SKRIPSI

PENERAPAN *BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS* (BERT) PADA ANALISIS SENTIMEN VAKSIN *BOOSTER COVID-19*

Oleh

Faizal Ariansyah
065118260

PROGRAM STUDI ILMU KOMPUTER

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS PAKUAN

BOGOR

2024

SKRIPSI

PENERAPAN *BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS* (BERT) PADA ANALISIS SENTIMEN VAKSIN *BOOSTER COVID-19*

Diajukan sebagai salah satu syarat untuk memperoleh

Gelar Sarjana Komputer Jurusan Ilmu Komputer

Fakultas Matematika dan Ilmu Pengetahuan Alam

Oleh

Faizal Ariansyah
065118260

PROGRAM STUDI ILMU KOMPUTER

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS PAKUAN

BOGOR

2024

HALAMAN PENGESAHAN

Judul : Penerapan *Bidirectional Encoder Representations from Transformers* (BERT) Pada Analisis Sentimen Vaksin *Booster COVID-19*
Nama : Faizal Ariansyah
NPM : 065118260

Mengesahkan,

Pembimbing Pendamping FMIPA UNPAK	Pembimbing Utama FMIPA UNPAK
Boldson H. Situmorang, MMSI	Prof. Dr. Sri Setyaningsih, M.Si

Mengetahui,

Ketua Program Studi Ilmu Komputer FMIPA UNPAK	Dekan FMIPA UNPAK
Arie Qur'ania, S.Kom., M.Kom.	Asep Denih, S.Kom., M.Sc., Ph.D.

PERNYATAAN KEASLIAN KARYA TULIS SKRIPSI

Dengan ini saya menyatakan bahwa:

Sejauh yang saya ketahui, karya tulis ini bukan merupakan karya tulis yang pernah dipublikasikan atau sudah pernah dipakai untuk mendapatkan gelar sarjana di Universitas lain, kecuali pada bagian-bagian di mana sumber informasinya dicantumkan dengan cara referensi yang semestinya.

Demikian penyataan ini saya buat dengan sebenar-benarnya. Apabila kelak dikemudian hari terdapat gugatan, penulis bersedia dikenakan sanksi sesuai dengan peraturan yang berlaku.

Bogor, 15 Maret 2024

Faizal Ariansyah

PERNYATAAN PELIMPAHAN SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

==========

Saya yang bertandatangan di bawah ini:

Nama	: Faizal Ariansyah
NPM	: 065118260
Judul Skripsi	: Penerapan *Bidirectional Encoder Representations from Transformers* (BERT) pada Analisis Sentimen Vaksin *Booster COVID-19*

Dengan ini saya menyatakan bahwa Paten dan Hak Cipta dari produk Skripsi dan Tugas Akhir di atas adalah benar karya saya dengan arahan dari komisi pembimbing dan belum diajukan dalam bentuk apapun kepada perguruan tinggi manapun.

Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan Paten, hak cipta dari karya tulis saya kepada Universitas Pakuan.

Bogor, 15 Maret 2024

Faizal Ariansyah
065118260

RIWAYAT HIDUP

Penulis dilahirkan di Lampung pada tanggal 17 Agustus 1999 dari pasangan Bapak Arifin dan Ibu Suhiroh sebagai anak pertama dari dua bersaudara. Penulis memulai pendidikan di Sekolah Dasar yang bertempat di SDN Makasar 09 Pagi di Jakarta, kemudian pada tahun 2011 masuk ke SMPN 287 Jakarta dan penulis adalah alumni dari SMAN 67 Jakarta yang lulus pada tahun 2017.

Pada tahun 2018 penulis melanjutkan pendidikan ke jenjang sarjana di Universitas Pakuan Bogor, Program Studi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam. Pada bulan Mei 2023 penulis memulai penelitian dengan judul Penerapan *Bidirectional Encoder Representations from Transformers* (BERT) pada Analisis Sentimen Vaksin *Booster COVID-19*. Selama penelitian banyak hal baru yang dipelajari oleh penulis dalam menyelesaikan topik tersebut hingga akhirnya pada bulan Januari 2024 peneliti dapat menyelesaikan penelitian tersebut.

RINGKASAN

Faizal Ariansyah 2018. Penerapan *Bidirectional Encoder Representations from Transformers* (BERT) pada Analisis Sentimen Vaksin *Booster COVID-19.* Dibawah bimbingan Prof. Dr. Sri Setyaningsih, M.Si dan Boldson H. Situmorang, MMSI.

Covid-19 merupakan wabah besar yang terjadi di seluruh dunia, termasuk Indonesia, yang berdampak pada semua aspek kehidupan. Untuk meminimalkan angka kematian dan mengurangi tingkat infeksi, WHO sangat menyarankan pemerintah untuk mulai menjalankan vaksin *COVID-19*. Tidak hanya itu, penyuntikan vaksin *booster* juga diperlukan untuk membantu meningkatkan kekebalan masyarakat dalam menghadapi virus *COVID-19* yang sering bermutasi. Keharusan memiliki vaksin *booster* menimbulkan banyak pro dan kontra di masyarakat. Untuk memahami sudut pandang yang berasal dari bahasa Indonesia, penelitian ini melakukan analisis sentimen respon masyarakat terhadap *booster COVID-19* melalui *Twitter* atau *X*. Analisis sentimen dalam penelitian ini menggunakan algoritma *Knowledge Discovery in Database* (KDD) dan *Bidirectional Encoder Representations from Transformers* (BERT) sebagai mesin pembelajaran untuk memroses klasifikasi dan pemodelan data *tweet*. Penelitian ini mendapatkan hasil *graphic loss good fit* dengan akurasi penelitian 85% berdasarkan *confusion matrix* dari 80% data latih dan 20% data uji di antara 1827 data *tweet*. Topik pemodelan dibagi menjadi topik positif yang mencakup topik disiplin, fasilitas, efektivitas dan prestasi sedangkan topik negatif yang mencakup topik efek samping, kekhawatiran dan hilang kepercayaan. Semua penulis memberikan kontribusi yang sama pada penelitian ini.

KATA PENGANTAR

Puji syukur kehadirat Allah SWT, karena rahmat dan hidayah-Nya penulis dapat menyelesaikan skripsi ini yang berjudul: "Penerapan *Bidirectional Encoder Representations from Transformers* (BERT) Pada Analisis Sentimen Vaksin *Booster COVID-19*". Penulisan tugas akhir ini merupakan salah satu syarat memperoleh gelar Sarjana Komputer di Program Studi Ilmu Komputer FMIPA UNPAK Bogor.

Dalam penulisan tugas akhir ini, penulis dengan senang hati ingin mengucapkan terima kasih kepada:

1. Prof. Dr. Sri Setyaningsih, M.Si selaku Pembimbing Utama yang telah memberikan bimbingan, dorongan moril dan motivasi kepada penulis.
2. Boldson H. Situmorang, MMSI selaku Pembimbing Pendamping yang telah memberikan bimbingan, semangat dan motivasi kepada penulis.
3. Arie Qur'ania, S.Kom., M.Kom selaku Ketua Program Studi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Pakuan Bogor.
4. Orang tua dan adik yang tiada henti memberikan semangat serta doa untuk penulis dari memulai hingga menyelesaikan penelitian ini dengan sebaik mungkin.
5. Iftinan Nurjihan Zuhry, S.Si, M.Sc selaku pasangan yang senantiasa memberikan arahan kepada penulis untuk menyelesaikan penelitian ini sedari awal dan membantu memberikan *support* dalam bentuk apapun yang membuat penulis dapat menyelesaikan penelitian ini.
6. Albert Bincar Panorangan selaku rekan terbaik penulis ketika perkuliahan berlangsung sampai saat ini yang senantiasa menemani penulis dalam melaksanakan perkuliahan hingga dapat menyelesaikan penelitian ini.
7. Milky, Jelly dan Simba selaku kucing yang selalu memberi motivasi kepada penulis untuk selalu bersemangat dalam menyelesaikan penelitian ini.
8. Seluruh pihak yang tidak dapat penulis sebutkan satu persatu yang telah memberikan segala dukungan, semangat, bantuan secara langsung maupun tidak langsung.

Penulis menyadari keterbatasan waktu dan kemampuan dalam penulisan tugas akhir ini. Oleh karena itu, segala kritik dan saran yang membangun akan diterima dengan senang hati. Mudah-mudahan Allah SWT akan membalas semua kebaikan kepada semua pihak yang membantu. Akhir kata, semoga laporan ini dapat bermanfaat bagi kita semua. Aamiin.

Bogor, 15 Maret 2024

Faizal Ariansyah

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR TABEL

Halaman

DAFTAR LAMPIRAN

Halaman

BAB I
PENDAHULUAN

1.1 Latar Belakang

Vaksin *booster* adalah vaksin yang dianjurkan oleh badan organisasi kesehatan dunia (WHO) dan juga pemerintahan Negara Kesatuan Republik Indonesia (NKRI) untuk menekan angka penularan dan kematian yang diakibatkan dari merebaknya virus *COVID-19* di Indonesia sebagai dampak dari proses mutasi virus tersebut. Menurut Kemenkes, per tanggal 16 Maret 2022 jangkauan vaksin *booster* masih sangat minim dengan nilai 6,42% secara Nasional yang terbagi ke dalam beberapa kategori, 8,71% pada masyarakat umum dan rentan, 9,38% pada lansia, 12,62% pada petugas pelayanan publik, 0,46% pada kelompok usia 12-17 tahun, dan 0,98% pada usia pendidik. Pemerintah telah menyediakan vaksin Pfizer dengan tingkat efikasi yang sangat baik berdasarkan penelitian yang telah dilakukan yaitu 94,6% dan diyakini memiliki efek samping yang minim. Data penerimaan vaksin di Indonesia per 11 Desember 2023 yaitu, dosis satu sebanyak 203.878.540 dosis, dosis dua 174.965.967 dosis dan *booster* sebanyak 72.931.361 dosis (Balaputra, 2022).

