IMPLEMENTASI *TEXT MINING* DAN ANALISIS SENTIMEN TERHADAP FORMULA E MENGGUNAKAN *NAÏVE BAYES CLASSIFIR* (NBC) DAN *SUPPORT VECTOR MACHINE* (SVM)

(Studi Kasus : Data Opini Twitter Tentang Formula E yang Diselenggarakan di Indonesia)

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana Program Studi Statistika

Disusun Oleh:

Diana Nabilla

16611118

PROGRAM STUDI STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2023

HALAMAN PERSETUJUAN PEMBIMBING

TUGAS AKHIR

Judul	:	Implementasi Text Mining dan Analisis Sentimen Terhadap Formula E Menggunakan Naïve Bayes Classifier (NBC) dan Support Vector Machine (SVM). (Studi Kasus : Data Opini Twitter Tentang Formula E yang Diselenggarakan di Indonesia).
Nama Mahasiswa	:	Diana Nabilla
NIM	:	16611118

TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK DIUJIKAN

Yogyakarta, 21 Februari 2023

Mengetahui,	Menyetujui,
Ketua Prodi Statistika	**Pembimbing**
(Dr.Atina Ahdika,S.Si.,M.Si.)	**(Rahmadi Yotenka,S.Si.,M.Sc.)**

HALAMAN PENGESAHAN

TUGAS AKHIR

IMPLEMENTASI *TEXT MINING* DAN ANALISIS SENTIMEN TERHADAP FORMULA E MENGGUNAKAN *NAÏVE BAYES CLASSIFIER* (NBC) DAN *SUPPORT VECTOR MACHINE* (SVM)

(Studi Kasus : Data Opini Twitter Tentang Formula E yang Diselenggarakan di Indonesia).

Nama Mahasiswa : Diana Nabilla

NIM : 16611118

TUGAS AKHIR INI TELAH DIUJIKAN
PADA TANGGAL : 27 FEBRUARI 2023

Nama Penguji	Tanda Tangan
1. Abdullah Ahmad Dzikrullah, S.Si.,M.Sc.	.....
2. Sekti Kartika Dini,S.Si.,M.Si.	.....
3. Rahmadi Yotenka,S.Si.,M.Sc.	.....

Mengetahui,

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

(Prof. Riyanto, S.Pd., M.Si., Ph.D.)

KATA PENGANTAR

Assalamu'alaikum Wr.Wb

Alhamdulillahirabbil`aalamin, puji syukur kehadirat Allah SWT yang telah melimpahkan rahmat, hidayah, serta karunianya sehingga penulis dapat menyelesaikan tugas akhir ini. Tidak pula shalawat serta salam tetap kita harturkan kepada baginda rasulloh Muhammad SAW yang banyak memberikan tauladan kepada ummatnya.

Atas izin Allah SWT, penulis dapat menyelesaikan tugas akhir berjudul Implementasi *Text Mining* dan Analisis Sentimen Terhadap Formula E Menggunakan *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM). (Studi Kasus : Data Opini Twitter Tentang Formula E yang Diselenggarakan di Indonesia). Tugas akhir ini merupakan syarat untuk memperoleh gelar sarjana Statistika pada program studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia.

Selama pelaksanaan dan penulisan tugas akhir ini penulis banyak mendapatkan dukungan dan bantuan dari berbagai pihak. Oleh karena itu penulis akan menyampaikan terimakasih kepada pihak-pihak yang telah memberikan dukungan dan bantuan nya:

1. Allah SWT yang telah memberikan rahmat, hidayah, dan karunianya sehingga penulis dapat menlaksanakan dan menyusun laporan tugas akhir ini.
2. Kedua orang tua, ayahanda Nurudin serta ibunda Halimah yang telah memberikan dukungan baik secara lahir dan moril sehingga penulis dapat menyelesaikan tugas akhir ini.
3. Kepada kakak serta adik, M. Wildan Khabibi serta Moch. Alfito Akmal yang sangat membantu saya dalam hal pengertian, perhatian, serta memberikan semangat motivasi untuk menyelesaikan tugas akhir ini.

4. Bapak Prof. Riyanto,S.Pd.,M.Si.,Ph.D. selaku dekan Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia.
5. Ibu Dr. Atina Ahdika, S.Si., M.Si. selaku ketua Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia.
6. Bapak Rahmadi Yotenka, S.Si.,M.Sc. selaku dosen pembimbing yang telah membantu dan memberikan masukan serta motivasi dari awal sampai selesai tugas akhir ini.
7. Ibu Mujiati Dwi Kartikasari, S.Si.,M.Sc. selaku dosen pembimbing akademik yang telah membimbing, memberikan masukan serta motivasi dari awal kuliah sampai selesai tugas akhir ini.
8. Seluruh staff pengajar program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia yang telah memberikan ilmu pengetahuan yang bermanfaat bagi penulis.
9. Sahabat seperjuangan : Barlinda Titania, Yesan Tiara, Nilda Aulia, Tri Binti dan Firli Windika atas rasa kekeluargaan serta semua dukungan, bantuan, dan memberikan semangat bagi penulis
10. Semua pihak yang telah membantu yang tidak bisa penulis sebutkan satu per satu.

Penulis menyadari masih banyak kekurangan dalam penulisan tugas akhir ini, maka segala kritik dan saran yang bersifat membangun sangat dibutuhkan untuk memperbaiki loporan tugas akhir. Semoga laporan ini dapat bermanfaat bagi penulis serta pembaca.

Wassalamualaikum Wr.Wb.

Yogyakarta, 27 Februari 2023

Diana Nabilla

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya karya yang sebelumnya pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan orang lain, kecuali yang diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 27 Februari 2023

Diana Nabilla

INTISARI

IMPLEMENTASI *TEXT MINING* DAN ANALISIS SENTIMEN TERHADAP FORMULA E MENGGUNAKAN *NAÏVE BAYES CLASSIFIER* (NBC) DAN *SUPPORT VECTOR MACHINE* (SVM)

(Studi Kasus : Data Opini Twitter Tentang Formula E yang Diselenggarakan di Indonesia)

Diana Nabilla
Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Islam Indonesia

Formula E merupakan seri balap mobil *single-seater* pertama yang menggunakan tenaga listrik. Balapan Formula E menjadi balapan pertama di Indonesia yang diselenggarakan di Jakarta *International* E-Prix *Circuit* (JIEC) Ancol pada Sabtu, 04 Juni 2022. Formula E ramai diperbincangkan di media sosial, khususnya twitter. Semua orang dapat berpendapat atau beropini tentang Formula E sehingga memunculkan banyak opini negatif atau positif. Media sosial twitter menjadi salah satu tempat yang banyak digunakan masyarakat untuk merespon sebuah peristiwa atau berita. Penelitian ini diharapkan dapat bermanfaat untuk melakukan riset atas opini masyarakat yang mengandung sentimen positif, netral atau negatif. Metode yang dipakai adalah analisis sentimen dengan melakukan *text preprocessing* data menggunakan *tokenisasi*, *cleaning, filtering*, dan untuk pelabelan sentimen menggunakan *lexicon based*. Untuk proses klasifikasinya menggunakan *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM). Data yang digunakan adalah *tweet* berbahasa Indonesia dengan *keyword* Formula E dengan jumlah data yang digunakan dalam analisis sebesar 1829 *tweet*. Hasil dari penelitian ini adalah klasifikasi SVM kernel linear memiliki akurasi yang tinggi dibandingkan algoritma lainya, dengan nilai akurasi sebesar 95,63%, *specificity* sebesar 95,52%, nilai *precision* sebesar 97,36%, dan nilai *recall* sebesar 95,68%.

Kata Kunci : Analisis Sentimen, Formula E, NBC, SVM, Twitter.

ABSTRACT

IMPLEMENTATION OF TEXT MINING AND SENTIMENT ANALYSIS OF FORMULA E USING NAÏVE BAYES CLASSIFIER (NBC) AND SUPPORT VECTOR MACHINE (SVM)

(Case Study : Twitter Opinion Data About Formule E Held In Indonesia)

Diana Nabilla
Department of Statistics, Faculty of Matematics and Natural Sciences
Universitas Islam Indonesia

Formula E is the first single-seater car racing series to use electricity. The Formula E race will be the first race in Indonesia to be held at the Jakarta International E-Prix Circuit (JIEC) Ancol on Saturday, 04 June 2022. Currently Formula E is being widely discussed on social media, especially Twitter. Everyone can have an opinion about the E Formula so that it gives rise to many negative or positive opinions. Social media Twitter is one of the places that many people use to respond to an event or news. This research is expected to be useful for conducting research on public opinion that contains positive, neutral or negative sentiments. The method used is sentiment analysis by performing text preprocessing of data using tokenization, cleaning, filtering, and for sentiment labeling using a lexicon based. For the classification process using the naïve Bayes classifier (NBC) and support vector machine (SVM). The data used are tweets in Indonesian with the keywords Formula E with a total of 1829 tweets. The results of this study are that the linear kernel SVM classification has high accuracy compared to other algorithms, with an accuracy value of 95,63%, a specificity of 95,52%, a precision value of 97,36%, and a recall value of 95,68%.

Keywords: Fomula E, NBC, *sentiment analysis*, SVM, Twitter.

BAB I
PENDAHULUAN

1.1. Latar Belakang Masalah

Internet dan *smartphone* merupakan salah satu produk teknologi yang banyak dimanfaatkan oleh masyarakat. Manusia dituntut untuk selalu mengikuti perkembangan teknologi serta informasi, dimana data dan informasi memiliki peranan sangat penting. Masyarakat saat ini memasuki era digital, dimana perubahan terjadi sangat cepat dan kompleks. Salah satunya adalah pola komunikasi masyarakat yang didukung oleh banyaknya jenis-jenis media sosial. Pada masa lalu, manusia berinteraksi secara langsung atau *face to face communication*, namun sekarang berubah menjadi lebih pasif (Alyusi, 2016). Meskipun kenyataannya akan lebih sibuk dengan kegiatannya masing-masing serta lebih asik dengan media sosial. Hal tersebut merupakan salah satu ciri-ciri bahwa kemajuan teknologi serta informasi tidak dapat dipisahkan dari kehidupan manusia.

Gambar 1.1 Pengguna Media Sosial di Indonesia (Sumber: We are social & Hootsuite)

Menurut survey dari We are social & Hootsuite pada tahun 2022 indonesia memiliki pengguna aktif media sosial sebesar 191,4 juta atau 68,9 % dari populasi penduduk Indonesia sebesar 277,7 juta jiwa. Dari survei tersebut dapat disimpulkan bahwa banyak penduduk atau orang yang memiliki telpon seluler lebih dari satu. Sedangkan rata-rata penduduk Indonesia mengabiskan waktu mengakses media sosial selama 3 jam 17 menit.

Media sosial tidak hanya digunakan sebagai sumber informasi dan sarana komunikasi namun dapat digunakan untuk mengontrol opini masyarakat. Hal ini menyebabkan masyarakat dapat secara langsung mengakses media sosial serta dapat mengemukakan opini terhadap sebuah isu yang sedang berkembang (Pahlevi, 2021). Opini masyarakat tesebut dapat mengarah kepada kepentingan yang benar atau sebaliknya. Hal tersebut memberikan pengaruh yang signifikan terhadap sebuah peristiwa. Youtube, whattsap, instagram, facebook dan twitter merupakan platforms media sosial yang sekarang ini banyak digunakan oleh masyarakat Indonesia. Pengguna twittter menempati peringkat ke-6 di Indonesia dan peringkat ke-15 di Dunia menurut survey we are social & hootsuite per Februari pada tahun 2022.

Twitter merupakan salah satu media sosial yang dipilih masyarakat dalam merespon sebuah peristiwa atau berita yang sedang berkembang. Pengguna twitter berasal dari berbagai kalangan dan lapisan masyarakat yang menyebabkan keragaman opini terhadap sebuah peristiwa atau berita. Pada media sosial twitter, Formula E menjadi kata populer yang banyak dibahas oleh masyarakat saat ini.

Gambar 1.2 *Tranding Topic* Formula E (Sumber: Twitter)

Berdasarkan gambar 1.2 menjelaskan bahawa topik mengenai Formula E menjadi topik yang banyak dibahas oleh pengguna twitter beberapa waktu belakangan ini. Pada 31 Mei 2022 ada lebih dari 22.000 ribu lebih kicauan/*tweet* mengenai topik ini hanya dalam kurun waktu 24 jam. Formula E merupakan perhelatan pertama di Indonesia. Hal ini mengakibatkan banyak berita yang membahas Formula E. Kebanyakan berita membahas persiapan penyelenggaraan serta banyak dihubungkan dengan politik. Banyak opini yang menganggap Formula E dijadikan alat untuk mencari kelemahan lawan politik. Opini tersebut terus dibahas sampai menjelang terselenggaranya balapan dan ditambah dengan terjadinya kerusakan pada atap tribun Formula E. Robohnya atap tribun tersebut membuat banyak masyarakat yang meragukan kesiapan penyelenggaraan acara ini.

Hal ini diperparah dengan banyaknya berita hoax yang bermunculan. Sehingga opini mengenai Formula E sangat beragam dari negatif ataupun positif.

Formula E merupakan seri balap mobil *single-seater* pertama yang menggunakan tenaga listrik. Balap mobil Formula E berbeda denga F1. Salah satu perbedaan yang paling mendasar adalah suara mobil Formula E tidak sebising Formula 1. Formula E memakai mesin *full electric* sebagai sumber tenaganya, namun Formula 1 memakai mesin konvensional *Internal Combustion Engine* (ICE) dengan bantuan motor listrik, yang artinya disebut juga sebagai mesin *hybrid.* Walaupun berbeda balapan mobil Formula E sudah menjadi kejuaraan dunia serta setara dengan F1. Indonesia menjadi salah satu negara yang berkesempatan menjadi tuan rumah ajang balap Formula E. Gelaran balapan Formula E di Indonesia akan diselenggarakan di Jakarta *International* E-Prix *Circuit* (JIEC) Ancol pada 4 Juni 2022.

Indonesia mempunyai populasi penduduk berjumlah 273,5 juta jiwa serta berada diposisi keempat dunia setelah Amerika Serikat (worldometers, 2022). Besarnya populasi, persentase pengguna media sosial serta lamanya waktu yang dihabiskan untuk mengakses media sosial tersebut menghasilkan informasi yang tidak terhitung jumlahnya. Informasi yang terkandung dalam media sosial dapat berupa gambar, video, ataupun tanggapan masyarakat terhadap suatu peristiwa. Dalam menggali atau mencari informasi yang terkandung dalam ribuan data *tweet*, maka dilakukan analisis yang tepat, sehingga memberikan atau mengahasilkann informasi yang dapat membantu banyak pihak untuk mendukung suatu keputusan atau pilihan.

Pengambilan data dilakukan dengan menggunakan teknik *scraping* dimedia sosial twitter. Data yang diperoleh berupa teks, gambar, bahkan terdapat video atau gabungan dari keduanya, data seperti ini disebut dengan data tidak terstruktur. Data tidak terstruktur ini lebih sulit untuk dikelola dan dianalisis dibandingkan dengan data terstruktur. Dalam mengolah data tak terstruktur dikenal dengan *text mining*. *Text mining* merupakan merupakan proses untuk mengektraksi informasi tersembunyi pada data tekstual yang berjumlah besar (Nugraha, Harani, & Habibi, 2021). *Text mining* membantu mengubah data menjadi data yang lebih terstruktur sehingga diperoleh informasi yang lebih bernilai. *Text mining* merupakan salah satu

teknik dalam data mining yang menggunakan teks. Prinsipnya *text mining* melibatkan *information retrieval* (IR), *text analysis*, *information extraction* (IE), *clustering*, *categorization*, *visualization*, *database technology*, *natural language processing* (NLP), *machine learning*, dan data mining (Nugraha, Harani, & Habibi, 2020).

Analisis sentimen atau disebut juga dengan *opinion mining* merupakan bidang studi yang menganalisis pendapat, evaluasi, penilaian, sikap dan emosi terhadap suatu entitas seperti produk, layanan, organisasi, individu, masalah, peristiwa atau topik. *Opinion mining* diekspresikan dalam bentuk teks atau kalimat sehingga dapat ditentukan klasifikasinya sebagai sentimen positif, negatif atau netral. Terdapat banyak metode klasifikasi yang dapat digunakan dalam analisis sentimen yaitu *Naïve Bayes Clasification* (NBC) dan *Support Vector Machine* (SVM).

Metode klasifikasi *Support Vector Machine* (SVM) dipilih karena mampu menemukan *hyperplane* paling optimum yang memisahkan setiap kelas diantara kelas-kelas data yang dimiliki (Asnawi , Firmansyah, & dkk, 2021), dapat membuat *margin* pemisah yang jelas jika dibandingkan dengan *decision tree* maupun model algoritma yang lain karena SVM dapat menangani pemisahan data yang non linear (Wahyuni, Arifiyati, & Afandi, 2020), proses *learning* yang cepat (Budianto, Maryono, & Ariyuana, 2018). Hal tersebut terlihat dalam penelitian (Pertiwi, 2018) yang menunjukan metode SVM memiliki tingkat akurasi yang lebih tinggi sebesar 76,9% dibandingkan dengan metode K-NN. Pada metode *naïve bayes clasification* (NBC) merupakan pengklasifikasian probabilitas sederhana berdasarkan pada teorema bayes. Metode ini mudah diimplementasikan dan waktu pemrosesan yang cepat. Selain itu metode ini popular dan banyak digunakan karena kemudahan serta kesederhanaannya, namun memberikan hasil klasifikasi yang setara dengan *decision tree* dan *neural network*. Hal tersebut terlihat dalam penelitian (Asnawi , Firmansyah, & dkk, 2021) yang menunjukan metode *Naïve Bayes Classifier* (NBC) memiliki tingkat akurasi lebih tinggi sebesar 79.8% dibandingkan metode K-NN.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah :

1. Bagaimana gambaran umum mengenai Formula E yang diadakan di Indonesia?
2. Bagaimana hasil klasifikasi data opini masyarakat terhadap Formula E menggunakan media sosial twitter?
3. Bagaimana perbandingan akurasi data opini masyarakat terhadap Formula E menggunakan metode metode *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM)?

1.3. Batasan Masalah

Batasan dalam penelitian ini adalah :

1. Penelitian ini menggunakkan data hasil *scraping* dimedia *social* twitter.
2. Data yang digunakan untuk penelitian adalah data *scraping* terkait Formula E yang dilakukan pada 23-31 Mei 2022 berjumlah 15000 data. Sedangkan data yang digunakan dalam analisis sebesar 1829 data.
3. *Key word* atau kata kunci yang digunakan untuk *scraping* data adalah Formula E.
4. Metode yang digunakan adalah *Text Mining*, analisis sentimen, dan *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM).
5. Aplikasi yang digunakan dalam penelitian adalah R Versi 3.6.1, R Studio, dan excel.

1.4. Jenis Penelitian dan Metode Analisis

Data yang digunakan pada penelitian ini merupakan data sekunder mengenai *tweet*s atau kicauan terkait Formula E dari media sosial twitter. Data tersebut diperoleh dengan cara melakukan *scraping* dengan memanfaatkan API *Key* Twitter. Data yang diperoleh dianalisis menggunakan *text mining*, analisis sentimen serta menggunakan *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM).

1.5. Tujuan Penelitian

Tujuan dari penelian ini adalah :

1. Mengetahui gambaran umum mengenai Formula E yang diadakan di Indonesia.

2. Mengetahui klasifikasi data opini masyarakat terkait Formula E menggunakan media sosial twitter.
3. Mengetahui perbandingan akurasi data opini masyarakat terhadap Formula E menggunakan metode metode *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM).

1.6. Manfaat Penelitian

Manfaat penelitian ini adalah :

1. Menambahkan pengetahuan mengenai pengaplikasian *text mining,* analisis sentimen pada media sosial twitter.
2. Mengetahui tanggapan masyarakat mengenai sentimen positif dan negatif mengenai pagelaran Formula E di Indonesia untuk pertama kali.
3. Hasil penelitian ini diharapkan masyarakat tidak mudah terprovokasi oleh berbagai informasi atau berita khususnya mengenai Formula E yang tidak jelas sumbernya. Masyarakat diharuskan lebih cermat dalam memahami isi berita serta harus memeriksa sumber keaslian berita.
4. Penelitian ini juga diharapkan masyarakat dapat memahami pentingnya bermedia sosial yang baik. Dimana masyarakat tidak boleh menggunakan kata kasar, provokatif, porno atau SARA, aksi kekerasan dan sebagainya. Hal tersebut penting agar masyarakat terhindar dari berita-berita hoax yang dapat memecah belah masyarakat.

BAB II
TINJAUAN PUSTAKA

Penelitian terdahulu menjadi penelitian yang sangat penting dikarenakan menjadi dasar atau acuan dalam menyusun suatu penelitian. Dengan adanya penelitian terdahulu ini, maka dapat diketahui perbedaan serta keterkaitan dengan penelitian yang akan disusun. Selain itu, berfungsi juga untuk menghindari duplikasi dari penelitian yang akan disusun. Berikut ini adalah penelitian-penelitian terdahulu yang menjadi dasar atau acuan dalam penelitian ini:

Tabel 2.1 Tabel Penelitian Terdahulu.

Nama	Judul	Metode Penelitian	Hasil Penelitian
Amin Nur Rohim, Ahmad R. Pratama (2022)	Analisis sentimen publik di media sosial Instagram atas kinerja Presiden Joko Widodo.	*Naïve Bayes Classifier*	Data yang diperoleh sebanyak 1152 komentar pada media sosial instagram. Berdasarkan analisis yang dilakukan pada metode *naïve bayes classifiier* mampu melakukan analisis sentimen terkait komentar publik dengan baik dengan nilai *accuracy* 83%, *precision* 81% dan *recall* 84% dan hasil dari analisis sentimen terhadap data komentar menunjukan sentimen positif berdasarkan kemunculan kata yang didominasi oleh kata sehat, bangga, mantap, presiden dan keren.

