yang kedua adalah dengan menggunakan perisian *Tableau*. Ia adalah sejenis alat visualisasi data yang cemerlang. Hasil visualisasi *Tableau* adalah dieksportkan dalam akaun *Tableau* pengkaji. Ia juga ditunjukkan seperti papan pemuka berikut.




Rajah 5.12: Papan Pemuka *Tableau*



## 6. KESIMPULAN

Kesimpulannya, objektif kajian yang telah ditetapkan telah dicapai. Rumusannya boleh menkonklusikan bahawa gabungan parti politik Pakatan Harapan mendapat peratusan sokongan yang amat tinggi ding dengan gabungan parti politik yang lain. Sementara ini, gabungan parti politik Barisan Nasional menunjukkan sumber data pada media sosial yang paling banyak, menunjukkan gabungan ini merupakan parti yang paling kerap dibincangkan dan popular antara kalangan masyarakat. Selain itu, model pembelajaran mesin *Naïve Bayes*, *Support Vector Machine* dan *Random Forest* menunjukkan ketepatan yang tinggi dlama kajin ini. Hal ini menunjukkan kesesuaian ketiga-tiga model ini dalam analisis sentimen kajian ini.

## 7. RUJUKAN


Berita Harian. 2021. Malaysia ada 28 juta pengguna media sosial. Bernama. https://www.bharian.com.my/bisnes/teknologi/2021/09/867407/malaysia-ada-28-juta-pengguna-media-sosial [22 September 2021]

Adibah Yasmin, Afif Zuhair & Muhamad Helmy. 2016. Pengaruh Media Sosial Dan Komuniti Digital Dalam Pru-13. https://www.academia.edu/41832683/Pengaruh_Media_Sosial_and_Komuniti_Digital_dalam_PRU_13

Ankur Agrawal & Tim Hamling. 2017. Sentiment Analysis of Tweets to Gain Insights into the 2016 US Election. https://journals.library.columbia.edu/index.php/cusj/article/view/6359

Ussama Yaqub, Nitesh Sharma, and Rachit Pabreja. 2020. Location-based Sentiment Analyses and Visualization of Twitter Elcetion Data. https://www.researchgate.net/publication/340654105_Location-based_Sentiment_Analyses_and_Visualization_of_Twitter_Election_Data

Rafiqa Cahyani, Indri Sudanawati Rozas & Nita Yalina. 2020. Analisis Sentimen Pada Media Sosial Twitter Terhadap Tokoh Publik Peserta Pilpres. https://www.researchgate.net/publication/340490084_Analisis_Sentimen_Pada_Media_Sosial_Twitter_Terhadap_Tokoh_Publik_Peserta_Pilpres_2019

Tiobe Software Index. 2011. “Tiobe Programming Community Index Python”. https://www.tiobe.com/tiobe-index/

Anvar Shathik J. & Krishna Prasad K. A Literature Review on Application of Sentiment Analysis Using Machine Learning Techniques. 2020. https://www.researchgate.net/publication/343736541_A_Literature_Review_on_Application_of_Sentiment_Analysis_Using_Machine_Learning_Techniques

Twitter API Access Levels and Versions. https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api

Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. Forbes. Gil Press 2016. https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=12f086fa6f63

Analysis Of Data Pre-Processing Methods For Sentiment Analysis Of Reviews. Tuba Parlar Selma Ayşe Özel Fei Song. 2019. https://www.researchgate.net/publication/331944353_Analysis_of_data_pre-processing_methods_for_the_sentiment_analysis_of_reviews

Sinar Harian 2020. 4.2 juta rakyat Malaysia belum daftar sebagai pengundi.

https://www.sinarharian.com.my/article/100343/BERITA/Nasional/42-juta-rakyat-Malaysia-belum-daftar-sebagai-pengundi

Jenny L. Davis, Tony P. Love, Gemma Killen 2018. Seriously funny: The political work of humor on social media.
https://www.researchgate.net/publication/324016051_Seriously_funny_The_political_work_of_humor_on_social_media

MalaysiaNow. 2021. Most youths don't care about politics, politicians, survey finds.
https://www.malaysianow.com/news/2021/05/07/most-youths-dont-care-about-politics-politicians-survey-finds/

Tan Jing Xuan (A175711)
Mohd Ridzwan Yaakub
Fakulti Teknologi & Sains Maklumat,
Universiti Kebangsaan Malaysia