Media sosial merupakan perangkat lunak yang dapat membuat suatu kelompok atau individu melakukan komunikasi untuk berdiskusi, berkumpul, berbagi, bermain dan beragam aktivitas lainnya. Pada tahun 2003 hingga saat ini media sosial mengalami perkembangan yang begitu pesat dengan memunculkan banyak *platform* yang dapat dipilih masyarakat untuk digunakan sesuai fungsinya. Salah satu media sosial yang sering digunakan dalam beropini merupakan media sosial *twitter* atau *X*, karena pada media sosial tersebut biasanya berisikan berita-berita yang sedang hangat diperbincangkan di dalam maupun di luar negeri. Salah satu yang sedang hangat diperbincangkan belakangan ini adalah vaksin *booster COVID-19* (Sari *et al*., 2018).

Twitter atau *X* merupakan salah satu *platform* media sosial yang mengizinkan para penggunanya untuk berbagi pesan baik teks, foto maupun video. *Twitter* memiliki keunggulan yang salah satunya yaitu memudahkan para penggunanya dalam menerima informasi secara singkat dan padat melalui fitur *trending* yang difasilitasi. Informasi yang diberikan dapat dengan mudah disebarluaskan oleh para penggunanya dengan fitur yang telah disediakan oleh *twitter* itu sendiri (Askaria, 2019).

Vaksin *booster* merupakan salah satu program pemerintah dalam rangka menyukseskan program vaksinasi di Indonesia agar bisa segera menekan angka penyebaran virus *COVID-19* dan beralih ke masa *new normal* untuk pemulihan ekonomi pada skala Nasional. Media sosial menjadi salah satu media yang dimanfaatkan untuk mengkampanyekan perihal vaksin *booster* tersebut, salah satu media sosial paling mudah untuk menarik perhatian publik merupakan media sosial *twitter* dikarenakan media sosial tersebut memiliki fitur *trending* yang memungkinkan para penggunanya melihat informasi yang sedang hangat diperbincangkan termasuk perihal vaksin *booster* (Yousefinaghani *et al*., 2021).

Pada masa pandemi pemerintah tentunya sangat ingin menekan angka penyebaran *COVID-19* yang salah satunya yaitu dengan cara menerapkan vaksin untuk masyarakat Indonesia. Pemerintah telah menyiapkan dosis vaksin yang akan diberikan kepada warga negara Indonesia dengan dosis satu, dua dan *booster*, namun tentunya hal

tersebut tidaklah mudah mengingat negara Indonesia terdiri dari banyak sekali pulau, provinsi, kota dan kabupaten yang tentunya akan menjadi tantangan tersendiri untuk pemerintah dalam pendistribusian vaksin terutama untuk masyarakat yang tinggal di pedalaman desa. Penyuluhan tentang pentingnya vaksin *COVID-19* juga dirasa penting agar masyarakat dapat mengerti jenis-jenis vaksin yang disediakan oleh pemerintah. Vaksin *COVID*-19 menjadi topik ulasan masyarakat dalam media sosial salah satunya adalah *twitter*. Analisis sentimen adalah bidang penelitian yang memeriksa dan menganalisis opini publik, sentimen, penilaian, perilaku dan emosi terhadap suatu produk, pelayanan, organisasi, individu, berita, peristiwa, topik dan atribut-atribut yang menyertainya. Analisis sentimen sudah banyak sekali digunakan untuk melihat *progress* perusahaan-perusahaan besar di dunia. Analisis sentimen pada media sosial *twitter* dengan topik vaksin *booster COVID-19* ini merupakan salah satu metode untuk melihat respon masyarakat Indonesia (Narulita, 2019).

Bidirectional Encoder Representations from Transformers (BERT) merupakan salah satu metode klasifikasi yang dapat diaplikasikan pada sebuah analisis sentimen untuk mengetahui sentimen yang dihasilkan berdasarkan dari data *tweets* yang diambil pada media sosial *twitter*. Model BERT hanya memerlukan sedikit pembelajaran lagi untuk mencapai titik optimal dan telah dilatih dengan empat miliar kata dengan sekitar 250 juta bahasa Indonesia termasuk di dalamnya, jadi metode klasifikasi ini dinilai cocok dengan analisis sentimen (Faisal & Mahendra, 2022).

Penelitian tentang topik analisis sentimen *COVID-19* pernah dilakukan oleh Wang *et al*. (2020) dengan judul *COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via* BERT *Model* menghasilkan akurasi 75,65% dengan pembagian rata-rata pada *precision* 0,7133%, *recall* 0,7173% dan *F1-score* 0,7150%. Pada penelitian ini teknik *crawling* yang digunakan yaitu dengan *library SciPy* dan juga menghapus data duplikat yang ada. Pelabelan dilakukan secara manual yang terbagi menjadi sentimen positif, netral dan negatif. Pelatihan data dilakukan dengan empat kali pengulangan. Model yang dihasilkan dapat diimplementasikan sebagai *platform online* dalam memonitor sentimen masyarakat secara *real-time* jika terulang krisis di masa yang akan datang.

Penelitian tentang topik analisis sentimen dilakukan oleh Pratama dan Romadhony (2020) dengan judul Identifikasi Komentar Toksik Dengan BERT menggunakan dataset UGC mentah yang menghasilkan akurasi rata-rata *precision* 68,0110%, *recall* 65,3011% dan *F1-score* 66,2838%. Analisa menggunakan dataset UGC yang dinormalisasikan tanpa penghapusan *stopword* dan lematisasi menghasilkan akurasi rata-rata *precision* 67,1588%, *recall* 65,7703% dan *F1-score* 66,2337%. Analisa menggunakan dataset UGC yang dinormalisasikan dengan penghapusan *stopword* dan lematisasi menghasilkan akurasi rata-rata *precision* 65,3395%, *recall* 61,5547% dan *F1-score* 66,8886%. Analisa menggunakan dataset UGC yang dinormalisasikan dengan lematisasi dan tanpa penghapusan *stopword* menghasilkan akurasi rata-rata *precision* 67,5508%, *recall* 65,5085% dan *F1-score* 66,2222%.

Penelitian tentang topik analisis sentimen terhadap vaksin *COVID-19* pernah dilakukan oleh Yulita *et al*. (2021) berdasarkan penelitian dengan judul Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin *COVID-19* Menggunakan

Algoritma *Naive Bayes Classifier* menghasilkan akurasi yang baik yaitu 93% dari 3780 data *tweets*.

Penelitian tentang topik analisis sentimen *COVID*-19 pernah dilakukan oleh Naseem *et al*. (2021) berdasarkan penelitian dengan judul *COVIDSenti A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis* menghasilkan kesimpulan yaitu sentimen positif memuncak pada topik *lockdown* dan *stay at home* di bulan Februari 2020, namun opini tersebut berubah pada pertengahan Maret, dilihat dari sentimen negatif yang dominan pada dua topik tersebut. Perubahan sentimen tersebut kemungkinan dikarenakan banyaknya konspirasi, miskonsepsi dan misinformasi terhadap *covid-19*. Akurasi yang dihasilkan pada penelitian ini sebesar 94% dengan metode BERT.

Penelitian tentang topik analisis sentimen *COVID*-19 pernah dilakukan oleh Ainapure *et al*. (2023) berdasarkan penelitian dengan judul *Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches* menampilkan *lexicon* dan *deep learning-based approaches* bertujuan untuk memahami perasaan masyarakat terhadap pandemi covid-19 dan vaksinnya berdasarkan postingan pada media sosial *twitter* dalam bahasa Inggris. Analisis sentimen dilakukan dengan (i) *lexicon-based techniques* menggunakan alat sebagai berikut: *VADER* dan *NRCLex*, dan (ii) metode *deep learning* seperti *Bi-LSTM* dan GRU. *Tweets* diklasifikasikan menjadi sentimen positif, negatif, dan netral. Nilai sentimen dan berbagai efek emoji pada *tweets* vaksinasi dihitung. Berdasarkan hasil, dapat disimpulkan bahwa kebanyakan *tweet* dataset mengenai vaksinasi mengandung sentiment positif. Dengan menggunakan pendekatan *Bi-LSTM*, akurasi pada klasifikasi mencapai 92,7% dan 91,24% menggunakan GRU pada tweets covid-19. Untuk *tweets* vaksinasi, akurasi yg dihasilkan sebanyak 92,48% dengan *Bi-LSTM* dan 93,03% dengan model GRU.

Penelitian menggunakan metode BERT pernah dilakukan oleh Ashraf *et al*. (2023) berdasarkan penelitian dengan judul *BERT-Based Sentiment Analysis for Low-Resourced Languages: A Case Study of Urdu Language* ini mendalami analisis sentimen yang dikhususkan pada penggunaan Bahasa yang terbatas, yaitu Bahasa Urdu. Untuk melakukan klasifikasi sentimen yang efektif pada penggunaan Bahasa ini, studi ini menggunakan dataset baru yang berjudul UDSA-23, dan memanfaatkan basis BERT untuk membuat penyematan kata. Berbagai klasifikator *deep learning-based* dilatih pada penyematan-penyematan kata tersebut. Pendekatan yang diusulkan (USA-BERT) menangani analisis sentiment untuk teks Urdu dengan memproses tinjauan yang diberikan terlebih dahulu. Ini meliputi tokenisasi, dan pembuatan vektor berikutnya untuk tinjauan yang telah diproses sebelumnya. Vektor-vektor ini digunakan untuk pelatihan dan evaluasi model USA-BERT. Proses evaluasi menggunakan prinsip Pareto pada dua kumpulan data berbeda: kumpulan data UCSA-21 dan kumpulan data UDSA-23. Penilaian komparatif ini menggambarkan keunggulan USA-BERT. Performanya mengungguli metode yang ada dengan menunjukkan peningkatan dalam akurasi dan pengukuran f, mencapai peningkatan masing-masing hingga 26,09% dan 25,87% dengan akurasi akhir yang dihasilkan yaitu 89,53% dan 94,81%.

Pada penelitian ini akan dilakukan analisis sentimen terhadap respon masyarakat tentang vaksin *booster COVID*-19 yang didapat melalui data *tweets* dari media sosial *twitter* untuk mengetahui sentimen masyarakat yang dibagi menjadi sentimen positif dan negatif menggunakan metode BERT.

1.2 Tujuan

Tujuan dari penelitian ini adalah untuk melakukan analisis sentimen dengan media sosial *twitter* untuk mengetahui respon masyarakat baik positif maupun negatif terhadap vaksin *booster covid-19* dengan metode klasifikasi BERT.