Nama	Judul	Metode Penelitian	Hasil Penelitian
Dwi Normawati, Surya Prayogi (2021).	Implementasi *Naïve bayes classifier* dan *Confusion Matrix* pada analisis sentimen berbasis teks pada twitter.	*Naïve Bayes Classifier* dan *Confusion Matrix*	Penelitian ini menghasilkan pemaparan yang terstruktur pada proses dan hasil implementasi NBC dan pengujian performa menggunakan *confusion matrix* yang didapatkan akurasi sebesar 82%, presisi 93%, dan *recall* sebesar 52%.
Tanthy Tawaqalia Widowati, Mujiono Sadikin (2020).	Analisis Sentimen Twitter Terhadap Tokoh Publik dengan Algoritma Naïve Bayes dan *Support vector Machine.*	*Support Vector Machine* dan *Naïve Bayes.*	Data yang digunakan berasal dari *tweet*s dengan kata kunci "nadiem makariem", "kemendikbud" dan "pak nadiem". Dari hasil analisis diketahui bahwa untuk kasus yang diteliti, metode *naive bayes* menghasilkan kinerja yang lebih baik dengan *accuracy* 91.48%, *precision* 89.28% dan *recall* 91.58%.
Veny Amilia Fitri, Rachmadita Andreswari, M. Azani Hasibuan (2019).	*Sentimen analysis of media social twitter with case of ANTI-LBGT campaign in Indoonesia using Naïve Bayes, Decission Tree, and Random Forest Algorithm.*	*Naïve Bayes*, *Decision Tree* dan *Random Forest.*	Pada penelitian ini menunjukan pengguna twitter di Indonesia memberikan komentar yang lebih netral. Metode *naïve bayes* memperoleh akurasi sebesar 86,43%, sedangkan mengunakan alat di RapidMiner akurasinya lebih tinggi daripada metode lainya. Serta menggunakan *decission tree* dan *random forest* nilai akurasinya adalah 82,91%.

Nama	Judul	Metode Penelitian	Hasil Penelitian
Hennie Thuteru, Ade Iriani (2018).	Analisis sentimen perusahaan listrik cabang Ambon menggunakan metode *Support Vector Machine* dan *Naïve Bayes Classifier.*	*Support Vector Machine* dan *Naïve Bayes Classifier.*	Hasil penelitian menggunakan metode klasifikasi NBC dengan 2 fold yaitu sebesar 67.2%. Sentimen positif yang diperoleh dari klasifikasi NBC sebesar 67%, sentimen netral 19%, dan sentimen negatif 14%. Sedangkan pada metode klasifikasi SVM memiliki akurasi yang lebih baik ketika menggunakan 2 fold. Sentimen positif yang diperoleh dari penggunaan metode SVM sebesar 24%, sentimen netral 29%, dan sentimen negatif 47%. Penelitian ini menunjukan rata-rata tingkat akurasi metode klasifikasi SVM yang lebih baik dari pada metode NBC, yaitu sebesar 76.42%.

Berdasarkan pemaparan berbagai acuan penelitian pada tabel 2.1 penulis akan melakukan penelitian dengan mengimplementasikan *text mining* dan analisis sentimen. Data diperoleh dengan cara melakukan *scrapping* melalui media sosial twitter mengenai Formula E. Selanjutnya, melakukan analisis sentimen yang diklasifikasikan menjadi dua kelas sentimen yaitu negatif dan positif dengan menggunakan algortima *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM). Tidak hanya melakukan klasifikasi, namun penulis juga melakukan analisis deskriptif dan membuat visualisasi data menggunakan *Word Cloud* serta asosiasi kata.

Jika dibandingkan dengan penelitian-penelitian sebelumnya, penelitian ini memiliki fokus untuk membandingkan hasil klasifikasi dari masing-masing

algortima. Hal ini dilakukan untuk mengetahui algoritma yang efektif dalam melakukan klasifikasi data. Untuk mengukur kinerja dari algoritma *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM) adalah menggunakan *confusion matrix*.

BAB III
LANDASAN TEORI

3.1. Twitter

Twitter didirikan oleh 3 orang yaitu *Jack Dorsey, Biz Stone*, dan *Evan Williams* pada bulan maret tahun 2006 dan baru diluncurkan pada bulan Juli pada tahun yang sama (Woloeyo, 2010). Twitter adalah salah satu jenis layanan jejaring sosial (media sosial) dan juga *microblog* yang memungkinkan penggunanya mengirim dan membaca pesan. Para pengguna twitter hanya dapat menulis dengan jumlah karaktater 280 saja. *Microblog* menjadi layanan dasar twitter. *Microblog* adalah sebuah layanan *blog* dimana penggunanya bisa menuliskan pembaharuan (*update*) statusnya secara singkat dan mempublikasikannya (juju & studio, 2009). Jumlah karakter yang dapat ditulis di*mikroblog* hanya 280 karakter. Hal inilah yang membedakan *mikroblog* dengan layanan *blog* yang lainnya. Pada awalnya jumlah karakter yang dapat ditulis pada twitter hanya 140 karakter, namun pada 7 november 2017 ditambah menjadi 280 karakter (Rosen, 2017).

Berdasarkan (Kusuma, 2009) terdapat beberapa istilah yang sering ditemui pada twitter, yaitu:

- *Timeline* adalah daftar *tweet* terbaru dari pengguna twitter yang diikuti pemilik akun, termasuk *tweet* yang dibuat pemilik akun.
- *Direct Message* (DM) yaitu fasilitas berkirim pesan antar pengguna secara lebih *private*. DM hanya bisa dilakukan oleh pihak yang diikuti (di-*follow*).
- *Trending topics* adalah daftar tema yang tengah hangat diperbincangkan di kalangan pengguna twitter.
- *Tweet* merupakan informasi yang terdiri dari pesan 140 karakter. *Tweet* berisi berita terbaru atau ter*update* yang berkaitan dengan hal - hal yang pemilik akun sukai.
- *Reply tweet* atau sering disebut *response tweet* (RT) adalah komentar atau balasan atas *tweet*.
- *Retweet* adalah menyalin seluruh isi *tweet* dari akun lain.
- *Follow* adalah mengikuti akun dan informasi yang disampaikan oleh seorang pengguna.

- *Follower* adalah pengikut atau yang mengikuti akun seseorang.
- *Mention* (@) digunakan untuk menyebut *username* pihak yang akan diajak berkomunikasi. Penggunaan simbol ini berada diawal sebelum menuliskan username pihak yang dituju.
- *Hastags* atau tanda pagar atau tagar (#) adalah tanda yang digunakan untuk menandai kata kunci untuk topik diskusi atau informasi yang dibagikan agar mudah dicari.

3.2. Formula E

Formula E merupakan kejuaraan olahraga otomotif mobil *single seater* yang menggunakan energi listrik. Kejuaraan pertama kali digagas oleh presiden FIA Jean Todt dan Alejandro Agag yang juga merupakan kepala eksekutif dari Formula E. Formula E, secara resmi bernama ABB FIA Formula E *World Championship*.

Kejuaraan Formula E bermula saat Jean Todt dan Alejandro Agag bertukar pikiran mengenai ide menggelar balapan mobil Formula berbahan bakar listrik dalam sebuah restoran di Paris pada 3 Maret 2011. Misi awal pendiri Formula E adalah melakukan balapan dijalanan kota-kota yang merupakan tempat ikonik diberbagai dunia. Tidak hanya menunjukan pemandangan yang berbeda, balapan juga diikuti oleh pembalap dan tim terbaik yang menggunakan energi listrik sebagai sumber energi untuk masa depan.

Formula E pertama kali berlangsung pada 13 September 2014 di Beijing *Olympic green circuit, China*. Pada seri pertama Formula E menggunakan mobil *Gen 1* berteknologi *Spark-renault SRT 01E*. Teknologi tersebut dibuat *oleh spark racing technology* dan sasisnya dibuat oleh Dallara serta girboksnya dibuat oleh Hewlanad. Sedangkan pemasok bannya adalah dari Michelin. Mobil balap *Gen 1* mempunyai kekuatan sekitar 250 tenaga kuda (190 kW). Mobil ini mampu berakselerasi dari 0-100 km dalam 3 detik, dengan kecepatan maksimum sebesar 225 km/h. mobil gen 1 digunakan dalam Formula E hingga musim keempat.

Pada musim 2018-2019 Formula E memperkenalkan mobil generasi kedua (*Gen 2*) dengan kemajuan yang signifikan dari mobil sebelumnya. Perubahan terlihat pada peningkatan kekuatan menjadi 250 kW dan kecepatan maksimum sebesar 290 km/h. Kedatangan mobil *Gen 2* juga mngekhiri pergantian mobil ditengah balapan.

Kemudian pada musim kesembilan (2022/2021) Formula E memperkenalkan mobil terbaru yaitu *Gen 3*. Musim kesembilan digelar di Mexico pada 14 Januari 2023. Mobil *Gen 3* mempunyai kecepatan mencapai 322 km/h. Mobil *Gen 3* juga mempunyai efisiensi daya sekitar 95% sehingga menghasilkan tenaga maksimum hingga 350 kW. Selain itu Gen 3 juga menjadi mobil dengan *powertrain* depan dan belakang. *Powertrain* depan menambahkan tenaga 250 kW ke 350 kW ke belakang, yang artinya kemampuan regeneratif *Gen 3* menjadi 600 kW. Formula E mengklaim bahwa mobil *Gen 3* menjadi mobil paling efisien serta mampu mengisi daya dengan kecepatan yang sangat tinggi.

Gambar 3.1 Mobil *Gen* 3 (Sumber : (https://www.fiaformulae.com/, 2022))

Kemudian pada musim 2020/2021, Formula E resmi masuk dalam kejuaraan dunia FIA dan menjadi balapan mobil *single seater* pertama yang mengggunakan energi listrik, serta menjadi kejuaran dunia kedua setelah Formula 1.

3.2.1 Format Balapan

Kejuaraan Formula E saat ini terdiri dari 11 tim, 22 mobil dan 22 pembalap. Kejuaran ini setiap tahun mengalami perkembangan yang sangat pesat hingga setara dengan Formula 1. Balapan berlangsung di sirkuit jalan raya yang menjadi tempat ikonik disuatu negara dengan panjang lintasan 1,9 km hingga 3,4 km.

Formula E mempunyai format balapan yang unik. Formula E mempunyai konsep balapan hanya dalam sehari. Konsep tersebut artinya *free practice*, kualifikasi dan balapan (*E-Prix*) dilakukan pada hari yang sama. Format balapan Format E sangat berbeda dengan Formula 1 yang setiap sesinya digelar dalam 3 hari berturut-turut. Berikut merupakan sesi balapan Formula E:

a. *Practice*

Setiap balapan mempunyai dua sesi latihan. Setiap sesi latihan mempunyai waktu 30 menit. Sesi latihan ini merupakan pertama kalinya bagi pembalap dan tim akan turun ke trek sirkuit. Sesi ini dimanfaatkan oleh pembalap dan timnya untuk beradaptasi dengan *set-up* mobil.

b. Kualifikasi

Format kualifikasi Formula E memungkinkan tim dan pembalap untuk menunjukan keterampilan dan kecepatan mereka. Konsep yang dipakai Formula E adalah *battle head to head*. Kualifikasi diawali dengan fase grup. Dimana pembalap dibagi menjadi dua grup yaitu grup A dan grup B. Pembagian ini berdasarkan posisi pembalap di klasemen kejuaraan, pembalap yang berapa diposisi ganjil masuk ke grup A, sedangkan pembalap yang berada diposisi genap masuk ke grup B.

Total terdapat 22 pembalap yang akan terbagi menjadi dua grup (grup A dan grup B) dengan masing-masing grup berisi 11 pembalap. Terdapat pengecualian pada musim pertama balapan karena belum terdapat peringkat klasemen, maka setiap tim dapat mimilih satu pembalap kedalam setiap grup. Setiap grup mendapatkan waktu 10 menit untuk memperoleh putaran dengan waktu tercepat. Dari dua grup tersebut, empat pembalap tercepat masing-masing grup (total delapan pembalap) akan masuk ketahap *duel*, perempat serta final. Dua pemenang dari masing-masing grup kemudian bersaing satu sama lain secara *head to head* dalam sistem gugur selama delapan be sar tercepat. Kemudian yang tercepat dari masing-masing grup akan berhadapan di *final duel*.

Pemenang *final* kemudian memperoleh posisi pertama atau *pole position*, sedangkan *runner-up* berada diposisi kedua. Semi-finalis akan menempati posisi ketiga dan keempat dan yang kalah diperempat final menempati posisi kelima sampai kedelapan, berdasarkan waktu yang diperoleh. Pembalap lainnya dari penyisihan grup ditempatkan bergantian dari posisi kesembilan, dengan *polesitter* ditempat ganjil dan kelompok lain ditempat genap.

c. *E-prix* atau Balapan

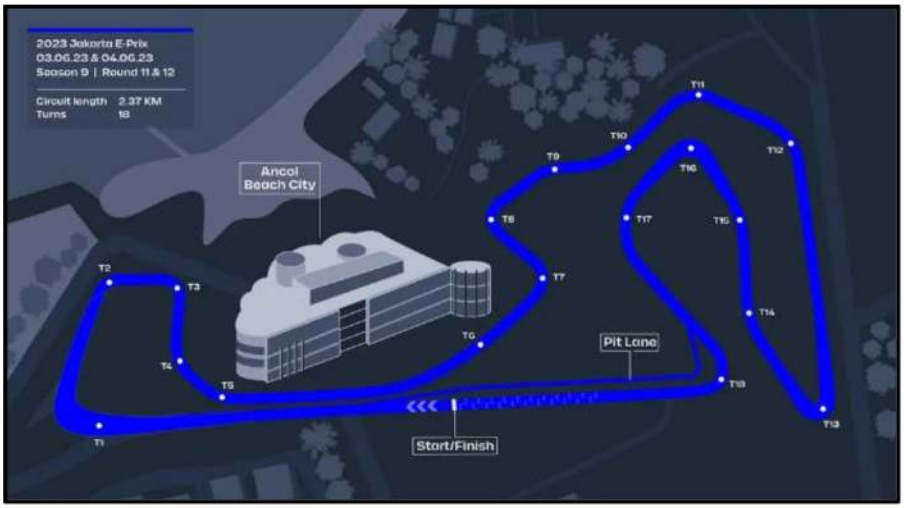
E-Prix atau balapan dimulai dengan *start* berdiri, artinya mobil tidak bergerak sampai lampu menyala hijau. Balapan akan berlangsung selama 45 menit ditambah dengan 1 *lap*. Dalam Formula E tidak bergantung pada *lap* namun pada waktunya.

Ini artinya banyaknya *lap* pada setiap balapan berbeda-beda. Setelah pembalap telah melewati garis *finish* selama 45 menit, masih ada satu putaran lagi untuk menyelesaikan balapan.

Pada musim 2018/2019 diperkenalkan *attack mode*. *Attack mode* merupakan zona aktivasi yang memungkinkan pembalap mengambil tenaga ektra dengan resiko ditanggung oleh pembalap. Setelah melewati zona aktivasi pembalap akan memperoleh tambahan daya 30 kW.

Pada Formula E terdapat pula *fanboost*, *fanboost* merupakan fitur dukungan dari para *fans* yang dapat membantu pembalap untuk menyalip atau untuk mejauhi pesaingnya dan dapat digunakan untuk mempengaruhi posisi saat *race*. Pembalap akan dipilih berdasarkan banyaknya *voting* yang diberikan. *Voting* dilaksanakan pada seminggu sebelum balapan hingga 15 menit sebelum balapan. *Fanboost* diberikan pada lima pembalap yang masing-masing mendapatkan tambahan daya 30 kW (Killowatt). Daya tersebut dapat mencapai kapasitas maksimal *power* mobil sebesar 250 kW.

3.3. E-Prix Jakarta



Gambar 3.2 Jakarta *International* E-Prix *Circuit*. (Sumber : (https://www.fiaformulae.com/, 2022))

E-Prix Jakarta merupakan balapan tahunan dari kejuaraan mobil bertenaga listrik *single seater* yang akan dilaksanakan di Jakarta, Indonesia. Balapan ini akan menjadi balapan pertama di Indonesia yang diselenggarakan langsung oleh badan FIA. Formula E akan digelar pada 4 Juni 2022 di *Jakarta International E-Prix Circuit* (JIEC) yang berada dikawasan Ancol. Sirkuit ini mempunyai panjang

lintasan 2,4 km, lebar lintasan 12 meter, lintasan lurus sepanjang 600 m, dan mempunyai 18 tikungan. Uniknya, *design* sirkuit JIEC terinspirasi dari kuda lumping.

Pagelaran ini sempat mengalami penundaan beberapa kali, salah satunya karena pandemi COVID-19 ditahun 2020. Penundaan juga terjadi akibat perubahan letak sirkuit, sebelumnya balapan direncanakan digelar disekitar kawasan Monumen Nasional. Namun komisi pengarah tidak menyetujui apabila sirkuit dibangun dikawasan tersebut. Penundaan terjadi sampai pada desember 2021 akibat lokasi, perizinan, logistik dan sebagainya. Akhirnya pada 22 Desember 2021, pembangunan sirkuit disetujui serta dibangun dikawasan Ancol. Pembangunan sirkuit dimulai pada 3 Februari dan dinyatakan selesai pada 1 Juni 2022.

Formula E yang dilaksanakan pada 4 Juni 2022 diikuti oleh 11 *team* dan 22 pembalap dari berbagai negara. Jakarta *E-Prix* merupakan balapan kesembilan dalam kalender kejuaraan Formula E 2021-2022. Mitch Evans keluar sebagai pemenang balapan Formula E Jakarta dari tim *Jaguar Racing* serta memperoleh 25 poin. Pada posisi kedua dan ketiga diperolah Jean Eric Vergne (*Tim Techeetah*) dan Edoardo Mortara (*Venturi*).

3.4. Data Mining

Perkembangan bidang ilmu yang mempelajari penggalian data (Data Mining) sedang banyak diminati oleh berbagai kalangan mulai dari akademisi, praktisi, industri informasi, dan masyarakat. Definisi data mining banyak dikemukakan oleh para ahli. Salah satu definisi yang terkenal oleh kusnawi (Gustiana, Priyanto, & dkk, 2021). Pengertian data mining adalah sebagai berikut (Kusnawi, 2007):

1. Data mining adalah serangkaian proses menggali atau menambang suatu kumpulan data untuk menghasilkan pengetahuan yang selama ini tidak diketahui secara manual.
2. Data mining atau *knowledge discovery in database* (KDD) merupakan proses pengambilan informasi yang tersebunyi, dimana informasi tersebut sebelumnya tidak dikenal. Proses dalam KDD meliputi proses pendekatan secara teknis seperti *clustering*, klasifikasi dengan meliputi metode yang merupakan irisan *artifisial intelligence* (AI), *machine learning* (ML), dan statistik.

Pada dasarnya data mining merupakan langkah analisis untuk mengambil atau mengekstrak kumpulan data sehingga memperoleh informasi yang dibutuhkan. Data mining dapat menangani data berskala besar serta dapat memanfaatkan data lampau untuk meningkatkan proses model pembelajarannya seperti pada penerapan klasifikasi (Gustiana, Priyanto, & dkk, 2021). Dalam prosesnya data mining menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* yang berfungsi untuk melakukan ekstraksi dan identifikasi informasi yang bermanfaat (Prasetyowati, 2017).

3.5. *Machine Learning*

Machine learning didefinisikan sebagai tipe kecerdasan buatan *artificial intelligence* (AI) yang berhubungan dengan pengembangan teknik-teknik yang dapat diprogramkan dan belajar dari data masa lalu (Daqiqil, 2021). *Machine learning* berfokus pada pengembangan program komputer yang dapat mengakses data dan mempelajari data untuk membuat model yang dapat digunakan dalam memecahkan kasus tertentu. Contoh penerapan *machine learning* dalam kehidupan adalah sebagai berikut

1. pada bagian *computer vision* contohnya adalah pengenalan wajah dan pelabelan wajah seperti pada facebook. Contoh lainya yaitu penerjemahan tulisan tangan menjadi teks.
2. pada bidang *information retrival* contohnya adalah mengubah suara menjadi teks, filter *spam* dan penerjemahan bahasa pada komputer.

Berdasarkan cara belajarnya *machine learning* dibagi menjadi 3 kelompok yaitu (Kusuma P. D., 2020):

1. *Supervised learning*

Secara bahasa, *supervised learning* adalah pembelajaran terarah atau terawasi. Pada *supervised learning* bertujuan membangun fungsi *input-output* atau hubungan *input* dan *output* berdasarkan data yang tersedia. Data yang disediakan untuk membangun fungsi disebut dengan data latih. Fungsi yang terbangun digunakan untuk memprediksi *output* berdasarkan data. Selain data latih terdapat pula data uji. Data uji digunakan untuk mengukur akurasi sistem atau fungsi yang telah dibangun.

2. *Unsupervised learning*

Unsupervised learning adalah metode belajar yang tidak terawasi. Pada pendekatan ini penerapannya tanpa ada data latih atau data *training*. Pengamatan yang diperoleh akan dikelompokan atau dimasukan dalam beberapa kelas sesuai yang dikehendaki oleh peneliti. Hal tersebut dapat dilakukan karena data yang dimiliki tidak mempunyai label yang dapat menandai data. Berbeda dengan *supervised learning*, pendekatan ini tidak mempunyai target *output* yang eksplisit sehingga dapat digunakan untuk kebutuhan pengelompokan.

Sebagai contoh seorang guru akan menerapkan kelompok belajar. Maka guru tesebut dapat membentuk kelompok berdasarkan jarak tempat tinggal atau sesuai dengan alasan lainnya. Pembentukan kelompok pada pendekatan *unsupervised learning* tanpa data latih. sehingga pengelompokan hanya berdasarkan pada kemiripan data yang dimiliki.

3. *Reinforcement learning*

Reinforcement learning adalah metode yang mempelajari aturan kontrol dengan cara berinteraksi dengan lingkungan yang masih asing. Pengalaman interaksi tersebut akan terakumulasi sehingga program dapat mengambil kesimpulan atau keputusan dikemudian hari menggunakan pola-pola yang telah dipelajarinya (Alpaydin, 2010). Pendekatan ini melatih program untuk mengambil keputusan spesifik berdasarkan kebutuhan dengan tujuan untuk memaksimalkan kinerja.

3.6. *Text Mining*

Menurut Feldman dan Sanger (2007), *Text mining* diartikan sebagai proses pengetahuan yang memungkinkan penggunanya berinteraksi dengan dokumen dari waktu kewaktu menggunakan berbagai macam analisis. *Text mining* digunakan untuk mengumpulkan, mempelajari pola-pola dokumen sehingga diperoleh informasi yang dibutuhkan. Secara umum kinerja *text mining* hampir sama dengan data mining. Perbedaan yang paling dasar terletak pada tipe datanya. Data mining bekerja pada data terstruktur sedangkan data mining bekerja pada data tidak terstruktur (Tan, 2020). Data tidak terstruktur biasanya diperoleh dari berbagai sumber diantaranya artikel, berita, hingga media sosial. Data teks yang diperoleh dari berbagai sumber tersebut terutama dari media sosial umumnya mempunyai struktur yang berbeda dan menggunakan bahasa yang tidak baku. Disinilah peran

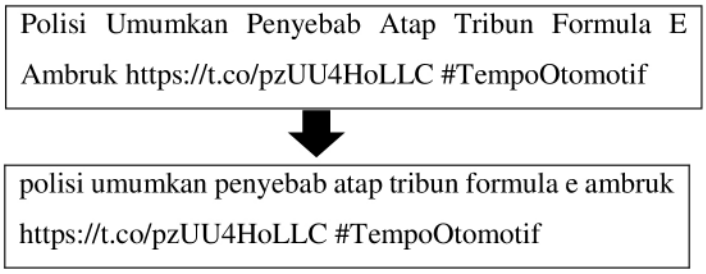
text mining yang menyediakan pengolahan data sampai memperolah informasi yang dibutuhkan.