1.3 Ruang Lingkup

Sistem ini dibatasi pada ruang lingkup sebagai berikut:

1. Analisis sentimen ini menggunakan *API Twitter* yang didapat dari *Twitter Developer Account*.
2. Analisis sentimen ini menggunakan Bahasa pemrograman *python.*
3. Klasifikasi sentimen dilakukan menggunakan metode *Bidirectional Encoder Representations from Transformers* (BERT).
4. *Text editor* yang digunakan dalam memroses analisis sentimen ini adalah *anaconda* untuk *crawling* dan *preprocessing*. *Google colaboratory* untuk klasifikasi teks dan *modeling*.
5. Analisis sentimen berbasis *lexicon* diklasifikasikan menjadi sentimen positif dan negatif.
6. Data *tweet* diambil melalui media sosial *twitter* pada periode 14 Mei 2022 – 24 Mei 2022 dan 29 Januari 2023 – 9 Februari 2023 dengan total data sebanyak 2100 data *tweets*.
7. *Topic modeling* dengan menggunakan metode *Bidirectional Encoder Representations from Transformers* (BERT).
8. *Topic modeling* sentimen positif terdiri dari disiplin, efektivitas, fasilitas dan pencapaian.
9. *Topic modeling* sentimen negatif antara lain efek samping, khawatir dan hilang kepercayaan.

1.4 Manfaat

Penelitian ini diharapkan dapat bermanfaat sebagai berikut:

1. Dapat menerapkan metode BERT untuk menganalisis sentimen terhadap komentar pengguna media sosial *Twitter*.
2. Dapat mengetahui kinerja metode BERT dalam pemodelan topik dan klasifikasi sentimen melalui pengujian validasi.
3. Mengetahui seberapa besar sentimen positif dan negatif masyarakat terhadap vaksin *booster Covid-19*.
4. Dapat ikut menyukseskan program vaksin *booster* yang telah dicanangkan oleh pemerintah pusat.
5. Pemerintah dapat mengetahui opini publik terhadap vaksin terutama vaksin *booster* berdasarkan analisis sentimen yang akan dilakukan.

BAB II
TINJAUAN PUSTAKA

2.1 Landasan Teori

2.1.1 Analisis Sentimen

Analisis Sentimen merupakan salah satu metode penelitian dari *text mining* yang berfungsi untuk melihat persepsi atau subjektivitas khalayak umum terhadap suatu topik pembahasan, kejadian, ataupun permasalahan yang sedang dibicarakan. Analisis sentimen merupakan pengklasifikasian suatu teks menjadi orientasi positif ataupun negatif. Analisis sentimen dibagi menjadi empat pendekatan yaitu: (1) *Machine Learning Approach*, (2) *Lexicon-based approach*, (3) *Role-based approach*, (4) *Statistical Model Approach* (Rachman & Pramana, 2020).

Analisis sentimen adalah proses untuk memahami, mengekstraksi dan mengolah data dalam bentuk tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung pada kalimat opini. Besarnya pengaruh yang dihasilkan pada analisis sentimen menyebabkan penelitian dan aplikasi berbasis analisis sentimen berkembang pesat. Sekitar 20-30 perusahaan yang berbasis di Amerika Serikat memfokuskan layanan analisis sentimen untuk melihat opini masyarakat (Buntoro, 2017).

2.1.2 Analisis Sentimen Berbasis *Lexicon*

Lexicon Based adalah metode yang biasa digunakan pada sebuah analisis sentimen karena efisiensinya. Pada *lexicon based* sumber bahasa berasal dari kamus untuk pedoman analisis sentimen sehingga didapat pengklasifikasian berdasarkan opini misalkan opini positif, netral maupun negatif sehingga kalimat akan lebih mudah diklasifikasikan (Mahendrajaya *et al*., 2019).

2.1.3 Media Sosial *Twitter*

Media sosial di masa kini tidak lagi memandang usia maupun generasi, mulai dari kalangan usia muda sampai usia tua dalam pengoperasian sosial media. Salah satu media sosial yang sangat populer untuk beropini adalah media sosial *twitter*. Berdasarkan data dari *We Are Social* di tahun 2020 pengguna aktif media sosial *twitter* di Indonesia tecatat sebanyak 10,65 juta pengguna dan berada pada posisi lima sebagai *platform* media sosial yang paling sering diakses dan digunakan dengan presentase 56% dengan pengguna dari rentang usia 16-64 tahun (Alkatiri *et al*., 2020).

Twitter merupakan media sosial yang dimiliki dan dikembangkan oleh *Twitter Inc.* yang mengizinkan penggunanya untuk berbagi dan membaca pesan yang biasa disebut dengan kicauan (*tweets*). Twitter mengizinkan para penggunanya untuk mengirim kicauan dengan batasan 140 karakter per-*tweets*, namun pada tahun 2017 *twitter* menguji coba dengan menambah Batasan karakter hingga 280 karakter per-*tweets*. Pengguna *twitter* dibebaskan dalam mengirimkan *tweets* yang bisa disisipkan foto maupun video dan juga pengguna lain dibebaskan untuk berinteraksi dengan pengguna tersebut dengan cara membalas pada *tweets* yang sama (Rosalina *et al*., 2020).

2.1.4 Vaksin *Booster COVID-19*

Virus *COVID-19* atau *SARS Cov-2* telah ditetapkan oleh Badan Kesehatan Dunia (*WHO*) sebagai darurat kesehatan global atau yang dikenal juga sebagai

pandemi. Salah satu upaya pencegahan penularan agar virus *COVID-19* tidak semakin meluas ialah dengan pengembangan pembuatan vaksin. Vaksin dapat melindungi orang-orang baik yang sudah divaksin maupun belum divaksin yang diharapkan dapat mengurangi penyebaran virus dalam satu populasi. Penyebaran virus *SARS CoV-2* ini dari manusia ke manusia dan dapat terputus tanpa kekebalan 100% atau yang disebut *Herd Immunity* yang merupakan manfaat dari program vaksinasi (Sari & Sriwidodo, 2020).

Mutasi virus *COVID-19* yang begitu cepat membuat berbagai varian dari virus tersebut bermunculan, seperti yang paling terakhir terdeteksi dikenal oleh masyarakat dunia sebagai varian *omicron* atau B.1.1.529 yang pertama ditemukan di Afrika Selatan, 24 November 2021. Menurut penelitian, efektivitas vaksin *COVID-19* terhadap varian *omicron* dapat menurun seiring berjalannya waktu. Di Afrika Selatan efikasinya menurun hingga 80%. Penelitian terhadap *booster* vaksin jenis *Pfizer* menunjukkan efikasi sebesar 75% pasca dua minggu penyuntikan. Maka dari itu, para ahli virologi mendesak agar masyarakat sesegara mungkin melakukan vaksinasi dan juga *booster* untuk menekan penyebaran virus *COVID-19* yang sudah bermutasi menjadi berbagai varian (Amalia, 2021).

2.1.5 *Natural Language Processing*

Natural Language Processing (NLP) adalah salah satu dari sekian banyak cabang ilmu kecerdasan buatan atau *artificial intelligence (AI)* yang berfokus pada pengolahan bahasa alami. Penelitian menggunakan *Natural Language Processing (NLP)* ini telah digunakan untuk mendeteksi bahasa yang digunakan para pengguna di media sosial *twitter* dengan hasil akurasi yang lebih baik dibanding dengan menggunakan *machine learning* (Rohman *et al*., 2020).

Natural Language Processing (NLP) yang merupakan cabang dari *Artificial Intelligence (AI)* berkaitan dengan interaksi antara komputer dan bahasa alami manusia. *NLP* bertujuan untuk membuat mesin atau sistem yang mampu memahami arti atau makna dari bahasa manusia lalu membetikan respon yang sesuai. *NLP* juga dapat digunakan untuk mengukur sentimen dan menentukan ada di bagian mana dari bahasa manusia yang tergolong penting (Yunefri *et al*., 2021).

2.1.6 *Bidirectional Encoder Representations from Transformers* (BERT)

Bidirectional Encoder Representations from Transformers (BERT) sebuah metode NLP yang diciptakan oleh *google* dan resmi dirilis secara global pada tahun 2018 (Devlin *et al.*, 2019). BERT merupakan sebuah arsitektur jaringan yang telah dilatih dengan banyak sekali *dataset* dari banyak artikel dengan berbagai macam bahasa. Oleh karenanya dilakukan *training* pada setiap lapisan BERT yang sudah terdapat bobot-bobot yang sangat baik dari kasus-kasus NLP. *Training* tersebut memerlukan tambahan *layer* yang akan digunakan untuk proses klasifikasi. *Layer* tersebut terdiri dari *text, preprocessing, BERT encoder, dropout* dan *classifier* (Pratama & Romadhony, 2020).

2.1.7 *Twitter API*

Twitter menyediakan *API (Application Programming Interface)* yang dirancang untuk memudahkan para pengembang atau *developer* untuk mengambil data dari *twitter* dan kemudian mengolahnya. Data diambil dengan menggunakan program yang disebut *crawling* yang ada pada bahasa pemrograman *python* dengan *library*

tweepy. Data *tweets* dapat diambil berdasarkan *keyword* yang sudah diatur sebelumnya (Razaq *et al*., 2021).

2.1.8 *Python*

Python merupakan bahasa pemrograman multifungsi yang menggabungkan kemampuan dan kapabilitasnya dengan sintaks kode yang sangat terstruktur. Banyak fitur yang disediakan oleh *python*, salah satunya sebagai bahasa pemrograman yang dinamis. *Python* dapat digunakan untuk mengembangkan perangkat-perangkat lunak dan juga dapat dijalankan di berbagai sistem operasi (Buana, 2018).

Python sering difungsikan untuk merancang berbagai program seperti Aplikasi *Mobile*, *Website*, *IOT*, Program *CLI & GUI*, Analisis dan masih banyak lagi. *Python* tergolong sebagai bahasa pemrograman yang mudah dipelajari karena pola sintaks yang rapi dan mudah dipahami. Oleh karena itu pada *project* ini penulis memilih pengantar bahasa pemrograman *python* untuk melakukan analisis sentiment (Melinda *et al*., 2021).

2.1.9 *Deep Learning*

Deep learning merupakan metode pembelajaran mengenai suatu data dengan tujuan untuk membuat abstraksi data secara bertingkat. Abstraksi data tersebut membutuhkan sejumlah *layer* pengolahan data untuk mengimplementasikannya. Representasi tersebut diperoleh secara otomatis dari algoritma *machine learning* (Openg *et al*., 2022).

2.1.10 *Confusion Matrix*

Confusion matrix merupakan tabel untuk menampilkan jumlah klasifikasi data uji yang benar maupun yang salah (Normawati & Prayogi, 2021). Contoh *confusion matrix* untuk klasifikasi ditunjukkan pada Gambar 1.

		Observed	
		True	*False*
Predicted Class	*True*	*True Positive* (TP)	*False Positive* (FP)
	False	*False Negative* (FN)	*True Negative* (TN)

Gambar 1. *Confusion matrix*

Dimana:

1. TP: *True Positif*, yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem.
2. TN: *True Negatif*, yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem.
3. FN: *False Negatif*, yaitu jumlah data negatif namun terklasifikasi salah oleh sistem.
4. FP: *False Positif*, yaitu jumlah data positif namun terklasifikasi salah oleh sistem.

Menunjukkan nilai akurasi merupakan perbandingan antara data yang terklasifikasi benar dengan keseluruhan data yang telah diproses (Pratiwi *et al*., 2020).

2.2 Penelitian Terdahulu

Penelitian terkait topik analisis sentimen terhadap vaksin *booster covid-19* dan penerapan metode BERT pernah dilakukan sebelumnya oleh:

1. Nama penulis : Wang *et al.*
   Tahun : 2020
   Judul : *COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via* BERT *Model*
   Isi : Data yang diambil dari sentimen masyarakat China pada media sosial *Weibo* tentang *COVID-19* menghasilkan akurasi 75,65% dengan pembagian rata-rata pada *precision* 0,7133%, *recall* 0,7173% dan *F1-score* 0,7150%. Pada penelitian ini teknik *crawling* yang digunakan yaitu dengan *library SciPy* dan juga menghapus data duplikat yang ada. Pelabelan dilakukan secara manual yang terbagi menjadi sentimen positif, netral dan negatif. Pelatihan data dilakukan dengan empat kali pengulangan.
2. Nama penulis : Pratama dan Romadhony
   Tahun : 2020
   Judul : Identifikasi Komentar Toksik Dengan BERT
   Isi : Hasil analisa menggunakan dataset UGC mentah yang menghasilkan akurasi rata-rata *precision* 68,0110%, *recall* 65,3011% dan *F1-score* 66,2838%. Analisa menggunakan dataset UGC yang dinormalisasikan tanpa penghapusan *stopword* dan lematisasi menghasilkan akurasi rata-rata *precision* 67,1588%, *recall* 65,7703% dan *F1-score* 66,2337%. Analisa menggunakan dataset UGC yang dinormalisasikan dengan penghapusan *stopword* dan lematisasi menghasilkan akurasi rata-rata *precision* 65,3395%, *recall* 61,5547% dan *F1-score* 66,8886%. Analisa menggunakan dataset UGC yang dinormalisasikan dengan lematisasi dan tanpa penghapusan *stopword* menghasilkan akurasi rata-rata *precision* 67,5508%, *recall* 65,5085% dan *F1-score* 66,2222%.
3. Nama penulis : Yulita *et al.*
   Tahun : 2021
   Judul : Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin *COVID-19* Menggunakan Algoritma *Naive Bayes Classifier*
   Isi : 3780 data *tweets* dengan hasil positif dengan persentase sebesar 60,3% dari 2278 data, netral 34,4% dari 1299 data, dan negatif 5,4% dari 203 data. Penelitian tersebut menghasilkan akurasi yang baik yaitu 93%.
4 Nama penulis : Naseem *et al.*
   Tahun : 2021

	Judul	:	*COVIDSenti A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis*
	Isi	:	Sentimen positif memuncak pada topik *lockdown* dan *stay at home* di bulan Februari 2020, namun opini tersebut berubah pada pertengahan Maret, dilihat dari sentimen negatif yang dominan pada dua topik tersebut. Perubahan sentimen tersebut kemungkinan dikarenakan banyaknya konspirasi, miskonsepsi dan misinformasi terhadap *covid-19*. Akurasi yang dihasilkan pada penelitian ini sebesar 94% dengan metode BERT.
5	Nama penulis	:	Ainapure *et al.*
	Tahun	:	2023
	Judul	:	*Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches* menampilkan *lexicon* dan *deep learning-based approaches*
	Isi	:	Analisis sentimen dilakukan dengan (i) *lexicon-based techniques* menggunakan alat sebagai berikut: *VADER* dan *NRCLex*, dan (ii) metode *deep learning* seperti *Bi-LSTM* dan GRU. *Tweets* diklasifikasikan menjadi sentimen positif, negatif, dan netral. Nilai sentimen dan berbagai efek emoji pada *tweets* vaksinasi dihitung. Berdasarkan hasil, dapat disimpulkan bahwa kebanyakan *tweet* dataset mengenai vaksinasi mengandung sentiment positif. Dengan menggunakan pendekatan *Bi-LSTM*, akurasi pada klasifikasi mencapai 92,7% dan 91,24% menggunakan GRU pada tweets covid-19. Untuk *tweets* vaksinasi, akurasi yg dihasilkan sebanyak 92,48% dengan *Bi-LSTM* dan 93,03% dengan model GRU.
6	Nama penulis	:	Ashraf *et al.*
	Tahun	:	2023
	Judul	:	*BERT-Based Sentiment Analysis for Low-Resourced Languages: A Case Study of Urdu Language*
	Isi	:	klasifikasi sentimen yang efektif pada penggunaan Bahasa ini, studi ini menggunakan dataset baru yang berjudul UDSA-23, dan memanfaatkan basis BERT untuk membuat penyematan kata. Berbagai klasifikator *deep learning-based* dilatih pada penyematan-penyematan kata tersebut. Pendekatan yang diusulkan (USA-BERT) menangani analisis sentiment untuk teks Urdu dengan memproses tinjauan yang diberikan terlebih dahulu. Proses evaluasi menggunakan prinsip Pareto pada dua kumpulan data berbeda: kumpulan data UCSA-21

dan kumpulan data UDSA-23. dengan akurasi akhir yang dihasilkan yaitu 89,53% dan 94,81%.

2.2.1 Tabel Perbandingan Penelitian

Tabel penelitian ini menampilkan penelitian sebelumnya dengan tujuan yang sama yaitu melakukan analisis sentimen namun dengan menggunakan metode klasifikasi yang berbeda, berikut dapat dilihat pada Tabel 1.

Tabel 1. Perbandingan Penelitian

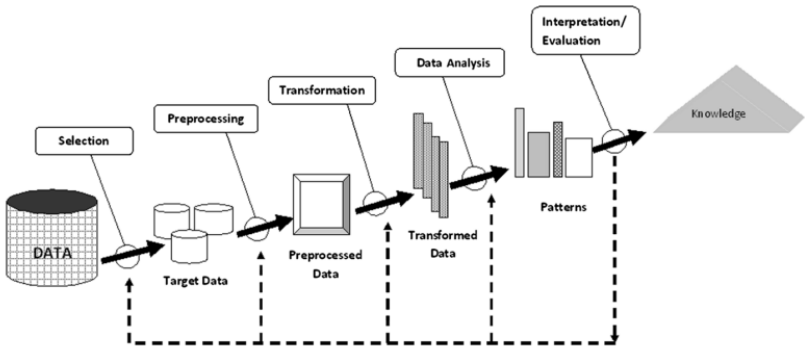
Peneliti	Metode Penelitian			Klasifikasi			Vaksin		Akurasi (%)
	Naive Bayes	BERT	Bi-LSTM	-1	0	1	Dosis 1 & 2	Booster	
Wang *et al.* (2020)	–	√	–	√	√	√	–	–	75,65%
Winda Yulita *et al.* (2021)	√	-	–	√	√	√	√	–	93%
Febrian Adhi Pratama & Ade Romadhony (2020)	-	√	–	√	√	√	–	–	68%
Naseem *et al.* (2021)	-	√	–	√	√	√	–	–	94%
Ainapure *et al.* (2023)	-	–	√	√	√	√	–	–	93%
Ashraf *et al.* (2023)	-	√	–	√	√	√	–	–	95%
Faizal Ariansyah (2022)	-	√	–	√	√	√	–	√	85%

Keterangan:

-1 : Negatif
0 : Netral
1 : Positif

BAB III
METODE PENELITIAN

Analisis sentimen ini akan melalui beberapa metode diantaranya pengumpulan data *tweets* yang bisa didapat dari *crawling* data dengan menggunakan *Twitter API*, *preprocessing* yang terdiri dari *case folding, tokenization, stopwords, normalization* dan *stemmer*; *topic modeling*, klasifikasi BERT, evaluasi dan visualisasi data. Setiap prosesnya tergambar seperti pada Gambar 2.



Gambar 2. Metode KDD (Logan, 2019)

3.1 *Selection* atau Pengumpulan Data

Data *tweets* diambil dari media sosial *twitter* dengan memanfaatkan *Twitter API* yang dapat diperoleh dari *Twitter Developer* dengan cara *crawling*.

3.1.1 *Crawling Data*

Crawling adalah metode pengambilan data yang berasal dari *website* tertentu. Pemilihan data dapat dilakukan secara spesifik sesuai dengan kebutuhan data (Budiarto, 2021).

3.1.2 *Scraping Data*

Mengubah format yang dihasilkan pada saat *crawling* menjadi format yang akan disimpan dan dipakai untuk proses selanjutnya.

3.1.3 *Cleaning Data*

Cleaning data merupakan proses pembersihan data terutama pada data duplikat atau data kosong.

3.1.4 Pelabelan Data atau *Labelling*

Pelabelan data dilakukan secara manual oleh pakar bahasa (Dra. Tiarma R, M.Pd, SMA Negeri 67 Jakarta) dan menggunakan *library* sastrawi yang diolah kembali oleh BERT dengan bahasa pemrograman *python*.