3.6.1 *Text Pre-processing*

Text pre-processing atau pra-proses merupakan tahap terpenting sebelum melakukan analisis lainnya. *Text pre-pocessing* dapat didefinisikan sebagai proses mempersiapkan data mentah sebelum dilakukan proses lainnya. Data teks mentah yang diperoleh biasanya merupakan data tidak terstruktur serta terdapat banyak *noise* seperti tanda baca, imbuhan, angka, karakter-karakter khusus dan lain sebagainya. Pada tahap ini, teks dibersihkan sehingga bentuk dasar dari masing-masing kata. Secara umum tahapan yang dilakukan pada *text pre-processing* adalah sebagai berikut:

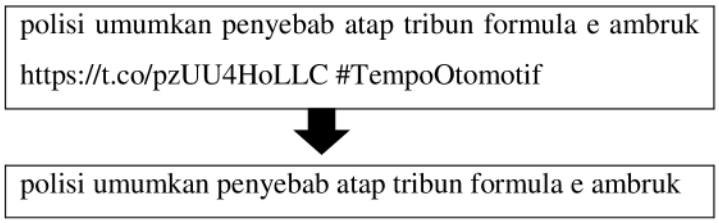
1) *Case folding*

Case folding atau dibeberapa buku disebut dengan "*case convension*" merupakan proses penyeragaman *case* atau huruf yang terdapat dalam sebuah dokumen. Tidak semua dokumen konsisten dalam menggunakan huruf kapital. Oleh sebab itu, dibutuhkan konversi seluruh abjad yang ada dalam dokumen menjadi huruf kecil. Dalam melakukan penyeragaman huruf menjadi huruf kecil menggunakan fungsi lower().



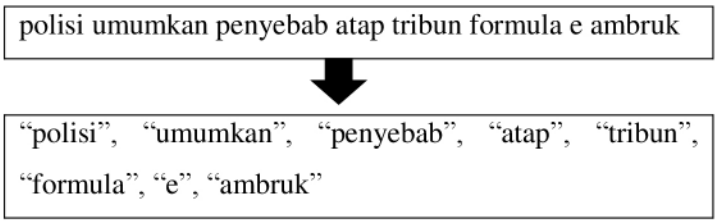
2) *Cleaning*

Cleaning adalah proses membersihkan teks dari karakter-karakter yang tidak diperlukan untuk mengurangi *noise*. Proses *cleaning* tergantung pada jenis datasetnya. Untuk data twitter ada beberapa karakter yang perlu dihilangkan seperti nama akun, *hastag*, karakter angka, URL, *emoticon*, dan tanda baca.



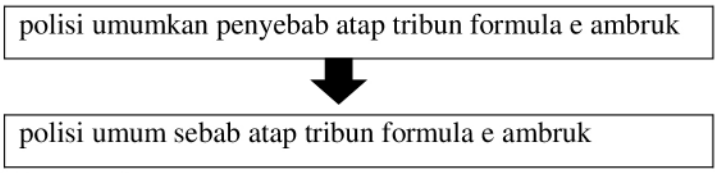
3) *Tokenization*

Tokenization merupakan tahapan yang dilakukan untuk melakukan pemotongan *string input* berdasarkan kata penyusunnya. Proses tokenisasi akan memecah sekumpulan teks atau karakter kedalam satuan kata. Tahapan ini akan membagi teks dari kalimat berdasarkan spasi dan tanda baca untuk tahap analisa teks selanjutnya.



4) *Stemming*

Stemming adalah proses untuk menemukan kata dasar. Proses *stemming* berbeda tergantung bahasa yang digunakan. Dalam Bahasa Indonesia mengenal adanya awalan, akhiran serta sisipan, sedangkan dalam Bahasa Inggris hanya mengenal akhiran. *Stemming* dalam Bahasa Indonesia dilakukan dengan cara menghilangkan semua imbuhan baik berupa awalan (*prefixes*), akhiran (*suffixes*), ataupun kombinasi awalan dan akhiran (*confixes*) yang terdapat dalam setiap kata dalam teks. Dimana dari hasil proses tersebut akan memperoleh sebuah informasi mengenai banyaknya kata yang muncul dalam sebuah dokumen.



5) *Stopword Removal*

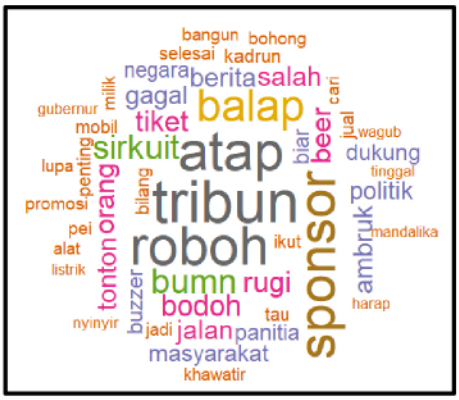
Stopword removal adalah proses untuk menghilangkan kata yang sering muncul namun tidak mempunyai arti atau makna spesifik atau peranya tidak diperlukan dalam analisa teks. Tujuan dari tahap *stopword removal* adalah untuk mengurangi *noise* pada data yang akan diklasifikasikan. Proses *stopword removal* dapat menggunakan *pacakage library* sastrawi yang menyediakan daftar *stopword* dalam Bahasa Indonesia.

Formula E dan Jakarta international stadium jis telah menjadi magnet bagi elektabilitas dan popularitas anies

Formula E Jakarta international stadium jis magnet elektabilitas popularitas anies

3.7. *Word Cloud*

Word cloud adalah salah satu metode untuk menampilkan data teks secara visual. Secara umum, *word cloud* merupakan representasi grafis dari sebuah dokumen yang dilakukan dengan *plotting* kata-kata yang sering muncul pada ruang dua dimensi. Semakin sering kata muncul dalamm teks yang dianalisis, semakin besar ukuran kata yang muncul dalam gambar yang dihasilkan.



Gambar 3.3 Tampilan *Word Cloud*

3.8. Asosiasi Kata

Asosiasi kata merupakan salah satu metode untuk mencari sebuah nilai hubungan suatu kata. Nilai asosiasi dihitung berdasarkan kata-kata yang sering muncul serta dianggap penting oleh peneliti. Hasil asosiasi teks menunjukan besarnya nilai asosiasi antar kata dan seberapa sering kata-kata tersebut muncul bersamaan dalam satu kalimat. Semakin sering kata tersebut muncul bersamaann dalam satu kalimat maka, semakin besar pula nilai asosiasi. Terdapat beberapa nilai asosiasi menurut (Jonathan, 2006) yang dapat digunakan sebagai berikut:

- 0 : Tidak ada korelasi antara dua variabel
- 0 - 0,25 : Korelasi lemah
- 0,25 - 0,5 : Korelasi cukup
- 0,5 – 0,75 : Korelasi kuat

- 0,75-0,99 : Korelasi Sangat Kuat
- 1 : korelasi hubungan Sempurna

3.9. Pembobotan Fitur TF-IDF

Data setelah dilakukan *text preprocessing* masih berupa teks, data tersebut tidak dapat langsung diproses oleh beberapa algoritma *machine learning*. Hal itu disebabkan karena sebagian besar algoritma *machine learning* menerima masukan berupa numerik. Pemberian suatu nilai (dalam bentuk angka) untuk suatu *term*/kata disebut dengan pembobotan. Dalam penelitian ini dilakukan pembobotan dengan menggunakan algoritma *Term Frequency- Inverse Document Frequency* (TF-IDF). TF-IDF merupakan salah satu metode pembobotan yang sering digunakan. Mrtode ini terkenal efisien, mudah dan menghasilkan nilai yang akurat.

TF-IDF merupakan suatu hubungan kata (*term*) yang berada pada dokumen yang akan diberikan suatu nilai bobot (Sari, Ginting, Zebua, & Mesran, 2021). Metode ini menggabungkan dua konsep untuk perhitungan bobot yaitu frekuensi kemunculan sebuah kata didalam sebuah dokumen tertentu yang disebut dengan *term frequency* (TF) dan *inverse* frekuensi dokumen yang mengandung kata yang disebut *inverse document frequency* (IDF). Frekuensi kemunculan kata didalam dokumen yang diberikan menunjukan seberapa penting kata tersebut didalam dokumen. Hubungan antara sebuah kata dengan sebuah dokumen akan tinggi apabila frekuensi kata tinggi didalam dokumen dan frekuensi keseluruhan dokumen yang mengandung kata tersebut akan rendah pada kumpulan dokumen. Untuk menghitung nilai bobot menggunakan TF-IDF, proses yang dilakukan adalah sebagai berikut:

1. *Term Frequency* (TF)

Term frequency (TF) adalah frekuensi dari kemunculan sebuah *term* dalam dokumen yang bersangkutan. Semakin besar jumlah kemunculan suatu *term* (TF tinggi) dalam dokumen, maka semakin besar pula bobotnya atau akan memberikan nilai kesesuaian yang semakin besar.

Perhitungan *Term Frequency* (TF)

$$\boldsymbol{tf}_{\boldsymbol{i,j}} = \frac{\boldsymbol{n}_{\boldsymbol{i,j}}}{\sum \boldsymbol{n}_{\boldsymbol{i,j}}} \tag{3.1}$$

Keterangan :

$\boldsymbol{tf_{i,j}}$ = Frekuensi *term*

$\boldsymbol{n_{i,j}}$ = Banyaknya kata *i* dalam dokumen *j*

2. *Inverse Document Frequency* (IDF)

IDF *(Inverse Document Frequency)* merupakan ukuran kemampuan kata untuk membedakan kategori. IDF dalam sebuah kata dapat diperoleh dari jumlah total dokumen yang terdapat kata tersebut dibagi oleh jumlah total dokumen setelah hasil bagi logaritmik. Semakin sedikit jumlah dokumen yang mengandung *term* yang dimaksud, maka nilai IDF semakin besar.

Perhitungan *Document Frequency* (IDF)

$$idf = log \frac{N}{df_i} \tag{3.2}$$

Keterangan :

idf = *Inverse Document Frequency*

N = Jumlah semua dokumen dalam koleksi

df_i = Jumlah dokumen yang mengandung *term* (fi)

3. *Term Frequency- Inverse Document Frequency* (TF-IDF)

TF-IDF adalah ukuran statistik yang mengevaluasi seberapa relevan suatu kata dengan dokumen dalam kumpulan dokumen. TF-IDF merupakan salah satu metode pembobotan fitur yang banyak digunakan. Metode ini terkenal efisien, mudah dan mempunyai hasil yang akurat. TF-IDF dilakukan dengan cara mengalikan dua matrik yang dapat dipresentasikan TF-IDF = TF*IDF, dengan TF merupakan frekuensi kemunculan dari suatu term/kata dan IDF merupakan frekuensi *inverse* dokumen dari term/kata tersebut.

Dengan demikian perhitungan TF-IDF adalah sebagai berikut:

$$W_{i,j} = tf_{i,j}\ x\ idf_i$$

$$W_{i,j} = tf_{i,j}\ x\ log \frac{N}{df_i} \tag{3.3}$$

Keterangan :

$W_{i,j}$ = Bobot TF-IDF

idf_i = *Inverse Document Frequency*

$tf_{i,j}$ = Frekuensi suatu kata

TF-IDF dapat mengasosiasikan setiap kata dalam dokumen dengan angka yang mewakili seberapa relevan suatu kata dalam dokumen. Seberapa relevan kata dalam dokumen dapat digunakan dalam berbagai hal sebagai berikut:

- Pencarian informasi

  TF-IDF diciptakan untuk menelusuri dokumen dan dapat digunakan untuk memberikan hasil yang paling relevan dengan apa yang sedang ditelusuri. Contohnya adalah ketika seorang mencari "Ronaldo" dimesin pencariannya. Hasil yang akan ditampilkan dalam urutan relevansinya. Artinya artikel olahraga yang paling relevan akan diberikan peringkat lebih tinggi.

- Ekstraksi kata kunci

  TF-IDF juga berguna untuk mengekstrak kata kunci dari teks. Kata-kata dengan skor tertinggi dari sebuah dokumen adalah yang paling relevan dengan dokumen itu, dan oleh sebab itu dapat dianggap sebagai kata kunci untuk dokumen tersebut.

3.10. Analisis Sentimen

Analisis sentimen merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam kalimat opini (Brahimi, Touahria, & Tari, 2019). Tugas dasar yang dilakukan oleh analisis sentimen adalah mengelompokkan polaritas yang terdapat pada suatu dokumen teks, apakah pendapat yang dikemukakan bersifat negatif, positif, atau netral. Data yang diproses beragam, data tersebut dapat berupa *review* barang atau jasa, politik, orang sekelompok orang dan juga tentang suatu kebijakan. Analisis sentimen berfokus pada polaritas teks (positif, dan netral), namun juga dapat mendeteksi perasaan atau emosi seperti kemarahan, kebahagiaan, kesedihan dan lain-lain. Saat ini analisis sentimen banyak digunakan oleh peneliti sebagai salah satu cabang riset dalam ilmu komputer seiring dengan banyaknya informasi yang ada dimedia sosial.

3.11. Model Klasifikasi

Klasifikasi merupakan salah satu motode yang termasuk dalam kategori *supervised learning* yang mana tujuannya adalah memprediksi kelas label yang belum diketahui label kelasnya berdasarkan observasi dari data yang telah diketahui

labelnya. Oleh karena itu, untuk membuat model klasifikasi memerlukan pemagian data menjadi dua kelompok. Dua kelompok tersebut adalah sebagai berikut:

1. Data Latih (Data *Training*)

Model klasifikasi akan dibangun menggunakan data latih. Data latih akan menggeneralisasi dan mempelajari pola dan menerapkan pola ini kedalam data uji. Model akan menggunakan set data latih dan akan menghitung cara terbaik untuk memetakan data input kelabel kelas tertentu. Dengan demikian, set data latih harus mewakili masalah dan mempunyai banyak contoh untuk untuk setiap label kelas.

2. Data Uji (Data *Testing*)

Data uji digunakan untuk menguji performa model klasifikasi yang dihasilkan. Untuk menguji model klasifikasi, maka dilakukan proses evaluasi model klasifikasi.

Algoritma untuk melakukan klasifikasi ada banyak, terdapat beberapa algoritma yang banyak digunakan adalah *decision tree, naïve bayes classification* (NBC), K-*Nearest neighbour* (KNN), *suport vector machine* (SVM) dan lain sebagainya. Pada penelitian ini akan membandingkan dua algoritma yaitu *naïve bayes classification* (NBC) dan *suport vector machine* (SVM).

3.12. Naïve Bayes Classifier (NBC)

Algoritma naïve bayes adalah salah satu algoritma yang terdapat pada teknik data mining untuk klasifikasi. Menurut (Feldman & Sanger, 2007) , algoritma *naïve bayes classifier* (NBC) merupakan algoritma yang digunakan untuk mencari nilai probabilitas tertinggi untuk mengklasifikasi data uji pada kategori yan paling tepat. Algortima ini memanfaatkan teori peluang yang diperkenalkan oleh ilmuan inggris, Thomas Bayes. Cara kerja teori ini yaitu dengan memprediksi peluang terjadinya kejadian dimasa depan berdasarkan data sebelumnya. Pada teorema bayes mempunyai asumsi bahwa setiap varaiabel X bersifat bebas (*independence*), yang artinya sebuah variabel tidak ada kaitannya dengan varaibel yang lain (Rosi, Fauzi, & Perdana, 2017). Secara umum teorema bayes mempunyai persamaan seperti berikut (Indriati, Idris, & Fauzi, 2019):

$$P(B|A) = \frac{P(B)\ P(A|B)}{P(A)} \tag{3.4}$$

Dengan keterangan :

A = Data dengan *class* yang belum diketahui.

B = Hipotesis data A merupakan suatu *class* spesifik.

$P(B|A)$ = Probabilitas hipotesis B berdasar kondisi A (*conditional/ Posterior probability*).

$P(B)$ = Probabilitas hipotesis B (*prior probability*).

$P(A|B)$ = Probabilitas A berdasar kondisi pada hipotesis B.

$P(A)$ = Probabilitas dari A

Naïve Bayes Classifier (NBC) menerapkan konsep probabilitas yang dapat digunakan dalam penentuan kelompok kelas dokumen teks serta dapat mengolah data dengan jumlah besar dengan hasil akurasi yang tinggi (Roifa, 2018). Tingkat performa sistem klasifikasi dengan menggunakan *naïve bayes classifier* (NBC) bergantung pada data yang dimiliki serta data yang dipilih sebagai data latih. Jika data latih dapat mewakili semua atau seluruh data, maka model klasifikasi yang dibuat mempunyai performa yang bagus. Hal ini berlaku untuk sebaliknya, jika data latih kurang mewakili suatu dokumen, maka tingkat performa model klasifikasi kurang baik digunakan.

Algoritma *naïve bayes classifier* memberi nilai target kepada data baru menggunakan nilai V_{map}, yaitu nilai kemungkinan tertinggi dari seluruh anggota himpunan set domain V. Setiap data *tweet* dipresentasikan dengan pasangan atribut "$x_1, x_2, x_3, \dots, x_n$" dimana $x_1$adalah kata pertama, x_2 adalah kata kedua, x_3 adalah kata ketiga dan seterusnya. Sedangkan V adalah himpunan kategori sentimen. Pada saat klasifikasi, algoritma akan mencari probabilitas tertinggi dari semua kategori yang diujikan (V_{map}), dimana persamaan yang dapat dibangun sebagai berikut (Rizqiyani, Mulwinda, & Putri, 2017):

$$V_{map} = \arg max \frac{P(x_1, x_2, x_3, \dots, x_n|V_j).P(V_j)}{P(x_1, x_2, x_3, \dots, x_n)} \quad (3.5)$$

Untuk $P(x_1, x_2, x_3, \dots, x_n)$ nilainya konstan untuk semua kategori (V_j), sehingga persamaan yang dapat ditulis sebagai berikut:

$$V_{map} = \arg max \ P(x_1, x_2, x_3, \dots, x_n|V_j).P(V_j) \quad (3.6)$$

Dengan mengasumsikan bahwa setiap kata dalam $(x_1, x_2, x_3, \dots, x_n)$ adalah *independent*, maka $P(x_1, x_2, x_3, \dots, x_n|V_j)$ dalam persamaan 3.6 dapat ditulis sebagai berikut:

$$P(x_1, x_2, x_3, \dots, x_n|V_j) = \prod_{i=1} P(X_i|V_j) \tag{3.7}$$

Sehingga persamaan (3.6) dapat ditulis :

$$V_{map} = \arg max\ P(V_j) \prod_{i=1} P(X_i|V_j) \tag{3.8}$$

Keterangan :

V_j = kategori komentar j= 1,2,3,..,n. Dimana penulisan ini $j_1$kategori komentar sentimen positif, j_2 kategori komentar sentimen negatif, j_2 kategori komentar sentimen netral.

$P(X_i|V_j)$ = Probabilitas x_i pada kategori V_j.

$P(V_j)$ = Probabilitas dari (V_j).

Untuk $P(V_j)$ dan $P(X_i|V_j)$ dihitung pada saat pelatihan dimana persamaanya adalah sebagai berikut :

$$P(V_j) = \frac{|docs_j|}{|contoh|} \tag{3.9}$$

Dimana $|docs_j|$ adalah banyaknya dokumenn yang mempunyai kategori j dalam pelatihan, sedangkan $|contoh|$ banyaknya dokumen atau *tweet* dalam contoh yang digunakan untuk pelatihan. Untuk setiap probabilitas X_i untuk setiap kategori $P(X_i|V_j)$, dihitung pada saat *training*.

$$P(X_i|V_j) = \frac{n_k + 1}{n + |kosakata|} \tag{3.10}$$

Dimana n adalah jumlah kemunculan kata X_i dalam *tweet* atau dokumen yang berkategori V_j, sedangka n adalah banyaknya seluruh kata dalam *tweet* dengan kategori V_j dan $|kosakata|$ merupakan banyaknya kata dalam data latih (*training*).

Algoritma *naïve bayes classifier* mempunyai kelebihan yaitu mudah diimplementasikan serta waktu pengolahan yang cepat bila dibandingkan dengan metode klasifikasi lainnya. Sedangkan kekurangan pada metode ini adalah asumsi masing-masing variabel independent menyebabkan akurasi yang dihasilkan model

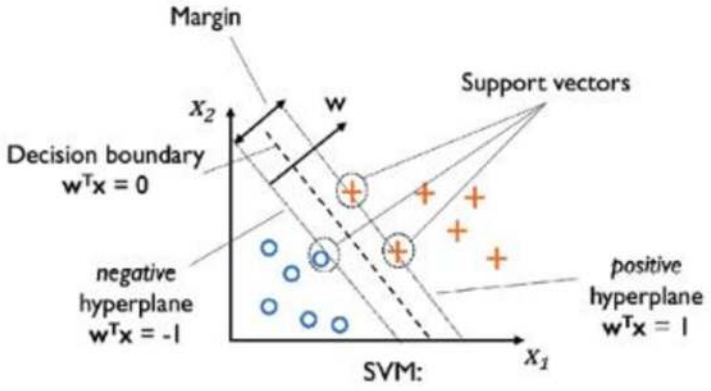
berkurang. Dalam kehidupan nyata, dapat dikatakan hampir tidak mungkin seperangkat variabel yang sepenuhnya independen.

3.13. Support Vector Machine (SVM)

Support vector machine (SVM) merupakan algoritma *supervised learning* yang digunakan untuk klasifikasi teks. *Support vector machine* (SVM) dikembangkan oleh Isabelle Guyon, Bernhard Boser dan Vladmir Vapnik pada tahun 1992 ketika diadakan di *Annual Workshop on Computional Learning Theory*. Klasifikasi menggunakan *support vector machine* (SVM) dilakukan dengan cara mencari *hyperplane* atau garis pembatas (*Decision Boundary*) yang memisahkan antara suatu kelas dengan kelas lainnya. SVM memiliki prinsip dasar *linear classifier* yang dapat bekerja pada kasus klasifikasi yang dipisahkan secara linear, namun juga dikembangkan agar dapat berkerja pada kasus *non-linear* dengan memasukkan konsep *kernel trick* pada ruang kerja berdimensi tinggi (Muttaqin & Kharisudin, 2021).

3.13.1 Kasus Data yang Terpisah secara Linear

Kasus data yang terpisah secara linear adalah penerapan metode SVM pada data yang dapat dipisahkan secara linear. Gambaran pada kasus data yang terpisah secara linear dapat dilihat padat gambar 2.3 berikut.



Gambar 3.4 Struktur *Support vector* Machine (Sumber : (Fransiska, Rianto, & Gufroni, 2020)).

Pada gambar 2.3 menjelaskan *hyperplane* digunakan untuk memisah data menjadi dua kelas dengan mengukur *margin hyperplane* tersebut dan mencari titik maksimalnya. *Margin* merupakan jarak terdekat dengan *pattern* terdekat dari masing-masing kelas. Sedangkan *pattern* yang paling dekat dengan *hyperplane* dinamakan *support vector*. Dalam klasifikasi SVM, *hyperplane* dengan *margin*

yang lebih besar menyebabkan akurasi klasifikasi lebih baik dibandingkan dengan margin yang lebih kecil. Mencari *margin* terbesar dikenal dengan *Maximum Marginal Hyperplane* (MMH). Data dinotasikan sebagai $\{x_i, x_{i+1}, \dots, x_n\}$ sedangkan label masing-masing kelas dinotasikan $y_i = \{+1, -1\}$untuk $i = 1,2,3, \dots, n$. Pada visualisasi bidang pemabatas pertama (persamaan 3.11) sedangkan bidang pembatas kedua membatasi kelas kedua (persamaan 3.12), sehingga diperoleh pertidaksamaan sebagai berikut:

$$x_i . w + b \geq +1, untuk\ y_i = +1 \tag{3.11}$$

$$x_i . w + b \leq -1, untuk\ y_i = -1 \tag{3.12}$$

Diketahui w merupakan bidang normal dan b merupakan posisi bidang relative terhadap pusat koordinat. *Margin* terbesar dapat ditemukan dengan memaksimalkan nilai jarak antara jarak dan titik terdekatnya, yaitu $\frac{1}{|w|}$ sama dengan meminimalkan $|w|^2$ dan jika kedua bidang pembatas direpresentasikan dalam pertidaksamaan (3.13).

$$y_i(x_i . w + b) - 1 \geq 0 \tag{3.13}$$

Sehingga dapat dirumuskan sebagai *quadratic programming* (QP) problem, yaitu mencari titik minimum persamaan (3.14).

$$\min \frac{1}{2}|w|^2$$

$$s.t y_i(x_i . w + b) - 1 \geq 0 \tag{3.14}$$

Permasalahan optimasi dapat diselesaikan dengan teknik komputasi, diantaranya dengan Formula lagrangian multiplier sehingga dinyatakan menjadi:

$$\min_{w,b} L_p(w, b, a) = \frac{1}{2}|W|^2 - \sum_{i=1}^{n} \alpha_i\, y_i((x_i . w + b) + \sum_{i=1}^{n} \alpha_i \tag{3.15}$$

$\alpha_i \geq 0$ merupakan nilai lagrange multiplier. Kemudian dengan meminimalkan L_p terhadap *w* dan *b*, maka $\frac{\partial}{\partial b} L_p(w, b, a) = 0$ diperoleh persamaan (6) dan dari $\frac{\partial}{\partial w} L_p(w, b, a) = 0$ diperoleh persamaan (3.17).

$$\sum_{i=1}^{n} \alpha_i\, y_i = 0 \tag{3.16}$$

$$w = \sum_{i=1}^{n} \alpha_i\, y_i x_i \tag{3.17}$$

Vector w bernilai besar hingga tak terhingga, tetapi nilai α_i terhingga, sehingga Formula lagrangian L_p (*Primal Problem*) diubah kedalam *dual problem* L_D. Dengan substitusi persamaan (7) ke L_p diperoleh L_D dengan konstain yang berbeda.

$$L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \, \alpha_j y_i y_j x_i x_j \tag{3.18}$$

Oleh sebab itu persoalan untuk picarian pemisah bidang terbaik dirumuskan sebagai berikut:

$$\min_{w,b} L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \, \alpha_j y_i y_j x_i x_j \tag{3.19}$$

$$s.t \sum_{i=1}^{n} \alpha_i \, y_i = 0, \qquad \alpha_i \geq 0 \tag{3.20}$$

Nilai α_i digunakan untuk menemukan w, setiap data latih memiliki niali α_i. Data latih yang mempunyai nilai $\alpha_i > 0$ merupakan *support vector* sedangkan sisanya mempunyai nilai $\alpha_i = 0$, sehingga fungsi keputusan yang dihasilkan hanya dipengaruhi oleh *support vector*.

Rumus untuk mencari *hyperplane* terbaik adalah dengan permasalahan *quadratic programming* sehingga nilai maksimum dari α_i selalu dapat ditemukan. Maka kelas pengujian x dapat ditentukan berdasarkan nilai fungsi keputusan:

$$f(x_d) = \sum_{i=1}^{ns} \alpha_i \, y_i x_i x_d + d \tag{3.21}$$

Keterangan :

x_d = Koefisien *Lagrange*.

x_i = *Support Vector.*

y_i = kelas data.

ns = Jumlah *Support Vector.*

x_d = Data yang Diklasifikasikan.

3.13.2 Kasus Data yang Tidak Terpisah secara Linear

Dalam klasifikasi banyak dijumpai bidang pemisah yang tidak dapat diambil dengan satu garis lurus. Klasifikasi yang tidak dapat dipisahkan secara linear memerlukan modifikasi pada Formula SVM dengan memanfaatkan kernel *trick*. Dengan kernel *trick*, akan dilakukan *mapping* data *input* ke *feature space* yang

dimensinya lebih tinggi sehingga membuat data *input* yang dihasilkan akan terpisah secara *linear* dan membentuk *hyperplane* yang optimal (Husada & Paramita, 2021). Beberapa kernel yang umum dipakai pada klasifikasi SVM adalah sebagai berikut (Praghakusma & Charibaldi, Juni 2021):

a. Kernel *Linear*

Kernel *Linear* merupakan fungsi kernel yang paling sederhana. Kernel ini banyak digunakan ketika data yang dianalisis sudah terpisah secara *linear*. Kernel *linear* sering digunakan untuk menyelesaikan masalah klasifikasi teks, karena mempunyai akurasi yang baik. Persamaan kernel *linear* sebagai berikut:

$$\mathrm{K}(x_i, x) = x_i^T x \tag{3.22}$$

b. Kernel *Polynomial*

Kernel *polynomial* merupakan fungsi kernel yang digunakan pada data yang tidak terpisah secara linear. Kernel *polynomial* baik digunakan untuk permasalahan ketika dataset *training* sudah dinormalisasi. Kernel *polynomial* memiliki persamaan seperti berikut:

$$\mathrm{K}(x_i, x_i) = \left(x_i . x_i^T\right)^d \tag{3.23}$$

c. Kernel *Radial Basic Function* (RBF)

Kernel *radial basic function* (RBF) atau *gaussian* merupakan fungsi kernel yang diimplementasikan ketika data tidak terpisah secara *linear*. Fungsi kernel ini memiliki dua parameter yaitu *gamma* dan *cost*. Parameter *cost* dirumuskan sebagai C berfungsi memaksimalkan SVM dalam menghindari kesalahan klasifikasi di setiap sampel dan dataset *training*. Parameter *gamma* memiliki fungsi untuk menentukan seberapa jauh pengaruh dari satu sampel *training* dataset pada garis pemisahnya. Berikut merupakan persamaan dari kernel *radial basic function* (RBF):

$$\mathrm{K}(x_i, x) = \exp(-\gamma |x_i - x|^2), \gamma > 0 \tag{3.24}$$

d. Kernel *Sigmoid*

Kernel sigmoid merupakan fungsi kernel yang digunakan ketika data tidak terpisah secara linear. Kernel ini merupakan pengembangan dari jaringan saraf tiruam yang mempunyai karakteristik kurva berbentuk "S". Kernel sigmoid dapat dinyatakan dalam persamaan berikut:

$$\mathrm{K}(x_i, x) = \tanh(\gamma x_i^T x + r) \tag{3.25}$$

3.14. Evaluasi Model

Setelah model klasifikasi dibangun, maka dilakukan proses evaluasi model klasifikasi. Evaluasi model merupakan sebuah proses untuk mengukur kinerja model klasifikasi yang dihasilkan. Salah satu cara untuk mengukur kinerja model klasifikasi adalah menggunakan *confusion matrix*.

3.14.1 Confusion Matrix

Confusion matrix merupakan matrik yang menyimpan informasi untuk mengetahui performa dari model yang digunakan dan digunakan sebagai acuan dari performa klasifikasi dari algoritma yang digunakan pada tahap evaluasi (Hidayatulloh , Yusuf, & dkk, 2019). Dalam *confusion matix* dapat diketahui bahwa model yang digunakan bekerja dengan baik atau tidak. Menurut (han & Kamber, 2011), *confusion matrix* merupakan alat yang berguna untuk menganalisis seberapa baik *classifier* mengenali *tuple* dari kelas yang berbeda. TP dan TN memberikan informasi ketika *classifier* benar, sedangkan FP dan FN memberitahu ketika *classifier* salah, *tuple* positif dikenali sebagai negatif dan tuple negatif dikenali sebagai positif. *Confusion matrix* berbentuk tabel matrik berukuran N x N dengan N merupakan jumlah kelas yang diprediksi (Daqiqil, 2021). Berikut merupakan tabel *confusion matrix*:

Tabel 3.1 *Confusion Matrix*

	Actual Positif	***Actual Negative***
Predicted Positif	*True Positife* (TP)	*False Positife* (FP)
Predicted Negative	*False Negative* (FN)	*True Negative* (TN)

Keterangan tabel diatas sebagai berikut:

a) *True Positife* (TP) : jumlah atau banyaknya data kelas aktual adalah positif namun kelas prediksinya merupakan kelas positif.
b) *False Negative* (FN) : jumlah atau banyaknya data kelas aktual adalah positif namun kelas prediksinya merupakan kelas negatif.
c) *False Positife* (FP) : jumlah atau banyaknya data kelas aktual adalah negatif namun kelas prediksinya merupakan kelas positif.
d) *True Negative* (TN) : jumlah atau banyaknya data kelas aktual adalah negatif namun kelas prediksinya merupakan kelas negatif.

Evaluasi kinerja dari suatu model klasifikasi dapat dilakukan mneggunakan pendekatan *confusion matrix* untuk memperoleh *accuracy, precision, recall*, F-1 *Score*. *Metric-metric* yang dapat digunakan adalah sebagai berikut (Azhari, Situmorang, & Rosnelly, 2021):

a. *Accuracy* (Akurasi)

Akurasi merupakan nilai rasio data *tweet* yang benar terdeteksi didalam pengujian. Dengan kata lain, akurasi adalah nilai yang menunjukan kedekatan anatara nilai prediksi sistem dengan nilai aktual.

Persamaan akurasi seperti berikut :

$$Akurasi = \frac{Prediksi\ Benar}{Prediksi\ benar + Prediksi\ Salah}$$

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.26}$$

b. *Precision* (Presisi)

Nilai *precision* (presisi) merupakan nilai sensitifitas atau nilai ketepatan sistem anatara informasi yang diberikan oleh sistem untuk menunjukan seberapa benar data kelas atau kelas positif.

Persamaan *precision* seperti berikut :

$$Precision = \frac{Jumlah\ Prediksi\ yang\ Benar\ untuk\ Kelas\ Positif}{Total\ Prediksi\ untuk\ Kelas\ Positif}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.27}$$

c. *Recall*

Recall merupakan nilai yang menunjukan tingkat keberhasilan dan sensitifitas model dalam menemukan kembali sebuah informasi.

Persamaan *recall* seperti berikut :

$$Recall = \frac{Jumlah\ Prediksi\ yang\ Benar\ untuk\ Kelas\ Positif}{Total\ Data\ untuk\ Kelas\ Positif}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.28}$$

d. F-1 *Score*

F-1 *Score* menggambarkan perbandingan rata-rata *precision* dan *recall* yang dibobotkan. *Accuracy* tepat kita gunakan sebagai acuan performa algoritma jika dataset memiliki jumlah data *false* negatif dan *false* positif yang sangat mendekati

(*Symmetric*). Namun jika jumlahnya tidak mendekati, maka sebaiknya kita menggunakan F-1 *Score* sebagai acuan.

Persamaan F-1 *Score* seperti berikut :

$$\mathrm{F}-1\ Score\ = 2 \times \frac{Recall\ \times\ Precision}{Recall +\ Precision} \tag{3.29}$$

BAB IV
METODOLOGI PENELITIAN

4.1. Populasi Penelitian

Populasi yang digunakan pada penelitian ini adalah seluruh data *tweet* atau kicauan yang berkaitan dengan Formula E melalui media sosial twitter. Jumlah sampel data yang digunakan adalah 1829 *tweet*s. Sampel tersebut diambil pada tanggal 23 Mei 2022 sampai 31 Mei 2022 dengan cara melakukan *scraping* menggunakan *software Rstudio*.

4.2. Variabel penelitian

Variabel yang digunakan pada penelitian ini ada tiga yaitu, *tweet*s atau kicauan, data terlabel, serta data klasifikasi. Berikut ini adalah definisi yang ditampilkan pada tabel 3.1.

Tabel 4.1 Variabel Penelitian

Variabel	Definisi
Data Teks	Data *tweet* atau kicauan yang dipublikasikan dimedia sosial twitter dengan *keyword* "Formula E"
Scoring	Penilaian yang diberikan terhadap tweet berdasarkan *stopword* positif dan negatif
Klasifikasi	Teks yang telah memiliki kategori sentimen , positif, dan netral.

4.3. Jenis dan Sumber Data

Data yang digunakan pada penelitian ini merupakan data sekunder mengenai *tweet*s atau kicauan terkait Formula E dari media sosial twitter. Data tersebut diperoleh dengan cara melakukan *scraping* dengan memanfaatkan API Key Twitter.

4.4. Metode Analisis Data

Proses analisis dalam penelitian ini menggunakan *Software Microsoft Excel* 2016, R*studio versi* 3.5.1. dan *software API Key Twitter*. penelitian ini menggunakan beberapa metode sebagai berikut:

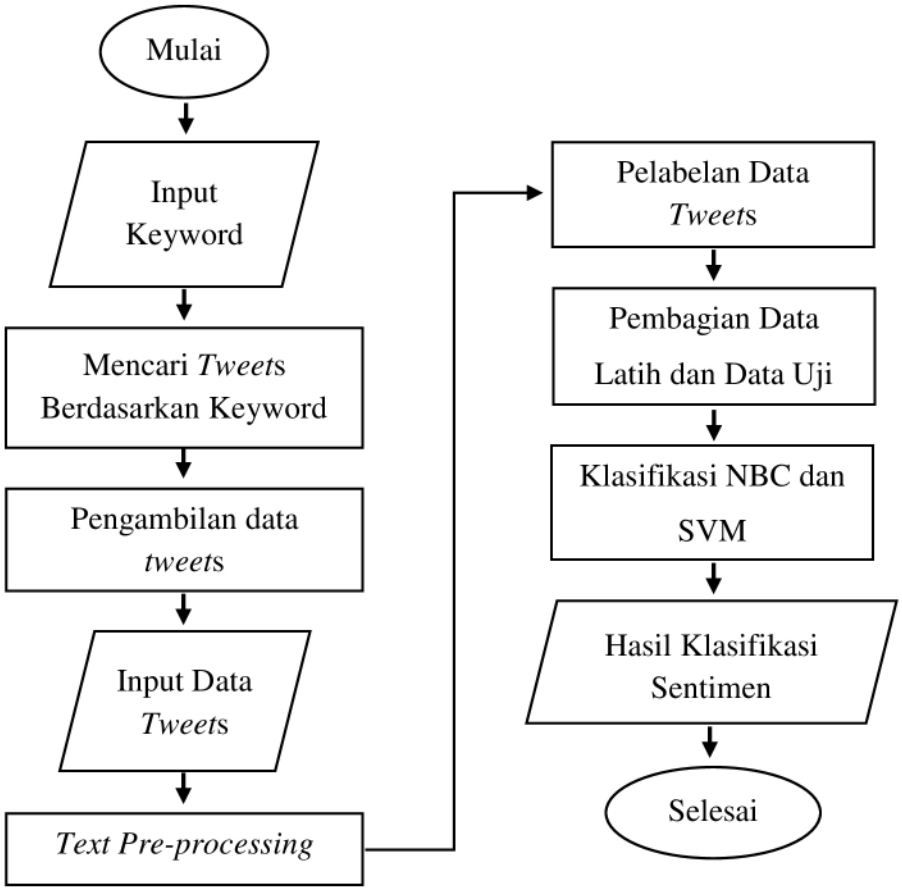
1. *Web scraping* digunakan untuk mengambil data *tweet*s pengguna media sosial twitter secara *online* dengan menggunakan API (*Application Programming Interface*).

2. *Text mining*, digunakan untuk mengubah data menjadi data yang lebih terstruktur sehingga diperoleh informasi data yang bernilai. *Text mining* digunakan pada data yang berjumlah besar.
3. *Sentimen analysis*, digunakan untuk melakukan klasifikasi pada data *tweet*, sehingga menghasilkan polaritas sentimen negatif atau sentimen positif.

Klasifikasi *machine learning* dengan menggunakan algoritma *Naïve Bayes Classification* dan *Support Vector Machine* yang digunakan utuk melakukan klasifikasi data *tweet* berdasarkan sentimen negatif atau sentimen positif serta melihat tingkat akurasi dalam melakukan klasifikasi teks.

4.5. Tahapan Penelitian

Langkah atau tahapan melakukan analisis dapat digambarkan dalam *flowchart* melalui gambar 4.1 berikut:



Gambar 4.1 *Flowchart*

Penjelasannya sebagai berikut:

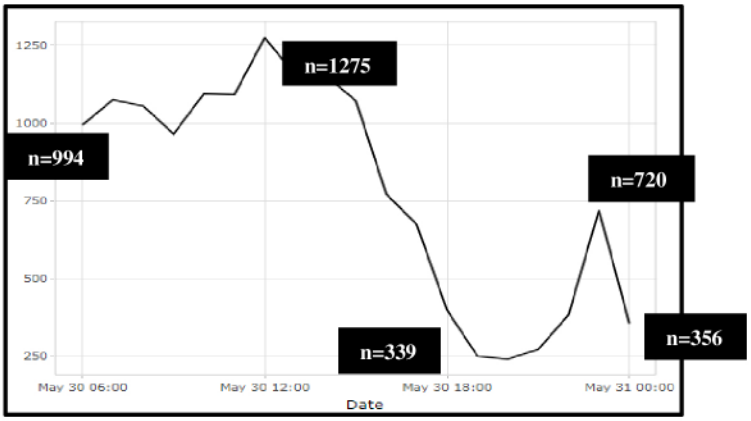
1. Langkah pertama adalah mengumpulkan data *tweet* dari media sosial twitter. pengumpulan data dalam penelitian ini dilakukan dengan cara *scraping* pada media sosial twitter dengan topik yang telah ditentukan sebelumnya.
2. Data yang dikumpulkan, kemudian masuk kedalam tahap *text preprocessing*. Pada tahap ini dilakukan *cleaning*, *case folding*, tokenizing, dan *stemming*. *Text pre-pocessing* dilakukkan untuk mengubah data menjadi lebih terstruktur.
3. Setelah mendapatkan kumpulan kata-kata, langkah berikutnya yaitu melakukan *sentiment analysis*. Pada tahap ini kata-kata akan diklasifikasikan kedalam sentimen positif atau negatif . Pelabelan kalimat positif dan dibuat dengan menggunakan kamus *lexicon*.
4. Tahap selanjutnya yaitu melakukan klasifikasi *Naïve Bayes Classificaion* dan *Support Vector Machine*. Namun terlebih dahulu melakukan pembagian data, data dibagi menjadi dua yaitu data latih (Data *Training*) dan data uji (Data *Testing*). Pada penelitian ini menggunakan perbandingan data latih 80 % dan data uji 20 %.
5. Langkah selanjutnya yaitu melakukan perbandingan hasil klasifikasi masing-masing algoritma *Naïve Bayes Classificaion* dan *Support Vector Machine*.

BAB V
HASIL DAN PEMBAHASAN

Berdasarkan kajian penelitian sebelumnya, maka pada bab ini akan menjelaskan mengenai implementasi dari analisis sentimen yang menggunakan algoritma *naïve bayes classifier* (NBC) dan *support vector machine* (SVM).

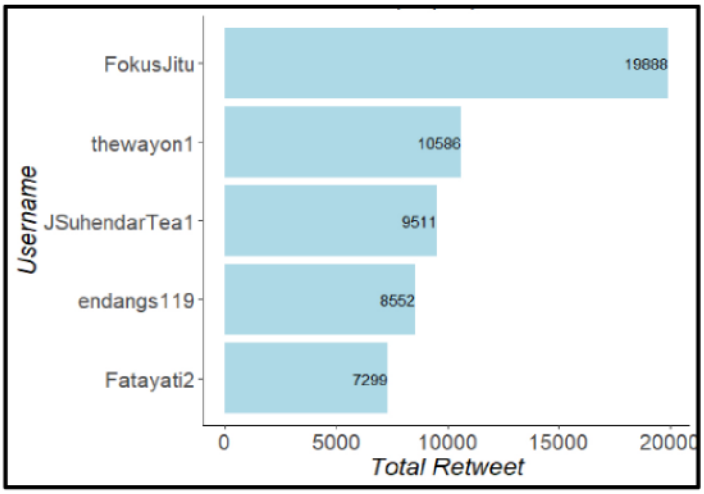
5.1. Analisis Deskripitif

Analisis deskriptif merupakan suatu metode analisis statistik yang bertujuan memberikan gamabaran mengenai subjek penelitian berdasarkan data variabel yang diperoleh. Analisis deskriptif dapat ditampilkan dalam bentuk tabel distribusi frekuensi, tabel histogram, nilai rata-rata, nilai standar deviasi dan lain-lain. Dibawah ini merupakan hasil frekuensi *tweet* berdasarkan jumlah *tweet* per jam dengan jumlah *tweet* sebanyak 15.000.



Gambar 5.1 Frekuensi *Tweet* Per Jam

Grafik diatas merupakan *line chart* data dari frekuensi *tweet* mengenai Formula E per jam. Dari grafik tersebut dapat diketahui bahwa pada 30 Mei pukul 12:00 merupakan jumlah *tweet* tertinggi dari data dengan jumlah *tweet* per jam sebanyak 1275 *tweet*. Jumlah *tweet* terendah pada 31 Mei pukul 00:00 dengan jumlah *tweet* per jam sebanyak 356 *tweet*. Banyaknya *tweet* mengenai Formula E pada 30 Mei diakibatkan dengan adanya kejadian mengenai robohnya atap tribun Formula E.