3.2 *Preprocessing*

Preprocessing merupakan permulaan dari semua tahapan dalam proses klasifikasi. Proses dalam *preprocessing* terdiri dari *case folding, normalization, tokenization, stopwords removal* dan *stemmer*. Berikut ini dijabarkan tiap-tiap tahap *preprocessing*:

3.2.1 *Case Folding*

Case folding merupakan proses penggantian huruf yang terkombinasi antara *lowercase* dan *uppercase* menjadi sama rata *lowercase*.

3.2.2 *Normalization*

Normalization merupakan fungsi untuk menyeragamkan *term* yang memiliki makna sama namun penulisan berbeda, yang dapat diakibatkan oleh kesalahan penulisan, penyingkatan kata, ataupun "bahasa gaul".

3.2.3 *Tokenization*

Tokenization merupakan proses pemisahan sebuah kalimat menjadi kata-kata.

3.2.4 *Stopwords Removal*

Stopwords merupakan penghilangan kata-kata yang tidak berguna pada sebuah kalimat dan sering muncul.

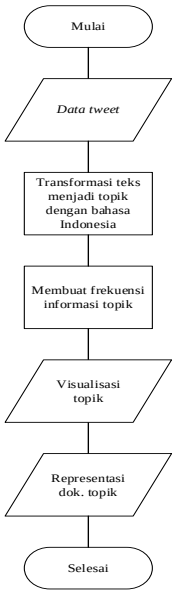
3.2.5 *Stemmer*

Stemmer merupakan pengubahan kata menjadi bentuk dasar, proses ini menyesuaikan dengan struktur yang digunakan pada pemrosesan *stemming*.

3.3 *Transformation* atau *Topic Modeling*

Topic modeling adalah proses pemodelan topik dari distribusi kata yang membentuk sebuah topik dengan topik tertentu. Terdapat dua tahap pada pemodelan topik. Tahap pertama dilakukan pemodelan topik berdasarkan penambahan dan pengurangan jumlah topik. Tahap kedua dilakukan pemodelan berdasarkan banyaknya iterasi. Hasil dari kedua tahap tersebut akan dianalisa dengan cara melakukan perbandingan kata di setiap *cluster*-nya. Topik akan divisualisasikan setelah melewati tahap pemodelan oleh BERT (Alfanzar *et al.*, 2020).

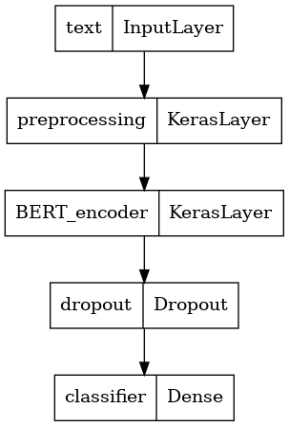
Modeling dilakukan dengan menggunakan BERT sebagai media pembagian topik yang paling sering dibicarakan pada data *tweet*. Alur algoritma *modeling* dapat dilihat pada Gambar 3.



Gambar 3. *Flowchart modeling*

3.4 *Data Analysis* atau Klasifikasi BERT

BERT merupakan sebuah arsitektur jaringan yang sudah dilatih dengan berbagai *dataset* dan berbagai bahasa (Putri *et al.*, 2020). Pada penelitian ini penulis akan menggunakan teknik *fine-tuning* dengan model IndoBERT yang merupakan model yang menggunakan arsitektur $BERT_{base}$. Model ini telah dilatih dengan empat miliar kata yang terkandung sekitar 250 juta kalimat formal dan kalimat sehari-hari dalam bahasa Indonesia (Wilie *et al*., 2020). Algoritma BERT dapat dilihat pada Gambar 4.

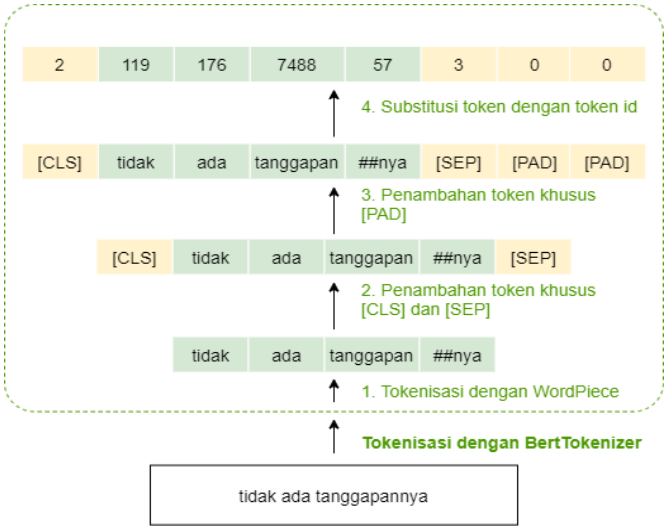


Gambar 4. Algoritma BERT

Dimulai *Text* atau *InputLayer* yaitu teks yang akan diklasifikasikan menggunakan algoritma BERT. Setelah melalu proses *preprocessing* selanjutnya dilakukan proses tokenasiasi oleh *BERT_encoder* yaitu proses yang dilakukan oleh BERT untuk mengklasifikasikan tiap-tiap kata. Selanjutnya, *dropout* yaitu menyisihkan tiap kata yang tidak terdapat pada kamus BERT dari kalimat yang sudah ditokenisasi pada tahap *BERT_encoder*. Terakhir, *classifier* yaitu tahap terakhir yang merupakan tahap klasifikasi dari kalimat yang sudah diolah pada proses-proses sebelumnya.

Library yang digunakan pada penelitian ini adalah *library transformers* yang dikembangkan oleh *HuggingFace*. Ribuan model *pre-trained* telah disediakan oleh *library* ini yang salah satunya dapat digunakan untuk melakukan sebuah klasifikasi teks dalam 100 bahasa. *Transformers* di-*support* oleh dua *library deep learning* yaitu *PyTorch* dan *TensorFlow*.

BertTokenizer dibutuhkan untuk melakukan tokenisasi pada kalimat dan menghasilkan *input* sesuai yang diinginkan. Hal tersebut dilakukan guna menemukan *vocabulary* yang spesifik dengan model yang dipakai. Kalimat yang menjadi representasi input pada BERT diproses melalui tokenizer dan melalui prosedur *stuck encoder* atau mekanisme seperti pada Gambar 5.



Gambar 5. Alur tokenisasi *stuck encoder* klasifikasi BERT

CLS sebagai token pembatas pada sisi kiri tokenisasi atau token pembukaan untuk kalimat awal, SEP sebagai token pembatas pada sisi kanan tokenisasi atau token penutup untuk akhir kalimat dan PAD sebagai token pembatas setelah token SEP dimana token PAD akan hilang pada tahap akhir tokenisasi yang bermanfaat untuk mengetahui panjang kalimat. Semua token tersebut merupakan batas tokenisasi yang diciptakan oleh *BertTokenizer*.

3.5 *Evaluation* atau *Confusion Matrix*

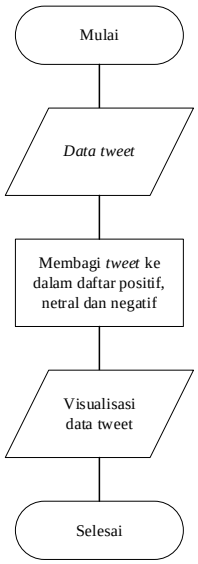
Training dilakukan menggunakan model yang telah dilatih sebelumnya dan hanya melakukan sedikit tambahan proses *training*. Cara ini lebih efisien karena *training* tidak harus dilakukan dari awal. *Training* dilakukan dengan mengunduh data latih yang telah diproses pada tahap sebelumnya. BERT berfungsi untuk menerima kalimat sebagai *input* dan akan terus melewati *stuck encoder*.

Setelah melewati tahap *encoder*, token akan menghasilkan *output* berupa vektor. Analisis sentimen akan menggunakan *output* pertama yaitu token [CLS] karena token tersebut dapat mengumpulkan rata-rata token kata untuk mendapatkan vektor pada kalimat. Vektor tersebut yang akan digunakan sebagai *input* untuk *classifier*. Lapisan terakhir pada *classifier* akan menghasilkan logits, yaitu *output* prediksi probabilitas kasar dari kalimat yang akan diklasifikasi. Data *tweets* yang digunakan untuk data latih dan data uji adalah sebesar 80%:20% data *tweets.*

3.6 *Knowledge* atau Visualisasi Data

Pengujian dilakukan dengan memanfaatkan *confusion matrix* yang didapat dari proses klasifikasi BERT terhadap data *training* dan data *testing* yang telah diklasifikasikan pada tahap sebelumnya untuk melihat berapa akurasi yang didapat dari proses klasifikasi tersebut. Dalam pengujian ini akan menghasilkan nilai *recall*, *precision* dan *accuracy*. Sementara itu hasil representasi akan ditampilkan dalam bentuk tabel yang menampilkan *True Positive*, *False Positive*, *True Negative* dan *False Negative*. Data *tweets* yang digunakan untuk data uji adalah sebesar 20% data *tweets* (Darwis *et al.*, 2020).

Visualisasi data *tweet* ditampilkan dengan menggunakan media *wordcloud*. Alur algoritma *wordcloud* dapat dilihat pada Gambar 6.



Gambar 6. *Flowchart wordcloud*

3.7 Waktu dan Tempat Penelitian

Waktu penelitian dilaksanakan setiap hari kerja dari Senin sampai Sabtu mulai jam 10.00 WIB sampai jam 16.00 WIB. Tempat penelitian dilaksanakan di Laboratorium Komputer Program Studi Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Pakuan.