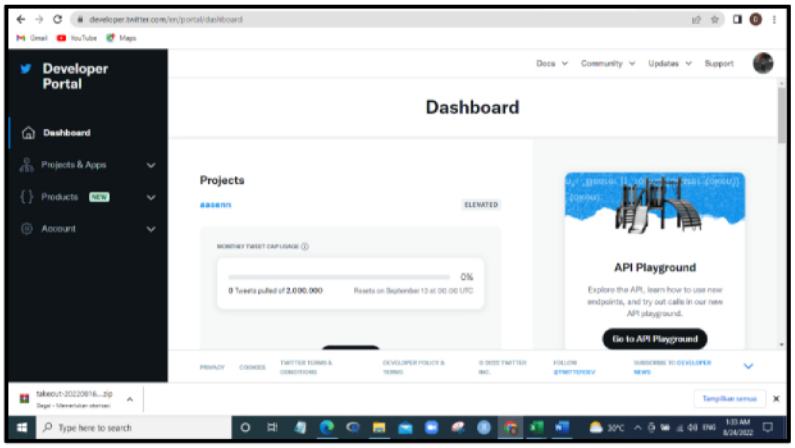


Gambar 5.2 Top 5 *Users Retweet*.

Grafik 5.2 merupakan grafik yang menunjukan banyaknya *tweet* yang yang diposting kembali dan dapat menambahkan komentar anda sendiri mengenai berita atau ulasan mengenai Formula E. Berdasarkan grafik tersebut dapat diketahui bahwa username dengan nama "fokusjitu" di*retweet* oleh 19888 akun twitter dan menjadi peringkat pertama. Peringkat kedua diperoleh dengan username "thewayon1" memperoleh *retweet* sebesar 10586. Peringkat ketiga dengan username "jsuhendartea1" memperoleh *retweet* sebesar 9511. Peringkat keemppat dengan username "engdangs119" memiliki *retweet* sebesar 8552. Sedangkan peringkat kelima dengan usename "fatayati2" memperoleh *retweet* sebesar 7299.

5.2. Mengumpulkan Data

Hal yang dibutuhkan untuk dapat mengakses data twitter adalah kode API twitter. API merupakan singkatan dari *application programming interface*. API digunakan untuk mengambil data dari twitter dan mengolahnya. Untuk mengakses layanan API tersebut anda harus mempunyai akun twitter dan twitter *developer*. Untuk memperoleh akun twitter *developer* harus melakukan *apply for access* dengan mengisi *form* registrasi. *Form* registrasi berisi pertanyaan terkait tujuan serta motivasi mendaftar akun *developer*. Setelah akun *developer* disetujui maka kita dapat membuat akses API twitter.



Gambar 5.3 Dashboard *Twitter Developer* (Sumber: Twitter Developer)

API twitter diakses melalui sebuah token. Token tersebut mempunyai sebuah fungsi yang hampir sama dengan *username* dan *password*. *Token* API berupa Api *Key*, Api *Secret Key* dan *Bearer Token*. Ketiga token ini yang digunakan untuk mengakses data twitter.

5.2.1 Scraping Data Twitter

Pada saat *scraping* atau pengambilan data kata kunci yang digunakan adalah "Formula+E" dengan menggunakna *syntax* "`tweets= searchTwitter('Formula+E',n = 15000, retryOnRateLimit = 10000, since="2022-05-23",until="2022-05-31")`".

Tabel 5.1 Data Penelitian

No	*text*	Created	*Retweet* Count
1	RT @aniesfor2024: Kesempatan terbaik disia siakan BUMN yg belum menjadi SPONSOR FORMUA E. Ada lebih dr 170 negara akan siarkan scr langsung…	5/30/2022 23:59	402
2	RT @tatakujiyati: Finally he got it right. Anggota PDIP, Gembong Warsono, mengatakan perhelatan Formula E harus tetap berkelanjutan. Menuru…	5/30/2022 23:59	116
3	RT @KOmengKO_25: Balapan GP Mandalika gw ga nonton. Balapan Formula E Ancol gw ga nonton. Balapan ESEMKA berapa pun tiketnya & dimana pun…	5/30/2022 23:59	98

No	*Text*	*Created*	*Retweet Count*
4	RT @ZettaZahra2: Atap Tribun E Formula tinggi & ada di lapangan (ruang terbuka). Tenda2 warung makanan kan rendah & di sekitarnya ada ban…	5/30/2022 23:59	3
5	Polisi Umumkan Penyebab Atap Tribun Formula E Ambruk https://t.co/pzUU4HoLLC #TempoOtomotif	5/30/2022 23:59	1

Tabel 5.1 merupakan beberapa hasil *scraping* data twitter mengenai Formula E. yang diperoleh berjumlah 3543 *tweets* dimulai dari tanggal 23 Mei 2022 sampai 31 Mei 2022. Tabel diatas terdiri dari *text, created* dan *retweet count*. Data *text* berisikan *tweet* atau opini yang dituliskan oleh pengguna twitter. Data *created* merupakan waktu *tweet* dibuat, berisikan tanggal dan waktu pembuatan. Sedangkan *retweet count* merupakan data yang berisikan banyaknya pengguna twitter yang ikut dalam menyebarkan ulang *tweet* atau opini tersebut.

5.3. Representasi Model

Representasi model dilakukan dengan membuat model dari data yang berbentuk kata-kata agar dapat diolah dan dihitung. Representasi melakukan penghitungan skor sentimen dan memberikan label kelas sentimen berdasarkan skor yang ada. Setiap opini twitter akan diberikan label sentimen berdasarkan skor sentimen yang telah dihitung.

5.3.1 Perhitungan Skor Sentimen

Perhitungan skor sentimen dengan contoh kalimat sebagai berikut :

Tabel 5.2 Perhitungan Skor Sentimen

Ulasan	Kata Positif	Kata Negatif
atap tribun arena sirkuit **roboh** akibat terjang **badai** lempar **senyum**	senyum	Roboh Badai
jumlah	1	2

Berdasarkan opini pada tabel 5.7 diketahui bahwa terdapat 2 kata negatif dan 1 kata positif berdasarkan kamus *lexicon*. Kata yang terdeteksi yaitu roboh dan badai. Sedangkan kata positif yang terdeteksi adalah senyum. Oleh karena itu perhitungan skor sentimen adalah:

$$\text{Skor} = \text{Jumlah Kata Positif} - \text{Jumlah Kata Negatif}$$

$$\text{Skor} = 1 - 2 = -1$$

Dalam penelitian ini, perhitungan skor sentimen dilakukan secara otomatis dengan bantuan Rstudio. Setelah menghitung skor sentimen selanjutnya adalah memberikan label sentimen pada setiap opini.

5.3.2 Pelabelan Kelas Sentimen

Setelah mendapatkan skor sentimen, pengolahan data selanjutnya yaitu memberikan pelabelan kelas sentimen. Pelabelan dilakukan dengan membagi data menjadi tiga kelas yaitu sentimen positif, dan netral dengan ketentuan sebagai berikut :

- Sentimen negatif = skor < 0
- Sentimen positif = skor > 0
- Sentimen netral = skor = 0

Kelas positif merupakan pernyataan yang mengandung dukungan, pujian, dan sebagainya. Kelas negatif merupakan pernyataan yang mengandung kritikan, hinaan terhadap diadakannya Formula E. Sedangakan kelas netral merupakan pernyataan yang tidak mengandung kata yang positif maupun , atau yang bersifat seimbang. Kelas sentimen netral dapat terjadi apabila dua kemungkinan yaitu :

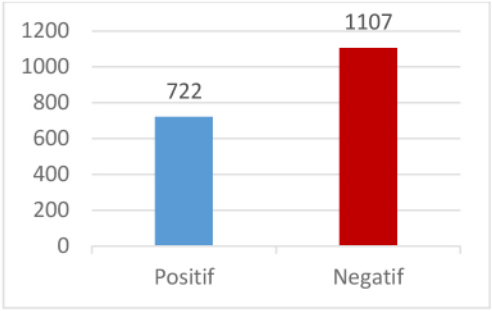
1. Tidak mengandung kata yang terdapat dalam kamus *lexicon*.
2. Jumlah skor positif dan negatif sama sehingga skor sentimen = 0

Berikut merupakan contoh pelabelan kelas sentimen, yang terdapat dalam tabel 4.7 dibawah ini:

Tabel 5.3 Contoh pelabelan Kelas Sentimen

Ulasan	Skor	Kelas Sentimen
atap tribun arena sirkuit **roboh** akibat terjang **badai** lempar **senyum**	-1	Negatif
harum nama bumn	1	Positif
rakyat makan	0	Netral

Hasil pelabelan kelas sentimen pada opini Formula E di media sosial twitter adalah sebagai berikut:



Gambar 5.4 Hasil Pelabelan Kelas Sentimen.

Berdasarkan grafik 5.4 diperoleh bahwa kategori sentimen positif terhadap Formula E sebesar 722 *tweet*. Sedangkan sentimen negatif terhadap Formula E seebesar 1107 *tweet*.

5.3.3 Sentimen Positif dan Negatif

Jakarta menjadi tuan rumah balapan Formula E pertama kali pada 4 Juni 2022. Dalam perjalanannya Formula E Jakarta banyak mendapat perhatian publik. Berawal dari berhasilnya Anies Baswedan dan petinggi Formula E membuahkan kesepakatan bahwa Jakarta menjadi tuan rumah Formula E pada 2020. Rencana tersebut tidak terlaksana akibat pandemi COVID-19 dan Jakarta tidak masuk dalam kalender Formula E. Penundaan Formula E juga diakibatkan oleh perubahan pembangunan sirkuit Formula E yang direncanakan dibangun dikawasan Monument Nasional. Penundaan berakhir pada 22 desember 2021, komite pelaksana Formula E Jakarta yang dipimpin Ahmad Sahroni secara resmi menunjuk kawasan Ancol sebagai sirkuit Formula E. Keputusan tersebut menggugurkan empat kandidat lokasi lain diantaranya Sudirman-Thamrin, Senayan, Pantai Maju, dan Jakarta International Expo. Penunjukan Ancol sebagai sirkuit juga tidak lantas bebas dari perhatian publik. Banyak sekali masyarakat yang memberikan opini terhadap Formula E dari perencanaan, persiapan sampai berakhirnya Formula E Jakarta. Berikut ini merupakan contoh sentimen negatif dan positif terhadap Formula E.

Tabel 5.4 Contoh Sentimen positif

Ulasan	Kelas Sentimen	Ulasan Setelah Cleaning
tiket Formula e laris manis buah dari getirnya perjuangan anies	Positif	tiket laris manis buah getir juang

Ulasan	Kelas Sentimen	Ulasan Setelah Cleaning
kita doakan bersama semoga anis baswedan tidak lengah hajatan besar Formula e berjalan aman dan lancar	Positif	Kita doa sama moga tidak lengah hajat besar jalan aman lancar
ya allah pak anis dan semua anak bangsa terlibat di Formula e mugi dirahmati dan diridloi allah	Positif	Allah semua anak bangsa rahmati ridho

Tabel 5.9 merupakan contoh ulasan terhadap Formula E yang masuk kedalam sentimen positif. Berdasarkan ulasan tersebut terdapat beberapa kata yang menunjukan orang, Lembaga, organisasi atau pihak-pihak terkait dalam penyelenggaraan Formula e yang sengaja dihilangkan. Kata-kata yang dihilangkan salah satunya yaitu bapak Anies Baswedan, Ancol, Mandalika, bapak Sandiaga Uno dan lain sebagainya. Hal tersebut dilakukan agar tidak memihak siapapun dan tidak menjadikan penelitian ini sebagai alat untuk menjatuhkan lembaga, orang, dan sebagainya. Penyelengaraan Formula E mengalami beberapa kali penundaan akibat pandemi COVID-19, perubahan letak sirkuit dan lain-lain. Hal tersebut merupakan salah satu faktor yang membuat Formula E banyak mendapat perhatian publik. Salah satu orang yang dikaitkan dengan Formula E adalah bapak Anis Baswedan. Dalam beberapa *tweet* banyak yang mengapresiasi bapak Anis Baswedan dan pihak-pihak terkait dengan terselenggaranya Formula E di Jakarta. Salah satu *tweet* yang menyatakan dukungannya yaitu kita doakan bersama semoga bapak Anis Baswedan tidak lengah hajatan besar Formula E berjalan aman dan lancar. Selain sentimen positif, terdapat juga *tweet* yang mengandung kritikan terhadap pagelaran Formula E. Berikut adalah contoh *tweet* yang mengandung sentimen negatif.

Tabel 5.5 Contoh Sentimen Negatif

Ulasan	Kelas Sentimen	Ulasan setelah cleaning
bumn sponsor said duga bumn jadi alat kepentingangan politik	Negatif	Sponsor alat penting politik
atap tribun roboh legislator pdip nilai terlalu dipaksakan	Negatif	Atap tribun roboh legislator nilai paksa

Ulasan	Kelas Sentimen	Ulasan setelah cleaning
Formula e adalah ambisi pribadi mewakili negara	Negatif	Ambisi pribadi mewakili negara
cebong hobinya busuk doang giliran mandalika mujinya setinggi langit ketika Formula e cuma bisa nyac	Negatif	Hobi busuk puji tinggi langit

Dalam sentimen negatif banyak yang membahas mengenai atap tribun yang roboh dikarenakan angin kencang dikawasan Ancol. pengguna twitter mengutarakan kekhawatiran dengan membuat *tweet* seperti penonton wajib memakai helm saat Formula E kepala benjol kejatuhan atap resiko ditanggung sendiri. Kejadian tersebut terjadi pada Jum'at, 27 Mei 2022 malam yang juga dibenarkan oleh direktur utama PT Jakarta Propertindo (Jakpro) Widi Amanasto. Kerusakan hanya terjadi pada satu *grandstand* saja serta tidak ada korban jiwa (CNN Indonesia, 2022). Ketua Panitia Formula E Ahmad Sahroni mengatakan atap tribun Formula e yang rusak langsung diperbaiki dan akan diselesaikan pada 2 Juni 2022.

Selain atap roboh, pengguna twitter juga mengomentari mengenai sponsor Formula e yaitu heineken. Heineken merupakan sponsor global Formula E yang berasal dari Belanda. Sponsor tersebut banyak dikritik karena dianggap tidak menghormati budaya Indonesia yang sebagaian besar beragama islam. Hal tersebut ditanggapi oleh Ahmad Sahroni selaku ketua pelaksana yang menyatakan bahwa tidak ada penjualan minuman beralkhohol dan tidak ada kampanye minuman berakhohol tersebut (Detik Otosport, 2022). Penyelenggara juga menyakinkah bahwa logo yang berkaitan dengan heineken tidak akan muncul sepanjang Formula E Jakarta.

Selain sponsor tersebut, BUMN juga menjadi salah satu topik yang banyak dikritik. Pengguna twitter membandingkan dengan pagelaran motor GP yang mendapat sponsor dari BUMN dan mendapat banyak dukungan dari pemerintah. Hal ini terbanding terbalik dengan Formula E yang tidak mempunyai sponsor dari BUMN dan kurangnya dukungan dari pemerintah. Salah satu *tweet* menyatakan BUMN menjadi alat kepentingan politik. Oleh sebab itu, peneliti menghapus beberapa kata yang berkaitan dengan lembaga, instansi, orang ataupun pihak-pihak

terkait. Hal tersebut dilakukan agar tidak memihak siapapun serta tidak menjadikan penelitian ini sebagai sarana untuk menjatuhkan orang, istansi dan sebagainya.

5.3.4 Visualisasi Data

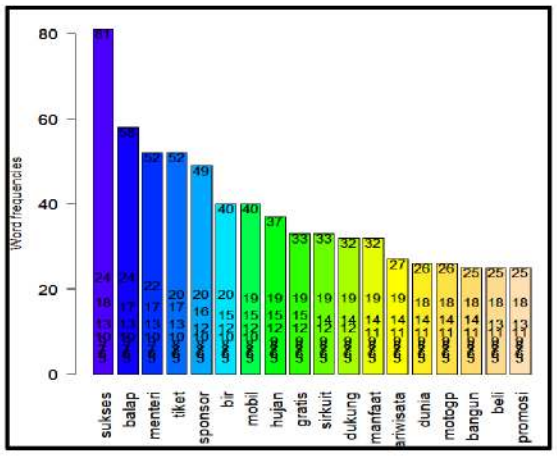
Kumpulan kata-kata yang sering muncul dapat divisualisasikan dalam bentuk *word cloud*. *Word cloud* memberikan gambaran terhadap kata-kata atau topik yang sering digunakan pengguna dalam memberikan opini terhadap Formula E. semakin tinggi frekuensi kata tersebut, maka semakin besar ukuran kata yang muncul dalam gambar yang dihasilkan. Pada gambar 5.2 merupakan *word cloud* dari keseluruhan *tweet* tentang Formula E.

Gambar 5.5 *Word Cloud* Seluruh *Tweet*.

Pada gambar 5.5 merupakan *word cloud* seluruh *tweet* Formula E. dari *word cloud* diketahui bahwa kata atau topik yang banyak dibicarakan mengenai Formula E melalui *platform* twitter yaitu sponsor, balap, tiket, roboh, tribun, dan lain sebagainya. Kemudian untuk sentimen dan sentimen positif dijelaskan dibawah ini.

a. Sentimen Positif

Sentimen positif merupakan pernyataan yang berisi dukungan atau pujian terhadap topik yang dibicarakan. Sentimen positif diperoleh melalui pelabelan data berdasarkan kamus *lexicon* ataupun manual. Hasil ekstrasi tersebut diidentifikasi berdasarkan frekuensi kata dalam ulasan. Berikut merupakan visualisasi sentimen positif mengenai Formula E.



Gambar 5.6 Frekuensi Kata sentimen Positif.

Gambar 5.6 merupakan visualisasi sentimen positif Formula E dengan mengunakan *barplot*. *Barplot* tersebut menunjukan seberapa banyak kata-kata atau topik yang paling banyak muncul. *Barplot* sentimen positif menunjukan kata "sukses" mempunyai frekuensi 81 kali, kata "balap" mempunyai frekuensi 58 kali dan seterusnya. Selain *barplot*, visualisasi sentimen positif juga dapat dilakukan dengan menggunakan *word cloud*. Berikut hasil visualisasi menggunakan *word cloud*:

Gambar 5.7 *Word Cloud* Ulasan Positif

Word cloud memberikan gambaran yang jelas mengenai kata-kata positif yang banyak digunakan. Pada *word cloud* semakin besar ukuran kata menunjukan kata tersebut paling banyak muncul serta mempunyai frekuensi yang besar. Dari *word cloud* tersebut diketahui kata-kata atau topik yang banyak dibahas mengenai Formula E adalah sukses, sponsor, balap, tiket, dan sebagainya. Selanjutnya, dilakukan pencarian asosiasi kata. Asosiasi kata diperoleh dengan menggunakan nilai akurasi terhadap kemungkinan suatu kata muncul bersamaan dengan kata lainya. Berdasarkan kata-kata yang banyak muncul memberikan asosiasi antar

kata pada masing-masing kelas sentiman secara bersamaan guna memperoleh informasi. Adapun asosiasi kata dari sentimen positif sebagai berikut.

Tabel 5.6 Asosiasi kata

Sukses	Promosi	Dukung
Alhamdulillah (0,40)	Menteri (0,37)	Syarat (0,37)
Axiata (0,28)	Pariwisata (0,31)	Tranparan (0,32)
Doa (0,20)	Listrik (O,24)	Cukai (0,22)
Reputasi (0,16)	Energi (0,17)	Optimal (0,22)

Berdasarkan tabel 5.11 penulis memilih beberapa kata yang cukup sering disebutkan untuk sentimen positif yaitu "sukses", "promosi", "dukung" yang menyangkut mengenai Formula E. kata "sukses" berasosiasi dengan kata "alhamdulillah", axiata, "doa", dan "reputasi" dengan nilai asosiasi ≥ 0,16. Berdasarkan asosiasi tersebut, diperoleh informasi tentang Formula E bahwa diharapkan Formula E dilaksanakan dengan sukses serta menbawa reputasi yang baik.

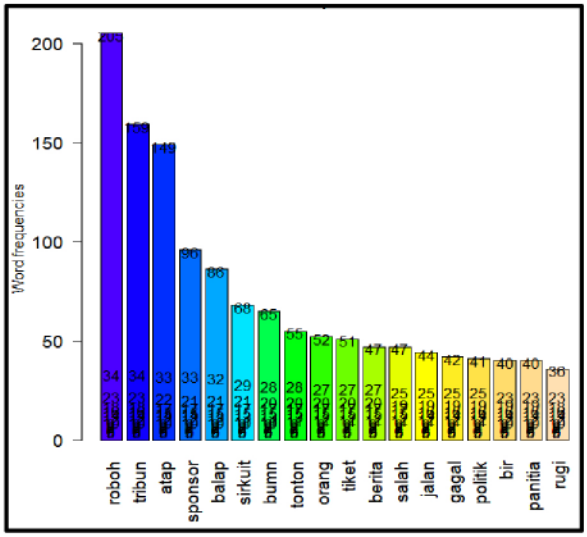
Kata "promosi" berasosiasi dengan kata seperti "menteri","pariwisata","Listrik", dan "energi" dengan nilai asosiasi ≥ 0,17. Berdasarkan asosiasi tersebut memberikan informasi bahwa Formula E memperoleh dukungan yang optimal dari cukai di Indonesia.

Kata "dukung" berasosiasi dengan kata seperti "syarat","transparan","cukai", dan "optimal" dengan nilai asosiasi ≥ 0,22. Berdasarkan asosiasi tersebut memberikan informasi bahwa Menteri pariwisata juga harus ikut mempromosikan Formula E tidak hanya sirkuit mandalika.

b. Ulasan Negatif

Sentimen negatif merupakan pernyataan yang mengandung kritikan, hinaan terhadap diadakannya Formula E. Sentimen negatif diperoleh melalui pelabelan data berdasarkan kamus *lexicon* ataupun manual. Hasil ekstrasi tersebut

diidentifikasi berdasarkan frekuensi kata dalam ulasan. Berikut merupakan visualisasi sentimen negatif mengenai Formula E.



Gambar 5.8 Frekuensi Kata Sentimen Negatif.