3.8 Alat dan Bahan

3.8.1 Alat

1. *Hardware*
   a. Laptop
   b. Intel Core i5-12500H 2,56GHz (16CPUs)
   c. SSD 512GB
   d. RAM 16GB 3200MHz
2. *Software*
   a. OS *Windows* 11 *Home Single Language* 64-*bit*
   b. *Jupyter Notebook*
   c. *Google Colaboratory*
   d. *Ms. Word 2021*
   e. *Ms. Excel 2021*
   f. *Ms. Visio 2021*

3.8.2 Bahan

1. Data *tweets* vaksin *booster*
2. Jurnal penelitian terdahulu

BAB IV
PERANCANGAN DAN IMPLEMENTASI

4.1 Perancangan

4.1.1 *Selection* atau Pengumpulan Data

Pengambilan data dilakukan dengan membuat *twitter developer account* lalu data akan diambil dengan bantuan *library tweepy* dan diseleksi dimana *tweet* yang terbuat dengan cara *retweet* akan dikecualikan untuk menghindari duplikasi data pada *jupyter notebook* dengan bahasa pemrograman *python*. Data akan menjadi data mentah yang akan diolah lebih lanjut melalui beberapa proses. Pengambilan data dilakukan dengan cara *crawling* data pada tanggal 14 Mei 2022 – 24 Mei 2022 dan 29 Januari 2023 – 9 Februari 2023 dengan total data sebanyak 2100 data *tweets* yang berbasis dari media sosial *twitter* dengan topik vaksin *booster COVID-19*. Data melalui proses *scraping* yang kemudian diubah menjadi file csv. *Cleaning data* dilakukan dengan menambahkan *source code -retweet* pada *text editor* untuk menghapus duplikasi data *tweet* yang berasal dari *retweet*. *Scraping data* disimpan dalam file CSV. Simulasi pengumpulan data dapat dilihat pada Gambar 7.

	edit_history_tweet_ids	text	created_at	id
0	[1623478468041375744]	Ayo segera vaksin Booster…\n#polresjembrana #...	2023-02-09T00:26:44.000Z	1623478468041375744
1	[1623477494384041984]	Jasa sertifikat vaksin.\nDosis 1,2 & boost...	2023-02-09T00:22:52.000Z	1623477494384041984
2	[1623477308324745216]	Sebanyak 500 dosis vaksin booster tahap 2 dibe...	2023-02-09T00:22:07.000Z	1623477308324745216
3	[1623472655587475456]	@RagilSemar Kewajiban vaksin booster 2 sejatin...	2023-02-09T00:03:38.000Z	1623472655587475456
4	[1623471741602332674]	Yang butuh sertifikat vaksin 1, 2 & Booste...	2023-02-09T00:00:00.000Z	1623471741602332674

Gambar 7. Simulasi pengambilan *data tweet*

4.1.2 *Preprocessing*

4.1.2.1 *Case Folding*

Mengubah semua huruf dalam data *tweets* menjadi huruf kecil. Karakter selain huruf akan dihilangkan dan dianggap *delimiter*. Simulasi *case folding* dapat dilihat pada Tabel 2.

Tabel 2. Simulasi *case folding*

Case Folding	
Percepat herd Immunity, Ayo ikut vaksin Booster. #vaksinbooster #vaksìnasicovid19 #vaksìncovid19 #vaksinaman #vaksinserentak #bersatulawancovid19 #polrestabanan https://t.co/acslWCjrBA	percepat herd immunity, ayo ikut vaksin booster. #vaksinbooster #vaksã¬nasicovid19 #vaksã¬ncovid19 #vaksinaman #vaksinserentak #bersatulawancovid19 #polrestabanan https://t.co/acslwcjrba

4.1.2.2 *Normalization*

Perbaikan kata yang diakibatkan dari kesalahan penulisan, penyingkatan kata ataupun dari “bahasa gaul”. Fungsinya adalah untuk penyeragaman kata dengan makna yang sama dan menghilangkan simbol maupun tanda baca agar hanya tersisa sebuah kalimat. Simulasi *normalization* dapat dilihat pada Tabel 3.

Tabel 3. Simulasi *normalization*

Normalization	
@awah_x @harapanbaru16 @sumandogaek ya sudah klo doyan vaksin ditenggak aja biar puas klo dirasa dienjus kurang. jangan lupa booster biar seterong	ya sudah kalau suka vaksin ditenggak saja biar puas kalau dirasa dienjus kurang jangan lupa booster biar kuat

4.1.2.3 *Tokenization*

Pemisahan kalimat menjadi masing-masing kata untuk dipilih agar memudahkan proses selanjutnya. Simulasi tokenization dapat dilihat pada Tabel 4.

Tabel 4. Simulasi *tokenization*

Tokenization	
vaksin booster untuk tingkat proteksi tubuh dari covid	['vaksin', 'booster', 'untuk', 'tingkat', 'proteksi', 'tubuh', 'dari', 'covid']

4.1.2.4 *Stopwords Removal*

Penghapusan kata yang dinilai terlalu sering muncul atau umum digunakan dan kata yang tidak mempunya informasi berharga pada sebuah kalimat. Simulasi *stopwords removal* dapat dilihat pada Tabel 5.

Tabel 5. Simulasi *stopwords removal*

Stopwords	
['tiap', 'orang', 'dan', 'tiap', 'jenis', 'vaksin', 'beda', 'beda', 'efeknya', 'aku', 'booster', 'az', 'tangan', 'pegal', 'banget', 'beberapa', 'hari', 'dan', 'sempat', 'demam', 'setelah', 'booster']	['orang', 'jenis', 'vaksin', 'beda', 'beda', 'efeknya', 'booster', 'az', 'tangan', 'pegal', 'banget', 'demam', 'booster']

4.1.2.5 *Stemmer*

Penghapusan imbuhan pada tiap-tiap kata pada sebuah kalimat dan mengubah kata pada *tweet* menjadi kata dasar. Simulasi *stemmer* dapat dilihat pada Tabel 6.

Tabel 6. Simulasi *stemmer*

Stopwords	
['satuan', 'tugas', 'penerima', 'vaksin', 'booster', 'corona', 'capai', 'juta', 'orang']	['satu', 'tugas', 'terima', 'vaksin', 'booster', 'corona', 'capai', 'juta', 'orang']

4.1.3 *Transformation atau Topic Modeling*

Untuk melihat sebaran topik yang paling sering dibicarakan pada sentimen positif dan negatif, setiap dokumen memiliki kemungkinan terdapat kesamaan topik.

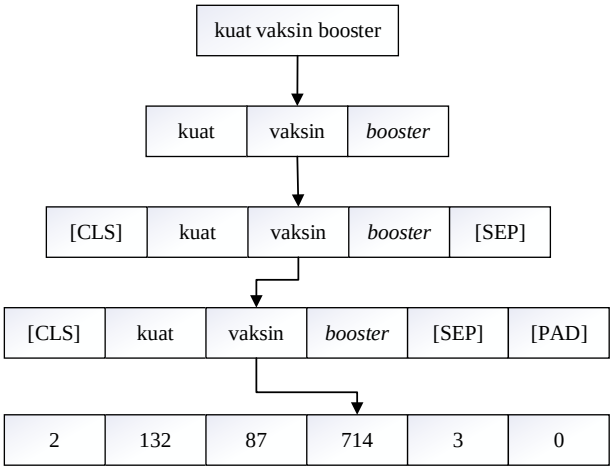
4.1.4 *Data Analysis* atau klasifikasi BERT

Pengklasifikasian akan dilakukan otomatis oleh BERT dengan menginputkan data *tweet* yang telah melalui proses *selection, preprocessing* dan *transformation*. Saat pengklasifikasian dilakukan, BERT memerlukan token [CLS] di awal kalimat dan [SEP] pada akhir kalimat. Token [PAD] digunakan untuk menentukan panjang kalimat. Hal tersebut guna membuat data dapat diklasifikasikan oleh BERT karena BERT membutuhkan token-token tersebut. Data *tweet* dibagi menjadi sentimen positif dan negatif.

Pelatihan oleh BERT Dimulai *Text* atau *InputLayer* yaitu teks yang akan diklasifikasikan menggunakan algoritma BERT. Setelah melalu proses *preprocessing* selanjutnya dilakukan proses tokenasiasi oleh *BERT_encoder* yaitu proses yang dilakukan oleh BERT untuk mengklasifikasikan tiap-tiap kata. Selanjutnya, *dropout* yaitu menyisihkan tiap kata yang tidak terdapat pada kamus BERT dari kalimat yang sudah ditokenisasi pada tahap *BERT_encoder*. Terakhir, *classifier* yaitu tahap terakhir yang merupakan tahap klasifikasi dari kalimat yang sudah diolah pada proses-proses sebelumnya.

4.1.4.1 *Stuck Encoder*

BERT mempelajari representasi input pada teks dengan menggunakan tokenisasi *WordPiece*. *Tokenizer* ini sudah menjadi populer karena bisa menghasilkan representasi vektor untuk kata-kata rancu dan sangat berguna. *Sub-words* dibentuk saat *training*, jadi bergantung pada *dataset training*. Ukuran maksimum kalimat pada BERT terbatas hanya sampai 512 token. Jadi, teks yang lebih dari 512 kata akan dipotong, karena BERT bukan untuk menjalankan tugas yang menghasilkan bahasa (Wijayanti *et al*., 2021). Simulasi *stuck encoder* dapat dilihat pada Gambar 8.



Gambar 8. Simulasi *stuck encoder*

4.2 Implementasi

4.2.1 *Selection* atau Pengumpulan Data

Selection dilakukan dengan menggunakan bahasa pemrograman *python*, *library* yang digunakan untuk melakukan pengumpulan data menggunakan *library tweepy* dan memasukkan kode *consumer key*, *consumer secret*, *access token*, *access token secret* dan *bearer token* yang didapat melalui *twitter developer account*. *Source code crawling*, *scraping* dan *cleaning data* dapat dilihat pada Gambar 9 dan 10.

```
import tweepy

consumer_key="HjLKFAAoJB0FFM4bNWvleraA3"
consumer_secret="VWgsL88nC7Fxx83LOUgbNnaIVY7ZP70iMIXx9J6Ugh9XedXH0N"
access_token="1477206820695412736-khWTjtRkMTs720cb3iQyzyP8p2965F"
access_token_secret="3QIQWq9wmtMAgHbkLrrRU6FcHcad4J7ScCjHqVSpPGPD6"
bearer_token ="AAAAAAAAAAAAAAAAAAAAAEUoaQEAAAAA3iEVXcssuUVdOep6lRi2P6wkGcQ%3DQATDx9NzwV7xWPWd4P8ZIYI6uuSPpqMLDvfLzM2SCO3MgAfN0W"

import requests

client = tweepy.Client( bearer_token=bearer_token,
                        consumer_key=consumer_key,
                        consumer_secret=consumer_secret,
                        access_token=access_token,
                        access_token_secret=access_token_secret,
                        return_type = requests.Response,
                        wait_on_rate_limit=True)

# query to search for tweets
query = "vaksin booster lang:id -is:retweet"

# your start and end time for fetching tweets
start_time = "2023-02-09T00:00:00Z"
end_time = "2023-02-09T00:30:00Z"

# get tweets from the API
tweets = client.search_recent_tweets(query=query,
                                     start_time=start_time,
                                     end_time=end_time,
                                     tweet_fields = ["created_at"],
                                     max_results = 100)
```

Gambar 9. *Source code* pengumpulan *data tweet* 1

```
import pandas as pd

# Save data as dictionary
tweets_dict = tweets.json()

# Extract "data" value from dictionary
tweets_data = tweets_dict['data']

# Transform to pandas Dataframe
df = pd.json_normalize(tweets_data)
```

```
df
```

	edit_history_tweet_ids	text	created_at	id
0	[1623478468041375744]	Ayo segera vaksin Booster...\n#polresjembrana #...	2023-02-09T00:26:44.000Z	1623478468041375744
1	[1623477494384041984]	Jasa sertifikat vaksin.\nDosis 1,2 & boost...	2023-02-09T00:22:52.000Z	1623477494384041984
2	[1623477308324745216]	Sebanyak 500 dosis vaksin booster tahap 2 dibe...	2023-02-09T00:22:07.000Z	1623477308324745216
3	[1623472655587475456]	@RagilSemar Kewajiban vaksin booster 2 sejatin...	2023-02-09T00:03:38.000Z	1623472655587475456
4	[1623471741602332674]	Yang butuh sertifikat vaksin 1, 2 & Booste...	2023-02-09T00:00:00.000Z	1623471741602332674

```
df.to_csv('Vaksin_Booster_0809Feb.csv', index=False)
```

Gambar 10. *Source code* pengumpulan *data tweet* 2

4.2.2 *Preprocessing*

4.2.2.1 *Case Folding*

Proses *case folding* dilakukan pada *jupyter notebook* untuk menyeragamkan *tweet* menjadi sama rata *lower case* dengan memanfaatkan *library string*. *Source code case folding* dapat dilihat pada Gambar 11.

```
df['text'] = df['text'].str.lower()
print('\nHasil case folding: ')
print(df.head())
```

Gambar 11. Source case folding

4.2.2.2 *Normalization*

Proses *normalization* dilakukan pada *jupyter notebook* dengan memanfaatkan *library string* dan bantuan *tools* "kamus alay" untuk mengubah kata-kata yang tidak baku menjadi kata dasar baku. *Normalization* juga dilakukan guna menghapus karakter selain karakter huruf dari a sampai z. *Source code* normalization dapat dilihat pada Gambar 12 dan 13.

```
def remove_data(Text):
    Text = Text.replace('\\t'," ").replace('\\n'," ").replace('\\u'," ").replace('\\',"")
    Text = ' '.join(re.sub("([@#][A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)"," ",Text).split())
    return Text.replace("http://", " ").replace("https://", " ")

df['text'] = df['text'].apply(remove_data)

def remove_number(Text):
    return re.sub(r"\d+", "", Text)

df['text'] = df['text'].apply(remove_number)

def remove_punctuation(Text):
    return Text.translate(str.maketrans("","",string.punctuation))

df['text'] = df['text'].apply(remove_punctuation)

def remove_whitespace_LT(Text):
    return Text.strip()

df['text'] = df['text'].apply(remove_whitespace_LT)

def remove_whitespace_multiple(Text):
    return re.sub('\s+',' ',Text)

df['text'] = df['text'].apply(remove_whitespace_multiple)

def remove_singl_char(Text):
    return re.sub(r"\b[a-zA-Z]\b", "", str(Text))

df['text'] = df['text'].apply(remove_singl_char)

df.head()
```

Gambar 12. *Source code normalization* 1

```
alay_dict = pd.read_csv(r'C:\Users\arian\Tugas Akhir\Tools\new_kamusalay.csv', encoding='latin-1', header=None)
alay_dict = alay_dict.rename(columns={0: 'original',
                                      1: 'replacement'})
alay_dict_map = dict(zip(alay_dict['original'], alay_dict['replacement']))
def normalize_alay(text):
    return ' '.join([alay_dict_map[word] if word in alay_dict_map else word for word in text.split(' ')])

df['text'] = df['text'].apply(normalize_alay)

print('\nHasil normalisasi: ')
print(df.head())
```

Gambar 13. *Source code normalization* 2

4.2.2.3 *Tokenization*

Proses *tokenization* dilakukan pada *jupyter notebook* untuk membuat kata-kata dalam kalimat *tweet* terpisah menjadi satuan kata dengan memanfaatkan *library nltk.tokenize*. *Source code tokenization* dapat dilihat pada Gambar 14.

```
def word_tokenize_wrapper(text):
    return word_tokenize(text)

df['text'] = df['text'].apply(word_tokenize_wrapper)
print('Hasil tokenizing: ')
print(df.head())
```

Gambar 14. *Source code tokenization*

4.2.2.4 *Stopwords Removal*

Proses *stopwords removal* dilakukan pada *jupyter notebook* untuk menghilangkan kata yang umum digunakan dan juga tidak mempunyai informasi berharga pada *tweet* dengan memanfaatkan *library nltk.corpus*. *Source code stopwords removal* dapat dilihat pada Gambar 15.

```
factory = StemmerFactory()
stemmer = factory.create_stemmer()

def stemmer_wrapper(term):
    return stemmer.stem(term)

term_dict = {}

for document in df['text']:
    for term in document:
        if term not in term_dict:
            term_dict[term] = ' '

print(len(term_dict))
print("----------------------")

for term in term_dict:
    term_dict[term] = stemmer_wrapper(term)
    print(term, ":" , term_dict[term])

print(term_dict)
print("----------------------")

def get_stemmed_term(document):
    return [term_dict[term] for term in document]

df['text'] = df['text'].apply(get_stemmed_term)
print('Hasil stemmer: ')
print(df.head())
```

Gambar 15. *Source code stemmer*

4.2.2.5 *Stemmer*

Proses *stemmer* dilakukan pada *jupyter notebook* guna menghapus imbuhan dan pengubahan kata pada *tweet* menjadi kata dasar dengan memanfaatkan *library Sastrawi.Stemmer.Stemmerfactory*. *Source code stemmer* dapat dilihat pada Gambar 16.

```
def stopward_removal(text):
    filtering =  stopwords.words('indonesian','english')
    x = []
    df = []
    def myFunc(x):
        if x in filtering:
            return False
        else:
            return True
    fit = filter(myFunc, text)
    for x in fit:
        df.append(x)
    return df

df['text'] = df['text'].apply(stopward_removal)
print('Hasil stopwords: ')
print(df.head())
```

Gambar 16. *Source code stopwords removal*

4.2.3 *Transformation* atau *Topic Modeling*

Proses *topic modeling* dilakukan pada *google colaboration* guna mengetahui topik yang menjadi *highlight* pada sentimen positif dan negatif vaksin *booster covid-19* dengan memanfaatkan *library BERTtopic*. *Source code topic modeling* dapat dilihat pada Gambar 17 dan 18.

```
model = BERTopic(language="indonesian")

topics, probs = model.fit_transform(docs)
```

Gambar 17. *Source code topic modeling* 1

Mengatur bahasa yang digunakan dalam pemodelan ke dalam bahasa Indonesia karena *tweet* yang berbahasa Indonesia dan mentransformasikan topik menjadi bentuk model sesuai dengan data *tweet* yang dimasukkan.

```
freq = model.get_topic_info()
print("Number of topics: {}".format( len(freq)))
freq.head()

#model.get_topic_freq()
```

Gambar 18. *Source code topic modeling* 2

Melihat jumlah dan kalimat topik yang muncul setelah dilakukan pemodelan data *tweet* oleh BERT.

4.2.4 *Data Analysis* atau klasifikasi BERT

Data *tweet* diklasifikasikan dengan algoritma BERT untuk menentukan akurasi akhir pada data *tweet* positif, netral dan negatif. *Source code* klasifikasi BERT dapat dilihat pada Gambar 19 dan 20. *Source code* dan akurasi pada tiap *epoch* dapat dilihat pada Lampiran 5.

```
from sklearn.model_selection import train_test_split

train_input, test_input, train_labels, test_labels = train_test_split(input_ids,
                                                                      labels,
                                                                      random_state=42,
                                                                      test_size=0.2)
train_mask, test_mask, _, _ = train_test_split(attention_mask,
                                               labels,
                                               random_state=42,
                                               test_size=0.2)

train_input, validation_input, train_labels, validation_labels = train_test_split(train_input,
                                                                                  train_labels,
                                                                                  random_state=42,
                                                                                  test_size=0.2)
train_mask, validation_mask, _, _ = train_test_split(train_mask,
                                                     train_mask,
                                                     random_state=42,
                                                     test_size=0.2)
```

Gambar 19. *Source code klasifikasi* BERT 1

Membagi data latih dan data uji, data validasi menyesuaikan data uji yaitu 80:20.

```
optimizer = torch.optim.AdamW(
    model.parameters(),
    lr = 2e-5,
    eps = 1e-8
)
```

Gambar 20. *Source code klasifikasi* BERT 2

Learning rate yang digunakan yaitu 2e-5 menggunakan rekomendasi dari penelitian terdahulu.

4.2.4.1 Stuck Encoder

Kalimat pada data *tweet* diubah menjadi token agar data dapat diklasifikasikan oleh BERT. *Source code stuck encoder* dapat dilhat pada Gambar 21.

```
input_ids = []

for sent in sentences:
  encoded_sent = tokenizer.encode(
      sent,
      add_special_tokens = True
  )
  input_ids.append(encoded_sent)

print("Original: ", sentences[217])
print("Token IDs: ", input_ids[217])
```

Gambar 21. *Source code stuck encoder*

Menambahkan token CLS, SEP dan PAD pada masing-masing kalimat (data *tweet*) untuk dilakukan *stuck encoder* oleh BERT *tokenizer*.