Gambar 5.8 merupakan visualisasi sentimen negatif Formula E dengan mengunakan *barplot*. *Barplot* tersebut menunjukan seberapa banyak kata-kata atau topik yang paling banyak muncul. *Barplot* sentimen negatif menunjukan kata "roboh" mempunyai frekuensi 205 kali, kata "tribun" mempunyai frekuensi 159 kali dan seterusnya. Selain *barplot*, visualisasi sentimen negatif juga dapat dilakukan dengan menggunakan *word cloud*. Berikut hasil visualisasi menggunakan *word cloud*:

Gambar 5.9 *Word Cloud* Ulasan Positif

Word cloud memberikan gambaran yang jelas mengenai kata-kata negatif yang banyak dibahas. Pada *word cloud* semakin besar ukuran kata menunjukan kata tersebut paling banyak muncul serta mempunyai frekuensi yang besar. Berdasarkan *word cloud* tersebut diketahui kata-kata atau topik yang banyak dibahas mengenai

Formula E adalah tribun, sponsor, roboh, bodoh, dan sebagainya. Selanjutnya, dilakukan pencarian asosiasi kata yang sering muncul secara bersamaan dan diperoleh hasil sebagai berikut:

Tabel 5.7 Asosiasi kata

Roboh	Tribun	bir
Atap (0,65)	Atap (0,69)	Haram (0,42)
Angin (0,24)	tonton (0,25)	Sponsor (0,30)
Sirkuit (0,20)	Khawatir (0,17)	Minum (0,18)
Badai (0,16)	Patah (0,15)	Heineken (0,18)

Berdasarkan tabel 5.12 penulis penulis memilih beberapa kata yang cukup sering dibahas terkait Formula E yaitu "roboh", "tribun", "bir". Kata "roboh" berasosiasi dengan kata seperti "atap", "angin", "sirkuit", dan "badai" dengan nilai asosiasi ≥ 0,16. Berdasarkan asosiasi tersebut, dapat diperoleh informasi bahwa sirkuit mengalami insiden badai angin yang menyebebkan atap tribun roboh.

Kata "tribun" berasosiasi dengan kata seperti "atap", "tonton", "khawatir", dan "patah" dengan nilai asosiasi ≥ 0,15. Berdasarkan asosiasi tersebut, dapat memberikan informasi bahwa tribun sirkuit patah dan roboh sehingga membuat khawatir penonton.

Kata "bir" berasosiasi dengan kata seperti "haram", "sponsor", "Minum", dan "Heineken" dengan nilai asosiasi ≥ 0,18. Berdasarkan asosiasi tersebut, dapat diperoleh informasi bahwa sponsor global Formula E salah satunya dari merk bir yaitu heineken yang dianggap haram oleh sebagian pengguna madia sosil.

5.4. Pengolahan Data

Data yang telah di*scraping* dari twitter selanjutnya diolah agar diperoleh informasi yang dibutuhkan. Data yang digunakan berupa opini masyarakat terkait Formula E di media sosial twitter sebanyak 1829 *tweet*. *tweet* tersebut diambil pada tanggal 23-31 Mei 2021 dengan cara melakukan *scraping* menggunakan *software Rstudio*. Pengolahan data tersebut diawali dengan *text preprocessing*.

5.4.1 *Text Prepocessing*

Data hasil *scraping* memiliki format yang sangat tidak terstruktur, sehingga data atau informasi tidak dapat diekstrak secara langsung. Oleh karena itu dibutuhkan *text preprocessing* yang bertujuan agar data menjadi lebih terstruktur. Dalam *text mining* sebegian waktu digunakan untuk persiapan data, meliputi membersihkan, mentransformasikan, dan mengatur ulang data. Tidak ada aturan yang pasti tentang tahapan dalam *text preprocessing*. Hal ini tergantung dari jenis data dan hasil yang diinginkan. Tahapan *preprocessing* merupakan tahap terpenting karena dapat mempengaruhi hasil analisis. Berikut tahapan *text preprocessing* yang dipakai dalam analisis:

a) Case Folding

Tahap *case folding* merupakan proses penyeragaman huruf yang terdapat dalam data. Dalam hal ini adalah mengubah huruf besar menjadi huruf kecil pada semua data. Tabel berikut merupakan contoh data *tweet*s hasil *case folding*.

Tabel 5.8 Contoh Data *Tweet*s Hasil *Case Folding*

Text	Hasil *Case folding*
RT @aniesfor2024: Kesempatan terbaik disia siakan BUMN yg belum menjadi SPONSOR FORMUA E. Ada lebih dr 170 negara akan siarkan scr langsung…	rt @aniesfor2024: kesempatan terbaik disia siakan bumn yg belum menjadi sponsor formua e. ada lebih dr 170 negara akan siarkan scr langsung…
RT @tatakujiyati: Finally he got it right. Anggota PDIP, Gembong Warsono, mengatakan perhelatan Formula E harus tetap berkelanjutan. Menuru…	rt @tatakujiyati: finally he got it right. anggota pdip, gembong warsono, mengatakan perhelatan Formula E harus tetap berkelanjutan. menuru…
RT @KOmengKO_25: Balapan GP Mandalika gw ga nonton. Balapan Formula E Ancol gw ga nonton. Balapan ESEMKA berapa pun tiketnya & dimana pun…	rt @komengko_25: balapan gp mandalika gw ga nonton. balapan Formula E ancol gw ga nonton. balapan esemka berapa pun tiketnya & dimana pun…

b) Cleaning Data

Cleaning merupakan proses membersihkan data dari karakter-karakter yang tidak digunakan dalam analisis. Proses ini sangat penting dengan tujuan untuk mendapatkan data yang sesuai untuk diolah pada tahap berikutnya. Hasil *scraping*

twitter mempunyai banyak karakteristik yang harus dihilangkan meliputi nama URL, nama akun, *hastag*, HTML, *mention*, *emoticon*, karakter angka, tanda baca. Tabel berikut merupakan contoh data *tweet*s hasil *Cleaning* Data.

Tabel 5.9 Contoh Data *Tweet*s Hasil *Cleaning*

Tweet setelah *Case Folding*	*Tweet* setelah Cleaning
rt @aniesfor2024: kesempatan terbaik disia siakan bumn yg belum menjadi sponsor Formula E. ada lebih dr 170 negara akan siarkan scr langsung…	kesempatan terbaik disia siakan bumn yg belum menjadi sponsor Formula E ada lebih dr negara akan siarkan scr langsung
Polisi Umumkan Penyebab Atap Tribun Formula E Ambruk https://t.co/pzUU4HoLLC #TempoOtomotif	polisi umumkan penyebab atap tribun Formula E ambruk
rt @mdy_asmara1701: tiket Formula E jakarta 2022 hampir habis, tersisa kelas ancol festival	tiket Formula E jakarta hampir habis tersisa kelas ancol festival
Rt @aniesmania: membludak<u+203c><u+fe0f>warga berebut ingin melihat show car Formula E di bundaran hi. mantap banget <u+0001f44d><u+0001f44d><u+0001f44d>	membludak warga berebut ingin melihat show car Formula E di bundaran hi mantap banget

c) Normalisasi Data

Media sosial twitter mempunyai batasan karakter yang dapat diposting oleh pernggunа twitter yaitu hanya 280 karakter saja. Dengan demikian pengguna twitter mempunyai kecenderungan untuk memakai singkatan. Singkatan merupakan sebuah kata yang sering digunakan pengguna twitter dalam menuliskan *tweet* yang pangjang. Normalisasi kata juga memperbaiki kata-kata yang salah pengejaan serta dalam bahasa gaul Berikut adalah contoh data *tweet*s hasil dari normalisasi kata.

Tabel 5.10 Contoh Data *Tweet*s Hasil Normalisasi Kata

Tweet setelah cleaning	*Tweet* setelah Normalisasi
polisi umumkan penyebab atap tribun Formula E ambruk	polisi umumkan penyebab atap tribun Formula E roboh

Tweet setelah cleaning	*Tweet* setelah Normalisasi
membludakWarga berebut ingin melihat show car Formula E di bundaran hi mantap banget	banyak warga berebut ingin melihat show car Formula E di bundaran hi mantap banget
ini contoh manusia goblok yang lain dia tidak tau feo selalu bawa sponsor global di semua sirkuit Formula E dan khusus jaka	ini contoh manusia bodoh yang lain dia tidak tau feo selalu bawa sponsor global di semua sirkuit Formula E dan khusus jakarta
Formula E dan jakaa international stadium jis tlah menjadi magnet bagi elektabilitas dan popularitas anies	Formula E dan Jakarta international stadium jis telah menjadi magnet bagi elektabilitas dan popularitas anies

d) Filtering

Didalam sebuah teks terdapat kata-kata yang tidak memiliki makna dan biasanya muncul dalam frekuensi yang besar. Pada tahap ini, kata-kata tersebut dihilangkan. *Stop word removal* merupakan proses untuk menghilangkan kata yang sering muncul namun tidak memiliki makna spesifik atau perannya tidak diperlukan dalam analisis teks. Data mengenai Formula E terdapat tambahan *stop word removal* dari peneliti meliputi Formula E, Jakarta, Indonesia, eprix, dan sirkuit. Berikut contoh *tweet*s hasil *stop word removal*.

Tabel 5.11 Contoh Data *Tweet*s Hasil *Filtering*

*Tweet*s setelah Normalisasi	*Tweet* setelah Filtering
kesempatan terbaik disia siakan bumn yg belum menjadi sponsor Formula E ada lebih dr negara akan siarkan scr langsung	kesempatan terbaik disia siakan bumn sponsor negara siarkan langsung
Formula E dan Jakarta international stadium jis telah menjadi magnet bagi elektabilitas dan popularitas anies	international stadium jis magnet elektabilitas popularitas anies
ini contoh manusia bodoh yang lain dia tidak tau feo selalu bawa sponsor global di semua sirkuit Formula E dan khusus jakarta	contoh manusia bodoh tau feo bawa sponsor global sirkuit khusus
saat pagelaran Formula E tak ada bumn milik negara dan atau kementrian dari pemerintahan jika nanti sukses	pagelaran bumn milik negara kementrian pemerintahan sukses

e) Stemming

Stemming merupakan proses untuk mengubah suatu kata menjadi kata dasarnya dengan menghilangkan kata imbuhan. Dalam bahasa Indonesia dikenal dengan awalan, akiran, dan sisipan. *Stemming* pada teks bahasa Indonesia dilakukan dengan menghilangkan semua imbuhan baik berupa awalan, akhiran, ataupun kombinasi awalan dan akhiran yang ada disetiap kata. Tujuannya adalah untuk mengurangi variasi kata yang memiliki kata dasar yang sama.

Tabel 5.12 Contoh Data *Tweet*s Hasil *Stemming*

Tweet setelah Filtering	*Tweet* setelah Stemming
kesempatan terbaik disia siakan bumn sponsor negara siarkan langsung	sempat baik sia sia bumn sponsor negara siar langsung
polisi umumkan penyebab atap tribun roboh	polisi umum sebab atap tribun roboh
tiket habis tersisa kelas festival	tiket habis sisa kelas festival
menghargai budaya lokal masyarakat melarang minuman beralkohol ajang	hargai budaya lokal masyarakat larang minum alkohol ajang

5.5. Analisis Klasifikasi

Setelah melakukan pelabelan kelas sentimen, pengolahan data dilanjutkan dengan melakukan analisis klasifikasi. Dalam penelitian ini algoritma yang digunakan adalah *naïve bayes classifier* (NBC) dan *support vector machine* (SVM).

5.5.1 *Naïve Bayes Classifier* (NBC)

Sebelum melakukan klasifikasi *Naïve Bayes Clasisifier* memerlukan data *training* dan data *testing*. Data *training* digunakan untuk membentuk sebuah model klasifikasi. model ini merupakan sebuah representasi data yang digunakan dalam memprediksi kelas data yang baru. Sedangkan data *testing* digunakan untuk mengukur tingkat akurasi model yang dihasilkan dari data *training*. Rasio yang umum digunakan adalah 90 % untuk data *training* dan 10 % untuk data *testing*, namun tidak menutup kemungkinan menggunakan rasio yang lain. Hal ini disebabkan besarnya data *training* akan mempengaruhi nilai akurasi. Pada penelitian ini digunakan nilai perbandingan data *training* dan data *testing* sebesar 90 % : 10 %. Peneliti menggunakan rasio perbandingan tersebut karena nilai akurasinya baik dibandingkan dengan rasio perbandingan lainnya. Data yang dihasilkan dari perbandingan tersebut adalah:

Tabel 5.13 Perbandingan Data *Training* dan Data *Testing*

Jenis Data	Presentase	Jumlah
Data *Training*	90%	1646
Data *Testing*	10%	183
Jumlah	**100%**	**1829**

Berikut perhitungan untuk mencari jumlah data *training* dan *testing* :

Data $Training = 1829 \times 90\% = 1646$

Data $Testing = 1829 \times 10\% = 183$

Dari tabel diatas diketahui bahwa data yang digunakan sebagai data *training* sebesar 1646 *tweet*. Sedangkan data *testing* yang digunakan dalam analisis sebesar 183 *tweet*.

Setelah melakukan penentuan perbandingan data *training* dan *testing* , maka dilakukan klasifikasi *Naïve Bayes Classifier* (NBC). Klasifikasi *Naïve Bayes Classifier* dilakukan untuk mengetahui tingkat akurasi dari prediksi yang telah dibuat. Hasil prediski kelas sentimen menggunakan alogoritma *Naïve Bayes Classifier* (NBC) menggunakan Rstudio sebagai berikut:

Tabel 5.14 Hasil Klasifikasi

Prediksi	Aktual	
	Negatif	Positif
Negatif	105	15
Positif	9	54

Berdasarkan tabel 5.14 yang merupakan *confusion matrix* klasifikasi menggunakan algoritma *Naïve Bayes Classifier* (NBC). *Confusion matrix* tersebut menunjukan bahwa untuk total sentimen negatif aktualnya sebanyak 114 data. Dari total tersebut terdapat sebanyak 105 data yang diklasifikasi secara tepat sebagai sentimen negatif (*True Negative*), namun terdapat sebanyak 9 data yang diprediksi positif namun aktualnya sentimen negatif *(False Positif*). Kemudian untuk total sentimen positif aktualnya sebanyak 79 data. Dari total sentimen tersebut terdapat 54 data yang diklasifikasi benar dan tepat sebagai sentimen positif (*True Positif*). Namun terdapat 15 data yang diprediksi sebagai sentimen negatif, namun aktualnya positif (*False Negative*). Dari tabel diatas juga diketahui bahwa terdapat 24 data yang salah klasifikasi.

Selanjutnya menghitung tingkat akurasi dari prediksi yan telah dibuat. Dalam algoritma *naïve bayes* tingkat akurasi dihitung dengan *confusion matrix.* Confusion matrix digunakan untuk mengevaluasi performa model yang telah dibentuk oleh algoritma klasifikasi. Pada penelitian *confusion matrix* dihitung dengan bantuan Rstudio dan didapatkan nilai sebagai berikut:

Tabel 5.15 Nilai Akurasi

Tingkat Akurasi	Nilai
Accuracy	86,89%
Spesificity	78,26%
Precision	87,5%
Recall	92,1%

Dari tabel diatas menjelaskan bahwa nilai akurasi sebesar 86,89%, *Specificity* sebesar 78,26%, *precision* sebesar 87,5%, dan *recall* sebesar 92,1%.

5.5.2 *Support Vector Maachine* (SVM)

Support vector berkerja dengan mencari *hyperplane* atau garis pembatas (*Decision Boundary*) terbaik yang mempunyai *margin* atau jarak yang besar untuk memisahkan antara suatu kelas dengan kelas lainnya. Sebelum melakukan analisis *Support Vector Machine* memerlukan pembagian data *training* dan data *testing*. Pembagian data merupakan kebutuhan *machine learning* agar dapat memproses suatu program sesuai dengan tujuannya. Data *training* digunakan untuk membuat model klasifikasi. Sedangkan data *testing* digunakan untuk mengukur performa atau kinerja dari data *training*. Proporsi yang digunakan pada penelitian ini yaitu 90 % untuk data *training* dan 10 % untuk data *testing*. Rasio perbandingan tersebut dipakai karena mempunyai nilai akurasi yang baik dibandingkan dengan rasio perbandingan lainnya. Dari proporsi tersebut diperoleh :

Tabel 5.16 Perbandingan Data *Training* dan Data *Testing*

Jenis Data	Presentase	Jumlah
Data *Training*	90%	1646
Data *Testing*	10%	183
Jumlah	**100%**	**1829**

Berikut perhitungan untuk mencari jumlah data *training* dan *testing* :

Data $Training = 1829 \times 90\% = 1646$

Data $Testing = 1829 \times 10\% = 183$

Dari tabel diatas diketahui bahwa data yang digunakan sebagai data *training* sebesar 1646 *tweet*. Sedangkan data *testing* yang digunakan dalam analisis sebesar 183 *tweet*.

Algoritma *Support Vector Machine* (SVM) mempunyai beberapa kernel antara lain *linear, polynomial, radial basis function* (RBF), dan *sigmoid.* Berikut hasil perbandingan dari kenel linear, *radial basic function* (RBF), dan sigmoid.

a. Kernel Linear

Kernel linear merupakan fungsi kernel yang baik digunakan ketika data sudah terpisah secara linear. Dalam melakukan analisis menggunakan fungsi kernel linear, dilakukan optimasi parameter C atau *Cost.* Nilai C bekerja sebagai optimalisasi svm untuk mengurangi kesalahan klasifikasi disetiap sampel data *training*. Semakin besar nilai C maka algoritma akan mengurangi kesalahan klasifikasi sebanyak mungkin dan menyeba kan *hyperplane* menjadi kecil. Setelah dilakukan analisis diperoleh parameter model sebagai berikut:

Tabel 5.17 Parameter SVM Kernel Linear

Parameter		
SVM-Type	:	C-Classification
SVM-Kernel	:	Linear
Cost	:	100
Number Of *Support vector*	:	762

Berdasarkan tabel 5.17, SVM-Type: C-Classification merupakan jenis dari metode SVM yaitu C-Classification. SVM-Kernel: Linear menunjukan bahwa kernel yang digunakan adalah kenel linear. *Cost*:100 menunjukan nilai parameter terbaik yang digunakan dalam analisis sebesar 100 dan banyaknya *support vector* atau pemisah *hyperplane* adalah 762 data.

Tabel 5.18 *Confusion Matrix* Kernel Linear

Prediksi	Aktual	
	Negatif	Positif
Negatif	111 (TN)	3 (FN)
Positif	5 (FP)	64 (TP)

Berdasarkan tabel 5.18 yang merupakan *confusion matrix* klasifikasi menggunakan algoritma *Support Vector Machine* (SVM) kernel linear. *Confusion matrix* tersebut menunjukan bahwa untuk total sentimen negatif aktualnya sebanyak 116 data. Dari total tersebut terdapat sebanyak 111 data yang diklasifikasi secara tepat sebagai sentimen negatif (*True Negative*), namun terdapat sebanyak 5 data yang diprediksi positif namun aktualnya sentimen negatif *(False Positif*). Kemudian untuk total sentimen positif aktualnya sebanyak 67 data. Dari total sentimen tersebut terdapat 64 data yang diklasifikasi benar dan tepat sebagai sentimen positif (*True Positif*). Namun terdapat 3 data yang diprediksi sebagai sentimen negatif, namun aktualnya positif (*False Negative*). Dari tabel diatas juga diketahui bahwa terdapat 8 data yang salah klasifikasi.

Dalam menghitung kinerja klasifikasi dapat menggunakan pendekatan *confusion matrix* melalui nilai dari akurasi, recall, precision, dan lain sebagainya. Berikut hasil dari nilai akurasi, precision, specificity, dan racall:

Tabel 5.19 Nilai Kinerja Klasifikasi

Tingkat Akurasi	Nilai
Accuracy	95,63%
Spesificity	95,52%
Precision	97,36%
Recall	95,68%

Pada tabel 5.19 diketahui nilai akurasi sebesar 95,63%, *specificity* sebesar 95,52%, nilai *precision* sebesar 97,36%, dan nilai *recall* sebesar 95,68%.

b. Kernel *Radial Basic Function* (RBF)

Fungsi kernel kedua yang digunakan dalam analisis adalah *Radial Basic Function* (RBF). Berbeda dengan kernel linear yang digunakan pada data yang sudah terpisah secara linear, namun pada kernel RBF digunakan ketika data tidak terpisah secara linear. Dalam melakukan analisis menggunakan RBF, dilakukan optimalisasi pada parameter C (*Cost*) dan γ (*Gamma*). Nilai *gamma* memiliki fungsi untuk menentukan seberapa jauh pengaruh dari satu sampel *training* dataset pada garis pemisahnya. Setelah melakukan analisis kernel RBF diperoleh nilai parameter sebagai berikut:

Tabel 5.20 Parameter SVM Kernel RBF

Parameter		
SVM-Type	:	C-Classification
SVM-Kernel	:	Radial
Cost	:	100
Number Of *Support vector*	:	1128

Berdasarkan tabel 5.20 menunjukan SVM-Type: C-Classification merupakan jenis dari metode SVM yaitu C-Classification. SVM-Kernel: Radial menunjukan bahwa kernel yang digunakan adalah kernel *radial basic function* (RBF). *Cost*:100 menunjukan nilai parameter terbaik yang digunakan dalam analisis sebesar 100 dan banyaknya *support vector* atau pemisah *hyperplane* adalah 1128 data.

Tabel 5.21 *Confusion Matrix* Kernel RBF

Prediksi	Aktual	
	Negatif	Positif
Negatif	111 (TN)	3 (FN)
Positif	15 (FP)	54 (TP)

Berdasarkan tabel 5.21 yang merupakan *confusion matrix* kernel RBF. *Confusion matrix* tersebut menunjukan bahwa untuk total sentimen negatif aktualnya sebanyak 126 data. Dari total tersebut terdapat sebanyak 111 data yang diklasifikasi secara tepat sebagai sentimen negatif (*True Negative*), namun terdapat sebanyak 15 data yang diprediksi positif namun aktualnya sentimen negatif *(False Positif*). Kemudian untuk total sentimen positif aktualnya sebanyak 57 data. Dari total sentimen tersebut terdapat 54 data yang diklasifikasi benar dan tepat sebagai sentimen positif (*True Positif*). Namun terdapat 3 data yang diprediksi sebagai sentimen negatif, namun aktualnya positif (*False Negative*). Dari tabel diatas juga diketahui bahwa terdapat 18 data yang salah klasifikasi.

Dalam menghitung kinerja klasifikasi dapat menggunakan pendekatan *confusion matrix* melalui nilai dari akurasi, recall, precision, dan lain sebagainya. Berikut hasil dari nilai akurasi, precision, specificity, dan racall:

Tabel 5.22 Nilai Kinerja Klasifikasi

Tingkat Akurasi	**Nilai**
Accuracy	90,16%

Tingkat Akurasi	Nilai
Spesificity	94,74%
Precision	97,36%
Recall	88,09%

Pada tabel 5.22 diketahui nilai akurasi sebesar 90,16%, *specificity* sebesar 94,74%, nilai *precision* sebesar 97,36%, dan nilai *recall* sebesar 88,09%.

c. Kernel Sigmoid

Fungsi ketiga yang digunakan dalam analisis yaitu kernel sigmoid. Kernel sigmoid merupakan fungsi kernel yang digunakan ketika data tidak terpisah secara linear.

Tabel 5.23 Parameter SVM Kernel Linear

Parameter		
SVM-Type	:	C-Classification
SVM-Kernel	:	Sigmoid
Cost	:	100
Coef.0	:	0
Number Of *Support vector*	:	1239

Berdasarkan tabel 5.23 menunjukan SVM-Type: C-Classification merupakan jenis dari metode SVM yaitu C-Classification. SVM-Kernel: Sigmoid menunjukan bahwa kernel yang digunakan adalah kernel Sigmoid. *Cost*:100 menunjukan nilai parameter terbaik yang digunakan dalam analisis sebesar 100 dan banyaknya *support vector* atau pemisah *hyperplane* adalah 1239 data.

Tabel 5.24 *Confusion Matrix* Kernel Sigmoid

Prediksi	Aktual	
	Negatif	Positif
Negatif	108 (TN)	6 (FN)
Positif	12 (FP)	57 (TP)

Berdasarkan tabel 5.24 yang merupakan *confusion matrix* kernel sigmoid. *Confusion matrix* tersebut menunjukan bahwa untuk total sentimen negatif aktualnya sebanyak 120 data. Dari total tersebut terdapat sebanyak 108 data yang diklasifikasi secara tepat sebagai sentimen negatif (*True Negative*), namun terdapat sebanyak 12 data yang diprediksi positif namun aktualnya sentimen negatif *(False*

Positif). Kemudian untuk total sentimen positif aktualnya sebanyak 63 data. Dari total sentimen tersebut terdapat 57 data yang diklasifikasi benar dan tepat sebagai sentimen positif (*True Positif*). Namun terdapat 6 data yang diprediksi sebagai sentimen negatif, namun aktualnya positif (*False Negative*). Dari tabel diatas juga diketahui bahwa terdapat 18 data yang salah klasifikasi.

Dalam menghitung kinerja klasifikasi dapat menggunakan pendekatan *confusion matrix* melalui nilai dari akurasi, recall, precision, dan lain sebagainya. Berikut hasil dari nilai akurasi, precision, specificity, dan racall.

Tabel 5.25 Nilai Kinerja Klasifikasi

Tingkat Akurasi	Nilai
Accuracy	90,16%
Spesificity	90,48%
Precision	94,73%
Recall	90%

Pada tabel 5.25 diketahui nilai akurasi sebesar 90,16%, specificity sebesar 90,48%, nilai precison sebesar 94,73%, dan nilai recall sebesar 90%.

5.6. Perbandingan Metode NBC dan SVM

Hasil algoritma *naïve bayes classifier* (NBC) dan *support vector machine* (SVM) pada penelitian ini akan digunakan untuk menentukan algoritma terbaik dalam klasifikasi. Untuk menentukan algoritma terbaik dapat diketahui melalui nilai akurasi tertinggi. Berikut hasil perbandinganya:

Tabel 5.26 Perbandingan NBC dan SVM

	NBC	SVM Linear	SVM RBF	SVM Sigmoid
Akurasi	86,89%	95,63%	90,16%	90,16%

Berdasarkan tabel 5.26 menunjukan nilai akurasi tertinggi diperoleh oleh algoritma *support vector machine* (SVM) dengan kernel liner sebesar 95,63%. Oleh karena itu, pada kasus ini kernel linear merupakan algortima terbaik untuk melakukan klasifikasi.

BAB VI
PENUTUP

6.1. Kesimpulan

Berdasarkan analisis dan pembahasan yang telah dilakukan, dapat disimpukan sebagai berikut:

1. Penelitian ini melakukan *scraping* media sosial twitter tentang Formula E di Jakarta. Dari hasil *scraping* data yang digunakan analisis sebesar 1829 *tweet*. Setelah melakukan analisis diperoleh informasi bahwa 1107 *tweet* diklasifikasikan sebagai sentimen negatif , dan 722 *tweet* diklasifikasikan sebagai sentimen positif.
2. Pada sentimen positif terdapat sebanyak 722 *tweet* dengan kata atau topik yang paling banyak dibahas yaitu sukses, balap, sponsor, tiket, dukung dan sebagainya. Hal tersebut dapat diartikan bahwa Formula E dapat dilaksanakan dengan sukses dan mendapat dukungan dari masyarakat serta pemerintah sehingga Indonesia dapat lebih dikenal didunia internasional. Sedangkan sentimen negatif terdapat sebanyak 1107 *tweet* dengan kata atau topik yang banyak dibahas yaitu roboh, bodoh, tribun, bir dan sebagainya. Kata atau topik pada sentimen tersebut dapat diartikan bahwa Formula E yang diadakan di Jakarta ini mengalami insiden tribun sirkuit roboh akibat badai yang membuat penonton khawatir. Kemudian kata "bir" diartikan bahwa Indonesia yang penduduknya mayoritas muslim diharapkan sponsor dengan merk bir tersebut dapat menghargai budaya serta orang-orang muslim di Indonesia.
3. Hasil perbandingan algoritma *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM) dengan menggunakan rasio data *traning* dan *testing* sebesar 90%:10% diperoleh nilai akurasi algoritma *Naïve Bayes Classifier* (NBC) sebesar 86,89%, nilai akurasi *Support Vector Machine* (SVM) kernel linear sebesar 95,63%, nilai akurasi SVM *Radial Basic Function* (RBF) sebesar 90,16%, dan nilai akurasi SVM sigmoid sebesar 90,16%. Berdasarkan nilai akurasi tersebut disimpulkan algoritma SVM linear

mempunyai kinerja lebih baik dalam melakukan klasifikasi tentang Formula E dengan nilai akurasi sebesar 95,63%.

6.2. Saran

Berdasarkan beberapa kesimpulan diatas, adapun saran yang dapat diberikan adalah sebegai berikut:

1. Penelitian selanjutnya diharapkan dapat mempelajari data lebih *detail* sehingga diperoleh informasi yang lebih dalam serta dapat menggunakan berbagai bahasa, karena pada penelitian ini terbatas pada ulasan dalam Bahasa Indonesia.
2. Pada penelitian ini pelabelan kelas sentimen terbatas dengan menggunakan kamus *lexicon*, sehingga kata negasi tidak dapat diidentifikasi dengan baik. Oleh sebab itu, pada penelitian selanjutnya dapat melakukan penanganan terhadap kata negasi agar hasil pelabelan lebih akurat.
3. Pada penelitian ini pembagian data *training* dan *testing* menggunakan perbandingan 90%:10%. Namun, pembagian data juga dapat dilakukan dengan K-Fold *Cross Validation*. Dalam metode ini seluruh data digunakan sebagai data *training*.
4. Pada algoritma *Support Vector Machine* (SVM) mempunyai banyak parameter seperti *cost* (C) dan *gamma* (γ). pada penelitian selanjutnya diharapkan peneliti dapat melakukan optimalisasi parameter (*Tunning*) dengan cara melakukan *grid search* dengan melatih atau men*training* banyak model yang berbeda dari setiap nilai parameternya.

DAFTAR PUSTAKA

Hidayatulloh , A. F., Yusuf, A. A., Juwairi, K. P., & Nayoan , R. A. (2019). Identifikasi Konten Kasar pada *Tweet* Bahasa Indonesia. *J. Linguist Komputasional, 2*, 1-5.

Advertorial. (2017, April 28). *Ternyata Ini Dia Sepeda Motor Pertama di Indonesia!* Diambil kembali dari tribunnews.com: https://www.tribunnews.com/otomotif/2017/04/28/ternyata-ini-dia-sepeda-motor-pertama-di-indonesia

Ahmad, A. (2017). Mengenal Artificial Intelligence, Machine Learning, Neural Network, dan Deep Learning. *Yayasan Cahaya Islam, Jurnal Teknologi Indonesia*, 1-5.

Aliady, H., Tuasikal, N. J., & Widodo, E. (2018). IMPLEMENTASI SUPPORT VECTOR MACHINE (SVM) DAN RANDOM FOREST PADA DIAGNOSIS KANKER PAYUDARA. *Seminar Nasional Teknologi Informasi dan Komunikasi 2018 (SENTIKA 2018)* (hal. 278-285). Yogyakarta: SENTIKA 2018.

Alpaydin, E. (2010). *Introduction to Machine Learning Second Edition.* london: MIT Press.

Alyusi, S. D. (2016). *Media Sosial Interaksi, Identitas, dan Modal Sosial.* Jakarta: Kencana.

Andini, P. D. (2014). Penerapan Analisis Pohon Regresi pada Data Perlindungan Sosial.

Argadea, D. (2019, September 9). *Kawasaki Adalah Pabrikan Motor Mahal di Indonesia, Benarkah Demikian?* Diambil kembali dari Gridoto.com: https://www.gridoto.com/read/221846459/kawasaki-adalah-pabrikan-motor-mahal-di-indonesia-benarkah-demikian

Asnawi , M. H., Firmansyah, I., Novian, R., & Pontoh , R. S. (2021). Perbandingan Algoritma Naive Bayes, K-NN dan SVM dalam Pengklasifikasian Sentimen Media Sosial. *PROSIDING SEMINAR NASIONAL STATISTIKA X, Vol. 10*, 2599-2546. doi:https://doi.org/10.1234/pns.v10i.85

Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM, dan Naive Bayes. *Jurnal Media Informatika Budidarma*, 640-651.

Bain, L. J., & Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistic Second Edition.* California: Duxbury Press.

BPS. (2020). *Perkembangan Jumlah Kendaraan Bermotor Menurut Jenis, 1949-2018.* Diambil kembali dari bps.go.id: https://www.bps.go.id/dynamictable/2016/02/09/1133/perkembangan-jumlah-kendaraan-bermotor-menurut-jenis-1949-2018.html

Brahimi, B., Touahria, M., & Tari, A. (2019). Improving sentimentanalysis in Arabic: A combined approach. . *Journal of King Saud University -Computer and Information Sciences.*, online Available at: <https://doi.org/10.1016/j.jksuci.2019.07.011>.

Breiman, L. (2001). *Random Forests.* University of California: Statistics Department.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1993). *Classification and Regression Trees.* New York: Chapman and Hall.

Budianto, A., Maryono, D., & Ariyuana, R. (2018). Perbandingan K-Nearest Neighbor (K-NN) dan Support Vector Machihne (SVM) dalam karakteristik Plat Kendaraan Bermotor. *Jurnal Ilmiah Pendidikan Teknik Kejuruan* , 11.

Daeng, D. A. (2017, Desember 27). *Apa Modal Honda Sampai Jadi Raja Sepeda Motor?* Diambil kembali dari tirto.id: https://tirto.id/apa-modal-honda-sampai-jadi-raja-sepeda-motor-cClb

Daqiqil, I. (2021). *MACHINE LEARNING: Teori, Studi Kasus dan Implementasi Menggunakan PYTHON.* Pekanbaru: UR PRESS.

Davies, a. P. (2004). *Database System Third Edition.* New York: Palgrave macmillan.

Dewi, T. T. (2006). Penerapan Metode Regresi Berstruktur Pohon pada Pendugaan Lama Penyusunan Skripsi Mahasiswa (Skripsi).

Drajana, I. C. (2018). Prediksi Jumlah Produksi Coconut Oil Menggunakan k-Nearest Neighbor dan Backward Elimination. *TECNOSCIENZA*, 51-64.

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* New York: Cambridge University Press.

Fransiska, S., Rianto, R., & Gufroni, A. I. (2020). Sentiment Analysis Provider By. U on Google Play Store Reviews with TF-IDF and Support Vector. *Scientific Journal of Informatics, Vol.2 No.2*, 203-212.

Gustiana, Z., Priyanto, C., & dkk. (2021). *Data Mining dan Penerapan Algoritma.* Yayasan Kita Menulis: Yayasan Kita Menulis.

Hakim, A. (2001). *Statistika Deskriptif untuk Ekonomi dan Bisnis.* Yogyakarta: Ekonisia.

Halim, C., & Prasetyo, H. (2018). Penerapan Artificial Intelligence dalam Computer Aided Instructure (CAI). *Jurnal Sistem Cerdas 2018 Volume 01*, 45-51.

han, J., & Kamber, M. (2011). *Data Mining Concepts and Techniques Third* . Waltham:: Elsevier Inc.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques Third Edition.* Waltham USA: Morgan Kaufmann.

Harrington, P. (2012). *Machine Learning in Action.* New York: Manning Publication.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning: Data mining, Inference, and Prediction.* New York: Springer.

Hogg, R., & Craig, A. (2005). *Introduction to Mathematical Statistics (6th ed.).* New Jersey: Pearson Prentice Hall.

https://www.fiaFormulae.com/. (2022, November Minggu). *https://www.fiaFormulae.com/*. Diambil kembali dari https://www.fiaFormulae.com/: https://www.fiaFormulae.com/en

Husada, H. C., & Paramita, A. S. (Maret 2021). Analisis Sentimen Pada Maskapai Penerbangan di Platform Twitter Menggunakan Algoritma Support Vector Machine (SVM). *TEKNIKA, 10(1)*, 2549-8045. doi:DOI: 10.34148/teknika.v10i1.311

Indriati, Idris, H. K., & Fauzi, M. A. (2019). Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan SAMBAT Online Menggunakan Metode Naïve

Bayes dan Kombinasi Seleksi Fitur. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, Vol. 3, No. 3*, 2436-2442.

Isal. (2018, Januari 8). *Ini Penyebab Yamaha XMAX Punya Harga Bekas Lebih Mahal dari Barunya*. Diambil kembali dari gridoto.com: https://www.gridoto.com/read/221010396/ini-penyebab-yamaha-xmax-punya-harga-bekas-lebih-mahal-dari-barunya

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R.* New York: Springer.

Jonathan, S. (2006). *Metode Penelitian Kuantitatif dan Kualitatif.* Yogyakarta: Graha Ilmu.

juju, D., & studio, M. (2009). *TWITTER.* JAKARTA: PT ELEK MEDIA KOPUTINDO.

Kaswidjanti, W., Aribowo, A. S., & Wicaksono, C. B. (2014). Implementasi Fuzzy Inference System Metode Tsukamoto Pada Pengambilan Keputusan Pemberian Kredit Pemilikan rumah. *Telematika*, 137-146.

Koriaty, S., & Agustani, M. D. (2016). Pengembangan Model Pembelajaran Game Edukasi untuk Meningkatkan Minat Siswa Kelas X TKJ SMK Negeri 7 Pontianak. *Jurnal Edukasi Vol. 14 No. 2*, 227.

Kurnia, A. B. (2014). Penerapan Realistic Mathematics Education Dalam Pembelajaran Membaca Diagram Batang dan Garis Siswa SMP Kelas VII. *AdMathEdu Vol 4 No.2*.

Kusnawi. (2007). Pengantar Solusi Data Mining. *Seminar Nasional Teknologi 2007 (SNT).* Yogyakarta: STMIK AMIKOM Yogyakarta.

Kusuma. (2009). *Creative Project:Pintar Twitter.* Jakarta: Grasindo.

Kusuma, P. D. (2020). *Machine Learning:Teori, Program, dan Studi Kasus.* Yogyakarta: Penerbit DEEPUBLISH (Grup Penerbitan CV. BUDI UTAMA).

Kusumadewi, S., & Purnomo, H. (2004). *Aplikasi logika fuzzy untuk mendukung keputusan.* Yogyakarta: Graha ilmu.

Lewis, R. J. (2000). An Introduction to Classification and Regression Tree (CART) Analysis. *In Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California*.

Li, X. (2013). Comparison and Analysis between Holt Exponential Smoothing and Brown Exponential Smoothing Used for Freight Turnover Forecast. *Third International Conference on Intelligent System Design and Engineering Applications* (hal. 453-456). IEEE.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News Vol. 2/3*, 18-22.

Mayangningsih, Siswanto, & Mesterjon. (2013). Metode Logika Fuzzy Tsukamoto Dalam Sistem Pengambilan Keputusan Penerimaan Beasiswa. *Jurnal Media Infotama*, 140-165.

Muhammad, H. (2019, September 21). *Perkuat Bisnis, iCar Asia Akuisisi Carmudi Indonesia*. Diambil kembali dari Republika.co.id: https://republika.co.id/berita/py6p4x380/perkuat-bisnis-icar-asia-akuisisi-carmudi-indonesia

Muslim, M. (2020, Januari 20). *Data Penjualan Motor 2019 dari AISI Untuk Semua Merek*. Diambil kembali dari BMSPEED7.com: https://bmspeed7.com/data-penjualan-motor-2019/

Muttaqin, M. N., & Kharisudin, I. (2021). ANALISIS SENTIMEN APLIKASI GOJEK MENGGUNAKAN SUPPORT VECTOR MACHINE DAN K NEAREST NEIGHBOR. *UNNES Journal of Mathematics*, 10 (2).

Nanja, M., & Purwanto. (2015). Metode K-Nearest Neighbor Berbasis Forward Selection untuk Prediksi Harga Komoditi Lada. *Jurnal Pseudocode*, 53-64.

Nugraha, F. A., Harani, N. H., & Habibi, R. (2020). *Analisis sentimen Terhadap Pembatasan Sosial Menggunakan Deep Learning.* Bandung: Kreatif Industri Nusantara.

Nugraha, F. A., Harani, N. H., & Habibi, R. (2021). *Analisis Sentimen Terhadap Pembatasan Sosial Menggunakan DEEP LEARNING.* Bandung: Kreatif Industri Nusantara.

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How Many Trees in a Random Forest? *Lecture Notes in Computer Science*, 154-168.

Pahlevi, N. A. (2021). *Pengaruh Media Sosial dan Gerakan Massa Terhadap Hakim.* Surabaya: Cipta Media Nusantara (CMN).

Pal, N., Sundararaman, D., Arora, P., Kohli, P., & Palakurthy, S. S. (2018). How much is my car worth? A methodology for predicting used cars prices using Random Forest. *Future of Information and Communications Conference*.

Pertiwi, S. R. (2018). *PERBANDINGAN METODE K-NEAREST NEIGHBOR DAN SUPPORT VECTOR MACHINE DALAM ANALISIS SENTIMEN TWITTER TERHADAP STASIUN TELEVISI BERITA INDONESIA.* Yogyakarta: Universitas Gajah Mada.

Polamuri, S. R., Srinivas, K., & Mohan, A. K. (2019). Stock Market Prices Prediction using Random Forest and Extra Tree Regression. *International Journal of Recent Technology and Engineering*, 1224-1228.

PP. (2012). *Peraturan Pemerintah (PP) tentang Kendaraan.*

Praghakusma, A. Z., & Charibaldi, N. (Juni 2021). Komparasi Fungsi Kernel Metode Support Vector Machine untuk Analisis Sentimen Instagram dan Twitter (Studi Kasus : Komisi Pemberantasan Korupsi). *Jurnal Sarjana Teknik Informatika, Vol.9, No.2*, 33-42.

Prasetyowati, E. (2017). *DATA MINING (Pengelompokan Data untuk Informasi dan Evaluasi.* Pamekasan: Duta Media Publising.

Raharjo, B. (2017). *Belajar Otodidak Flask (Framework Python Untuk Pengembangan Aplikasi Web).* Bandung: INFORMATIKA.

Rahutomo, R., Perbangsa, A. S., Lie, Y., Cenggoro, T. W., & Pardamen, B. (2019). Artificial Intelligence Model Implementation in Web-Based Application for Pineapple Object Counting. *2019 International Conference on Information Management and Technology (ICIMTech)*, 525-530.

Rale, N., Solanki, R., Bein, D., Andro-Vasko, J., & Bein, W. (2019). Prediction of Crop Cultivation. *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)* (hal. 0227-0232). Las Vegas, NV, USA, USA: IEEE.

Ramadhan, M. M., Sitanggang, I. S., Nasution, F. R., & Ghifari, A. (2017). Parameter Tuning in Random Forest Based on Grid Search Method for Gender Classification Based on Voice Frequency. *International Conference on Computer, Electronics and Communication Engineering 2017*, (hal. 625).

Rasyad, R. (2013). *Metode Statistik Deskriptif untuk Umum.* Jakarta: Grasindo.

Rizqiyani, V., Mulwinda, A., & Putri, R. D. (2017). Klasifikasi Judul Buku dengan Algoritma Naive Bayes dan pencarian Buku pada Perpustakaan Jurusan Teknik Elektro. *Jurnal Teknik Elektro, Vol.9 No.2*, 60-65.

Roifa, A. N. (2018). *TEXT MINING DENGAN METODE NAÏVE BAYES CLASSIFIER UNTUK MENGKLASIFIKAN BERITA BERDASARKAN KONTEN.* Surabaya: Institut Teknologi Sepuluh November.

Rosen, A. (2017, September Senin). *Mencuit akan Jadi Lebih Mudah*. Diambil kembali dari Twitter: https://blog.twitter.com/in_id/topics/product/2017/Mencuit-akan-Jadi-Lebih-Mudah

Rosi, F., Fauzi, M. A., & Perdana, R. S. (2017). Prediksi Rating Pada Review Produk Kecantikan Menggunakan Metode Naïve Bayes dan Categorical Proportional Difference (CPD). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, Vol.2 No.5*, 1991–1997.

Santosa, B. (2007). *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis.* Yogyakarta: Penerbit Graha Ilmu.

Saputra, R. (2013). *Statistika terapan dalam ilmu kesehatan masyarakat.* Sumatera Barat: STIKES Perintis.

Sari, H., Ginting, G. L., Zebua, T., & Mesran. (2021). Penerapan Algoritma Text Mining dan TF-IDF untuk Pengelompokan Topik Skripsi Pada Aplikasi Repository STMIK Budi Darma. *Terapan Informatika Nusantara*, 414-432.

Sekaran, U. (2011). *Research Methods For Business (Metodologi Penelitian untuk Bisnis).* Jakarta: Salemba Empat.

Sholihin, M., Fuad, N., & Khamiliyah, N. (2013). Sistem Pendukung Keputusan Penentuan Warga Penerima Jamkesmas Dengan Metode Fuzzy Tsukamoto. *Jurnal Teknika*, 501-505.

Siburian, V. W., & Mulyana, I. E. (2018). Prediksi Harga Ponsel Menggunakan Metode Random Forest. *Annual Research Seminar (ARS) 2018* (hal. 144-147). Palembang: Fakultas Ilmu Komputer UNSRI.

Simarmata, J. (2006). *Pengenalan Teknologi Komputer Informasi.* Yogyakarta: Andi.

Singh, M., Verma, A., Parasher, A., Chauhan, N., & Budhiraja, G. (2019). Implementation of Database Using Python Flask Framework. *International Journal of Engineering and Computer Science*, 24894-24899.

Sinta, D. (2015). *METODE ENSEMBLE K-NEAREST NEIGHBOR UNTUK PREDIKSI HARGA BERAS DI INDONESIA.* BOGOR: SEKOLAH PASCASARJANA INSTITUT PERTANIAN BOGOR .

Sirait, S. E. (2015, Februari 24). *Penjualan Motor Bekas Diyakini Terus Tumbuh hingga 10 Tahun kedepan*. Diambil kembali dari okeotomotif: https://otomotif.okezone.com/read/2015/02/24/15/1109895/penjualan-motor-bekas-diyakini-terus-tumbuh-hingga-10-tahun

Sugiyono. (2017). *Metode Penelitian Kuantitatif, Kualitatif, dan R&D.* Bandung: Alfabeta.

Suyatno, T., Chalik, H., Sukada, M., Ananda, C. Y., & Marala, D. T. (1999). *Dasar-dasar perkreditan.* Jakarta: Gramedia pustaka utama.

Tan, A. &. (2020). Text Mining: the state of the art and the chalengges. IN PROCEEDINGS OF THE PAKDD 1999 ON KNOWLEDGE DISOCOVERY FROM ADVANCED DATABASES. Dalam E. D. Wahyuni, A. A. Arifiyanti, & M. I. Afandi, *KLASIFIKASI TEKS dengan PYTHON* (hal. 5). Sidoarjo: Indomedia Pustaka.

Taylor, O., Ezekiel, P., & Deedam-Okuchaba, F. (2019). A Model to Detect Heart Disease using Machine Learning Algorithm. *International Journal of Computer Sciences and Engineering, Volume-7, Issue-11*, 1-5.

Turban, E. (2005). *Decision Support Systems and Intelligent System.* Yogyakarta: Penerbit Andi.

Wahyono, T. (2018). *Fundamental of Python for Machine Learning (Dasar-dasar Pemrograman Python untuk Machine Learning dan Kecerdasan Buatan).* Yogyakarta: Penerbit Gava Media.

Wahyuni, E. D., Arifiyati, A. A., & Afandi, M. I. (2020). *Klasifikasi Teks dengan PYTHON.* Sidoarjo: Indomedia Pustaka.

Walpole, E., & Myers, H. (1995). *Ilmu Peluang dan Statistika untuk Insinyur dan Ilmuwan.* Bandung: ITB.

Wijaya, M., Junaedy, & Arfandy, H. (2019). Perancangan Chatbot Untuk Informasi Penerimaan Mahasiswa Baru pada STMIK Kharisma Makassar. *Jurnal Ilmu Komputer Kharisma.tech Vol 14 No 1*, 14-23.

Woloeyo, Y. J. (2010). *Twitter Best Social Networking.* Yogyakarta: ANDI YOGYAKARTA & ELCOM.

worldometers. (2022, 6 rabu). *worldometers*. Diambil kembali dari worldometers: https://www.worldometers.info/world-population/

Wulansari, M. J. (2018). *Analisis Faktor-Faktor Yang Mempengaruhi Seseorang Terkena Penyakit Diabetes Melitus Menggunakan Regresi Random Forest.* Yogyakarta: Universitas Islam Indonesia.

Zhang, Y., Wu, X., Gu, C., & Xie, Y. (2019). Predict Future Sales using Ensembled Random Forests. *ArXiv*.

Zhu, Y. (2018). Comparison of Model Performance for Basic and Advanced Modeling Approaches to Crime Prediction. *Intelligent Information Management*, 123-132.

LAMPIRAN

Lampiran 1 Data *Tweet*

text	created	*retweet*Count
RT @aniesfor2024: Kesempatan terbaik disia siakan BUMN yg belum menjadi SPONSOR FORMUA E. Ada lebih dr 170 negara akan siarkan scr langsung…	5/30/2022 23:59	402
RT @tatakujiyati: Finally he got it right. Anggota PDIP, Gembong Warsono, mengatakan perhelatan Formula E harus tetap berkelanjutan. Menuru…	5/30/2022 23:59	116
RT @KOmengKO_25: Balapan GP Mandalika gw ga nonton. Balapan Formula E Ancol gw ga nonton. Balapan ESEMKA berapa pun tiketnya & dimana pun…	5/30/2022 23:59	98
RT @ZettaZahra2: Atap Tribun E Formula tinggi & ada di lapangan (ruang terbuka). Tenda2 warung makanan kan rendah & di sekitarnya ada ban…	5/30/2022 23:59	3
Polisi Umumkan Penyebab Atap Tribun Formula E Ambruk https://t.co/pzUU4HoLLC #TempoOtomotif	5/30/2022 23:59	1

RT @Boediantar4: Formula E mengharumkan nama Indonesia tanpa BUMN ... https://t.co/rHqPcxw5dW	5/30/2022 23:59	325
RT @aniesmania: Membludak<U+203C><U+FE0F> Warga Berebut Ingin Melihat Show Car Formula E di Bundaran HI. Mantap banget <U+0001F44D><U+0001F44D><U+0001F44D> #Anies #AniesBaswedan https://…	5/30/2022 23:59	191
RT @forbescolombia: Afrocolombiana, doctora en Educación, ingeniera, docente e investigadora, Marelen Castillo pasó a segunda vuelta como f…	5/30/2022 23:59	576
@wong_dalan Lah loe nanya? Loe tanya teman loe dan junjungan loe, kok bisa dukung Formula E. Kemaren habis habisan loe nyinyir	5/30/2022 23:59	0
@CNNIndonesia Yakin gara2 itu, bukannya diserang netizen gara2 diem soal Formula E? Ngaku bang @sandiuno	5/30/2022 23:59	0
@gefersonk89 O problema é que há muita gente que só tem olhos para a Fórmula 1. De fato é a nata do automobilismo e… https://t.co/kC06KlFLn2	5/30/2022 23:59	0

Lampiran 2 *Scraping* Data

```
#BISMILLAH
#instal dan panggil library yang dibutuhkan untuk analisis
sentimen
library(twitteR)
library(rtweet)
library(stringr)
library(ggplot2)
library(NLP)
library(SnowballC)
library(plyr)
library(RColorBrewer)
library(wordcloud)
library(wordcloud2)
library(stopwords)
library(tm)
library(textclean)
library(RCurl)
library(dplyr)
library(tokenizers)
library(devtools)
install_github("nurandi/katadasaR")
library(katadasaR)
library(corpus)
library(tmap)

# Kode API Twitter
consumer_key <- "FfDAvSTYxUCpi3ZDg3LmDF5Uf"
consumer_secret                                          <-
"YkRGsxZtTcPEgnZfwK5hXttXlwh4FAFmRZdjy9KM1lNA105Wid"
```

```
access_token <- "1332748896213241856-
yno15meVJepPd6WMBQtWe1h213RlnY"
access_secret <-
"WtvnhJVGGcQdXpjEXWkGvd9ryg8ueDyyIQQ27eA5btsb4"
setup_twitter_oauth(consumer_key, consumer_secret,
access_token, access_secret)

#Crawlind Data Twitter
tweets = searchTwitter('LGBT+islam',n = 15000,
                       retryOnRateLimit = 10000,
since="2022-05-12",until="2022-05-25")
n.tweet <- length(tweets)
tweets.df <- twListToDF(tweets)
dim(tweets.df)
tweets[1:10]

#Menyimpan Data Hasil Crawling
dataframe=data.frame(tweets.df)
write.csv(dataframe,file = 'C:\\BISMILLAH\\lgbtislam.csv')
```

Lampiran 3 *Text Prepocessing*

```
uniq.text                                                    <-
read.csv('C:\\BISMILLAH\\PROSES\\FORMULADUP1.csv')
#Pre-Processing data
# Membangun corpus data
twetnabil=data.frame(uniq.text)
View(twetnabil)
tweet.corpus<-VCorpus(VectorSource(twetnabil$x))
# 1. CASEFOLDING
#Transform to lower case
tweet.corpus<-tm_map(tweet.corpus,
content_transformer(tolower))
casefolding<-
data.frame(text=sapply(tweet.corpus,as.character),
stringsAsFactors = FALSE)
View(casefolding)
write.csv(casefolding,                file                =
"C:\\BISMILLAH\\PROSES\\casefoldingbaru.csv")
# 2 CLEANSING DATA
#Remove URL
remove.URL<-function(x){gsub("http[^[:space:]]*", "", x)}
tweet.corpus<-
tm_map(tweet.corpus,content_transformer(remove.URL))
removeurl<-
data.frame(text=sapply(tweet.corpus,as.character),
                     stringsAsFactors = FALSE)
View(removeurl)
write.csv(removeurl,                  file                =
"C:\\BISMILLAH\\PROSES\\remove_urlbaru.csv")
#Unescape HTML
```

```
unescapeHTML<-
function(str){xml2::xml_text(xml2::read_html(paste0("<x>",
str, "</x>")))}
tweet.corpus<-
tm_map(tweet.corpus,content_transformer(unescapeHTML))
removehtml<-
data.frame(text=sapply(tweet.corpus,as.character),stringsAsF
actors = FALSE)
View(removehtml)
write.csv(removehtml,                                    file
="C:\\BISMILLAH\\PROSES\\remove_htmlbaru.csv")
#Remove Mention
removemention<-function(x){gsub("@\\w+", "",x)}
tweet.corpus<-
tm_map(tweet.corpus,content_transformer(removemention))
removemention<-
data.frame(text=sapply(tweet.corpus,as.character),
stringsAsFactors = FALSE)
View(removemention)
write.csv(removemention,file="C:\\BISMILLAH\\PROSES\\removem
entionbaru.csv")

#Menghapus hastag
removehastag<-function(x){gsub("#\\S+","",x)}
tweet.corpus<-
tm_map(tweet.corpus,content_transformer(removehastag))
removehastag<-
data.frame(text=sapply(tweet.corpus,as.character),
stringsAsFactors = FALSE)
View(removehastag)
```

```
write.csv(removehastag,                file                =
"C:\\BISMILLAH\\PROSES\\removehastagbaru.csv")

#Remove Emoticon
removeemoticon<-function(x){gsub("[^\x01-\x7F]","", x)}
tweet.corpus<-
tm_map(tweet.corpus,content_transformer(removeemoticon))
remove_emoticon<-
data.frame(text=sapply(tweet.corpus,as.character),
stringsAsFactors = FALSE)
View(remove_emoticon)

#Remove Punctuation(tandabaca)
tweet.corpus<-tm_map(tweet.corpus, removePunctuation)
removepunctuation<-
data.frame(text=sapply(tweet.corpus,as.character),
                              stringsAsFactors = FALSE)
View(removepunctuation)
#removeRT
removeRT<-function(x){gsub("rt","", x)}
tweet.corpus<-
tm_map(tweet.corpus,content_transformer(removeRT))
removeRT<-data.frame(text=sapply(tweet.corpus,as.character),
stringsAsFactors = FALSE)
View(removeRT)

#9. Normalisasi Kata
spell.lex<-read.csv('C:\\SKRIPSSI\\colloquial-indonesian-
lexicon.csv', sep = ",", header=T)
spell.correction = content_transformer(function(x, dict){
  words = sapply(unlist(str_split(x, "\\s+")),function(x){
    if(is.na(spell.lex[match(x, dict$slang),"formal"])){
```

```
      x = x
    } else{
      x = spell.lex[match(x, dict$slang),"formal"]
    }
  })
  x = paste(words, collapse = " ")
})
tweet.corpus=tm_map(tweet.corpus,          spell.correction,
spell.lex)
norm<-data.frame(text=sapply(tweet.corpus,
as.character),stringsAsFactors = FALSE)
View(norm)
#Remove Number
tweet.corpus<-tm_map(tweet.corpus, removeNumbers)
number<-data.frame(text=sapply(tweet.corpus,
as.character),stringsAsFactors = FALSE)
View(number)
write.csv(number,
file="C:\\BISMILLAH\\PROSES\\numberbaru.csv")

#NORMALISASI PERBAIKAN KATA EJAAN ATAU SLANG
stetam1 <- read.csv('C:\\BISMILLAH\\PROSES\\stetambaru.csv',
header=T)
old_stemm <- as.character(stetam1$old)
new_stemm <- as.character(stetam1$new)
stemmword <- function(x)Reduce(function(x,r)

gsub(stetam1$old[r],stetam1$new[r],x,fixed=T),seq_len(nrow(s
tetam1)),x)
dok_slangword=tm_map(tweet.corpus,
content_transformer(stemmword))
```

```
tweet.corpus<-data.frame(text=sapply(dok_slangword,
as.character), stringsAsFactors = FALSE)
View(tweet.corpus)

#11.idstop word
stopwords.id=readLines('C:\\BISMILLAH\\DATA       INPUT\\ID-
Stopwords.txt')
tweet.corpus<-tm_map(dok_slangword,removeWords,stopwords.id)
stop<-data.frame(text=sapply(tweet.corpus,
as.character),stringsAsFactors = FALSE)
View(stop)
write.csv(stop,
file="C:\\BISMILLAH\\PROSES\\FORMULASTOPbaru.csv")

#10. Stemming
stemming                                    =function(x){
paste(sapply(unlist(str_split(x,'\\s+')),katadasar),collapse
= " ")}
tweet.corpus=tm_map(tweet.corpus,
content_transformer(stemming))
stem<-data.frame(text=sapply(tweet.corpus,    as.character),
stringsAsFactors = FALSE)
View(stem)
write.csv(stem,
file="C:\\BISMILLAH\\PROSES\\stemmingbaru.csv")
write.csv(stem,file="C:\\BISMILLAH\\PROSES\\CLEANBARU1.csv")
```

Lampiran 4 Analisis Sentimen

```
#################################################################
#
kalimat2                                                    <-
read.csv('C:\\BISMILLAH\\PROSES\\CLEANBARU2.csv')
#ambil kata kata untuk skoring
#Scoring data berdasarkan kamus kata / ambil kata untuk
skoring
positif     <-     scan("C:\\BISMILLAH\\DATA     INPUT\\s-
pos.txt",what="character",comment.char=";")
negatif     <-     scan("C:\\BISMILLAH\\DATA     INPUT\\s-
neg.txt",what="character",comment.char=";")
kata.positif = c(positif)
kata.negatif = c(negatif)
score.sentimen    =    function(kalimat2,    kata.positif,
kata.negatif, .progress='none')
{
  require(plyr)
  require(stringr)
  scores = laply(kalimat2, function(kalimat, kata.positif,
kata.negatif) {
    kalimat = gsub('[[:punct:]]', '', kalimat)
    kalimat = gsub('[[:cntrl:]]', '', kalimat)
    kalimat = gsub('\\d+', '', kalimat)
    kalimat = tolower(kalimat)

    list.kata = str_split(kalimat, '\\s+')
    kata2 = unlist(list.kata)
    positif.matches = match(kata2, kata.positif)
    negatif.matches = match(kata2, kata.negatif)
    positif.matches = !is.na(positif.matches)
    negatif.matches = !is.na(negatif.matches)
```

```
    score = sum(positif.matches) - (sum(negatif.matches))
    return(score)
  }, kata.positif, kata.negatif, .progress=.progress )
  scores.df = data.frame(score=scores, text=kalimat2)
  return(scores.df)
}
#melakukan skoring text
hasil    =    score.sentimen(kalimat2$text,    kata.positif,
kata.negatif)
head(hasil)
#CONVERT SCORE TO SENTIMEN
hasil$klasifikasi<-                    ifelse(hasil$score<0,
"Negatif","Positif")
hasil$klasifikasi
View(hasil)

#Tukar Row
data <- hasil[c(3,1,2)]
View(data)
write.csv(data, file = "C:\\BISMILLAH\\PROSES\\DATA SENTIMEN
TERLABELI5.csv")
#Memisahkan twit
data.pos <- hasil[hasil$score>0,]
View(data.pos)
write.csv(data.pos,   file   =   "C:\\BISMILLAH\\PROSES\\DATA
POSITIFBARU.csv")
data.neg <- hasil[hasil$score<0,]
View(data.neg)
write.csv(data.neg,   file   =   "C:\\BISMILLAH\\PROSES\\DATA
NEGATIFBARU.csv")
```

Lampiran 5 Syntax R Algoritma Naïve Bayes Classifier

```
#######################################
#Naive bayes classifier
library(tm)
library(RTextTools)
library(e1071)
library(dplyr)
library(caret)
library(klaR)
library(MASS)
library(lattice)
library(lava)
library(naivebayes)
df<-      read.csv("C:\\BISMILLAH\\PROSES\\DATA      SENTIMEN
TERLABELI5.csv", stringsAsFactors = FALSE)
glimpse(df)
set.seed(1)
df$klasifikasi <- as.factor(df$klasifikasi)
corpus <- Corpus(VectorSource(df$text))
corpus
inspect(corpus[1:5])
dtm <- DocumentTermMatrix(corpus)
inspect(dtm[1:10, 1:20])
df.train <- df[1:2725,]
df.test <- df[2726:3406,]
dtm.train <- dtm[1:2725,]
dtm.test <- dtm[2726:3406,]
corpus.train <- corpus[1:2725]
corpus.test <- corpus[2726:3406]
fivefreq <- findFreqTerms(dtm.train, 10)
length((fivefreq))
```

```
fivefreq
dtm.train.nb       <-       DocumentTermMatrix(corpus.train,
control=list(dictionary = fivefreq))
dim(dtm.train.nb)
dtm.test.nb <- DocumentTermMatrix(corpus.test,
                                  control=list(dictionary  =
fivefreq))
#Fungsi Pelabelan
convert_count <- function(x) {
  y <- ifelse(x > 0,1,0)
  y <- factor(y, levels=c(0,1), labels=c("No", "Yes"))
  y
}
trainNB <- apply(dtm.train.nb, 2, convert_count)
testNB <- apply(dtm.test.nb, 2, convert_count)
trainNB

# Train the classifier
system.time(      classifier      <-      naiveBayes(trainNB,
df.train$klasifikasi, laplace = 1) )
# Use the NB classifier we built to make predictions on the
test set.
system.time( pred <- predict(classifier, newdata=testNB) )
# Create a truth table by tabulating the predicted class
labels with the actual class labels
table("Predictions"= pred,  "Actual" = df.test$klasifikasi)
# Prepare the confusion matrix
conf.mat <- confusionMatrix(pred, df.test$klasifikasi)
conf.mat
conf.mat$byClass
conf.mat$overall
conf.mat$overall['Accuracy']
```

Lampiran 6 Syntax R Algortima SVM

```
library(tm)
library(RTextTools)
library(caTools)
library(e1071)
library(dplyr)
library(caret)
library(lattice)
library(klaR)
library(MASS)
library(ROSE)
library(lava)


###SVM
cf    =    read.csv("C:\\BISMILLAH\\PROSES\\DATA    SENTIMEN
TERLABELI4.csv")
cf$klasifikasi = factor(cf$klasifikasi)
corpus2 = Corpus(VectorSource(cf$text))
dtm = DocumentTermMatrix(corpus2)
View(cf)

dtMatrix  =  create_matrix(cf["text"],  language  =  "id",
removeStopwords = FALSE,
                           removeNumbers = FALSE, stemWords =
FALSE, tm::weightTfIdf)

#kernel linear
container   =   create_container(dtMatrix,   cf$klasifikasi,
trainSize = 1:2714, testSize = 2715:3392, virgin = FALSE)

model = train_model(container, "SVM", kernel = "linear")
model
result = classify_model(container, model)
result$SVM_LABEL = factor(result$SVM_LABEL)
confussionMatrix                                              =
confusionMatrix(cf$klasifikasi[2715:3392],
result[,"SVM_LABEL"])
confussionMatrix
confussionMatrix$byClass
```

Lampiran 7 *Output Syntax*

- *Naïve Bayes Classifier*

```
> conf.mat
Confusion Matrix and Statistics

          Reference
Prediction Negatif Positif
   Negatif     105      15
   Positif       9      54

               Accuracy : 0.8689
                 95% CI : (0.8112, 0.9141)
    No Information Rate : 0.623
    P-Value [Acc > NIR] : 1.243e-13

                  Kappa : 0.7159

 Mcnemar's Test P-Value : 0.3074

            Sensitivity : 0.9211
            Specificity : 0.7826
         Pos Pred Value : 0.8750
         Neg Pred Value : 0.8571
             Prevalence : 0.6230
         Detection Rate : 0.5738
   Detection Prevalence : 0.6557
      Balanced Accuracy : 0.8518

       'Positive' Class : Negatif
```

- *Support Vector Machine*

Linear

```
Call:
svm.default(x = container@training_matrix, y = cc
    cost = cost, cross = cross, probability = TRU


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  100

Number of Support Vectors:  762
```

```
> confussionMatrix
Confusion Matrix and Statistics

          Reference
Prediction Negatif Positif
   Negatif     111       3
   Positif       5      64

               Accuracy : 0.9563
                 95% CI : (0.9157, 0.9809)
    No Information Rate : 0.6339
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9064

 Mcnemar's Test P-Value : 0.7237

            Sensitivity : 0.9569
            Specificity : 0.9552
         Pos Pred Value : 0.9737
         Neg Pred Value : 0.9275
             Prevalence : 0.6339
         Detection Rate : 0.6066
   Detection Prevalence : 0.6230
      Balanced Accuracy : 0.9561

       'Positive' Class : Negatif
```

Radial Basic Function

```
> model = train_model(container, "SVM",
> model

Call:
svm.default(x = container@training_matr
    cost = cost, cross = cross, probabi


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  100

Number of Support Vectors:  1128
```

```
> confussionMatrix
Confusion Matrix and Statistics

          Reference
Prediction Negatif Positif
   Negatif     111       3
   Positif      15      54

               Accuracy : 0.9016
                 95% CI : (0.849, 0.9407)
    No Information Rate : 0.6885
    P-Value [Acc > NIR] : 6.271e-12

                  Kappa : 0.7832

 Mcnemar's Test P-Value : 0.009522

            Sensitivity : 0.8810
            Specificity : 0.9474
         Pos Pred Value : 0.9737
         Neg Pred Value : 0.7826
             Prevalence : 0.6885
         Detection Rate : 0.6066
   Detection Prevalence : 0.6230
      Balanced Accuracy : 0.9142

       'Positive' Class : Negatif
```

Sigmoid

```
> model = train_model(container, "SVM", kernel
> model

Call:
svm.default(x = container@training_matrix, y =
    cost = cost, cross = cross, probability = T


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  sigmoid
       cost:  100
      coef.0:  0

Number of Support Vectors:  1239
```

```
> confussionMatrix
Confusion Matrix and Statistics

          Reference
Prediction Negatif Positif
   Negatif     108       6
   Positif      12      57

               Accuracy : 0.9016
                 95% CI : (0.849, 0.9407)
    No Information Rate : 0.6557
    P-Value [Acc > NIR] : 1.163e-14

                  Kappa : 0.787

 Mcnemar's Test P-Value : 0.2386

            Sensitivity : 0.9000
            Specificity : 0.9048
         Pos Pred Value : 0.9474
         Neg Pred Value : 0.8261
             Prevalence : 0.6557
         Detection Rate : 0.5902
   Detection Prevalence : 0.6230
      Balanced Accuracy : 0.9024

       'Positive' Class : Negatif
```