BAB V
HASIL DAN PEMBAHASAN

5.1 Hasil

5.1.1 *Selection* atau Pengumpulan Data

Data *tweet* diambil melalui media sosial *twitter* atau yang sekarang telah berganti nama menjadi *X* pada rentang tanggal 14 Mei 2022 – 24 Mei 2022 dan 29 Januari 2023 – 9 Februari 2023 dengan total data sebanyak 1827 data *tweets* setelah melalui proses *cleaning*. Hasil pengumpulan data dapat dilihat pada Tabel 7.

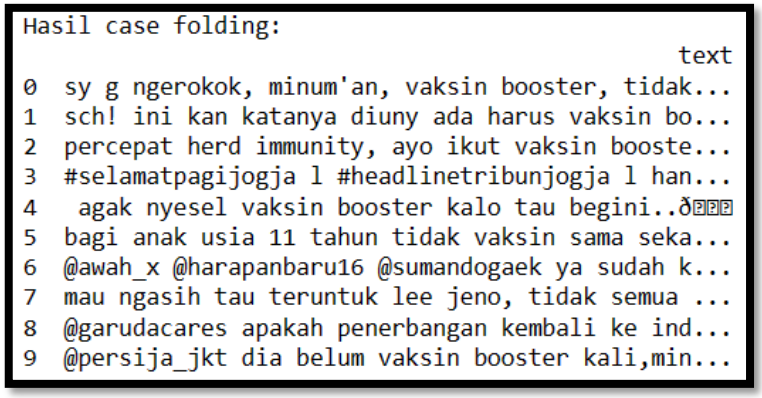
Tabel 7. Hasil pengumpulan *data tweet*

created_at	no_record	Text
2022-05-20T23:51:41.000Z	1,53E+18	Anak 5-11 Tahun di AS Sudah Bisa Vaksin Booster Pakai Pfizer Infrastruktur Jokowi Terbaik https://t.co/BXfdpyVIjl
2022-05-20T23:39:15.000Z	1,53E+18	@irmnfirmnsyah Urg mh geus vaksin booster jadi teu kudu antigen
2022-05-20T23:13:12.000Z	1,53E+18	@rapfunjelxx Kalo belum vaksin booster gabole
2022-05-20T23:06:22.000Z	1,53E+18	Anak di Amerika Bisa Vaksin Booster, di Indonesia Bagaimana? https://t.co/BDp4VnM633
2022-05-20T22:56:26.000Z	1,53E+18	@aliyuita Udah gw ulas juga. Efek samping drugs ya liver ancur. Vaksin itu drugs, tambah drugs lagi ketika meriang 3 hari akibat vaksin. Trus booster. Drugs, booster, drugs. Kenalah. Itu yg mati kemaren, antek receh kasebul, dieharder pembenci ulama, mati kena sirosis. Efek vaksin
2022-05-20T22:17:54.000Z	1,53E+18	@aeodroma kurang baik di tangan kiri. kemaren abis vaksin booster.
2022-05-20T22:15:39.000Z	1,53E+18	@iJ0NGSE0NG udah vaksin booster ya
2022-05-20T21:35:30.000Z	1,53E+18	Vaksin booster membuatku cenut² hot potato².

5.1.2 *Preprocessing*

5.1.2.1 *Case Folding*

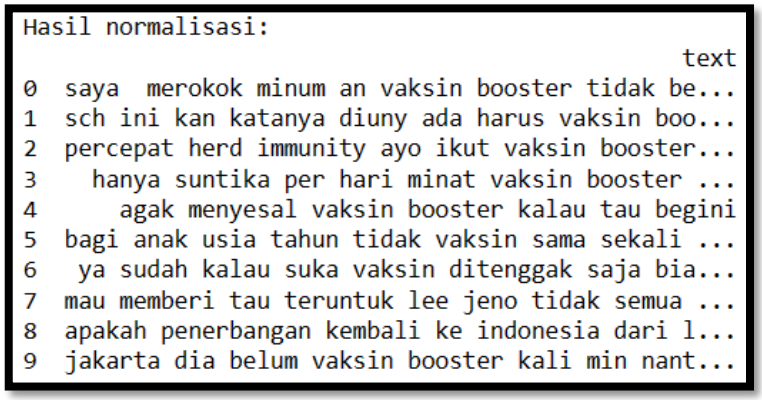
Data *tweet* yang telah dikumpulkan menjadi satu dataset kemudian akan melalui tahap-tahap *preprocessing* dimana *case folding* menjadi proses pertama yang dilewati dengan menyeragamkan semua kalimat menjadi *lower case*. Hasil *case folding* dapat dilihat pada Gambar 22.



Gambar 22. Hasil *case folding*

5.1.2.2 *Normalization*

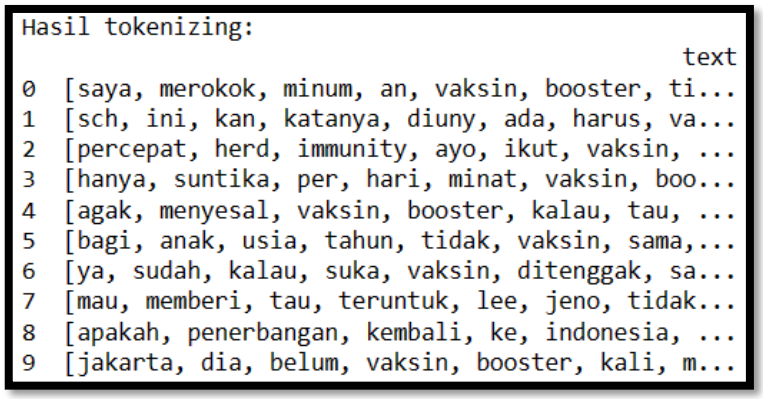
Setelah melalu proses *case folding* data *tweet* akan melalui proses *normalization* untuk menghapus simbol-simbol, tanda baca dan penyeragaman kata. Hasil *normalization* dapat dilihat pada Gambar 23.



Gambar 23. Hasil *normalization*

5.1.2.3 *Tokenization*

Setelah proses *normalization* data *tweet* melalui proses *tokenization* guna memisahkan kalimat menjadi kata per kata. Hasil dari *tokenization* dapat dilihat pada Gambar 24.



Gambar 24. *Hasil tokenization*

5.1.2.4 *Stopwords Removal*

Setalah kalimat dipisah menjadi kata per kata pada proses *tokenization* selanjutnya data *tweet* akan memasuki proses *stopwords removal* untuk menghilangkan kata yang sering muncul dan tidak memiliki informasi berharga. Hasil *stopwords removal* dapat dilihat pada Gambar 25.

```
Hasil stopwords:
                                                text
0  [merokok, minum, an, vaksin, booster, berpenya...
1  [sch, diuny, vaksin, booster, vaksin, booster,...
2  [percepat, herd, immunity, ayo, vaksin, booste...
3  [suntika, minat, vaksin, booster, menurun, pas...
4                    [menyesal, vaksin, booster, tau]
5  [anak, usia, vaksin, pacar, pesawat, domestik,...
6  [ya, suka, vaksin, ditenggak, biar, puas, dira...
7  [tau, teruntuk, lee, jeno, orang, kuat, lihat,...
8  [penerbangan, indonesia, negeri, diwajibkan, t...
9  [jakarta, vaksin, booster, kali, min, kasih, b...
```

Gambar 25. Hasil *stopwords removal*

5.1.2.5 *Stemmer*

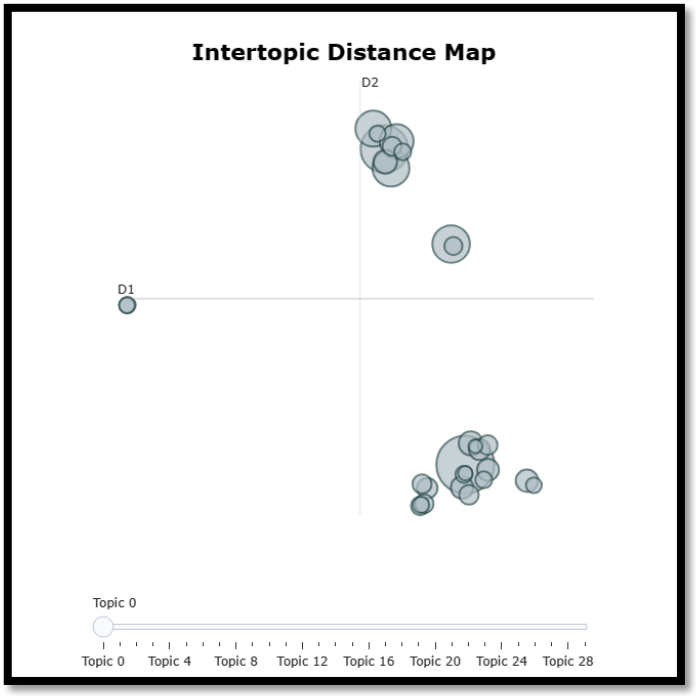
Setelah kata per kata yang serimg muncul dihilangkan pada proses *stopwords removal* selanjutnya data *tweet* akan memasuki tahap terakhir *preprocessing* yaitu *stemmer* untuk menghapus imbuhan yang terdapat pada kalimat. Hasil dari *stemmer* dapat dilihat pada Gambar 26.

```
Hasil stemmer:
                                                text
0  [rokok, minum, an, vaksin, booster, sakit, dtn...
1  [sch, diuny, vaksin, booster, vaksin, booster,...
2  [cepat, herd, immunity, ayo, vaksin, booster, ...
3  [suntika, minat, vaksin, booster, turun, pasca...
4                      [sesal, vaksin, booster, tau]
5  [anak, usia, vaksin, pacar, pesawat, domestik,...
6  [ya, suka, vaksin, tenggak, biar, puas, rasa, ...
7  [tau, untuk, lee, jeno, orang, kuat, lihat, je...
8  [terbang, indonesia, negeri, wajib, tes, pacar...
9  [jakarta, vaksin, booster, kali, min, kasih, b...
```

Gambar 26. Hasil *stemmer*

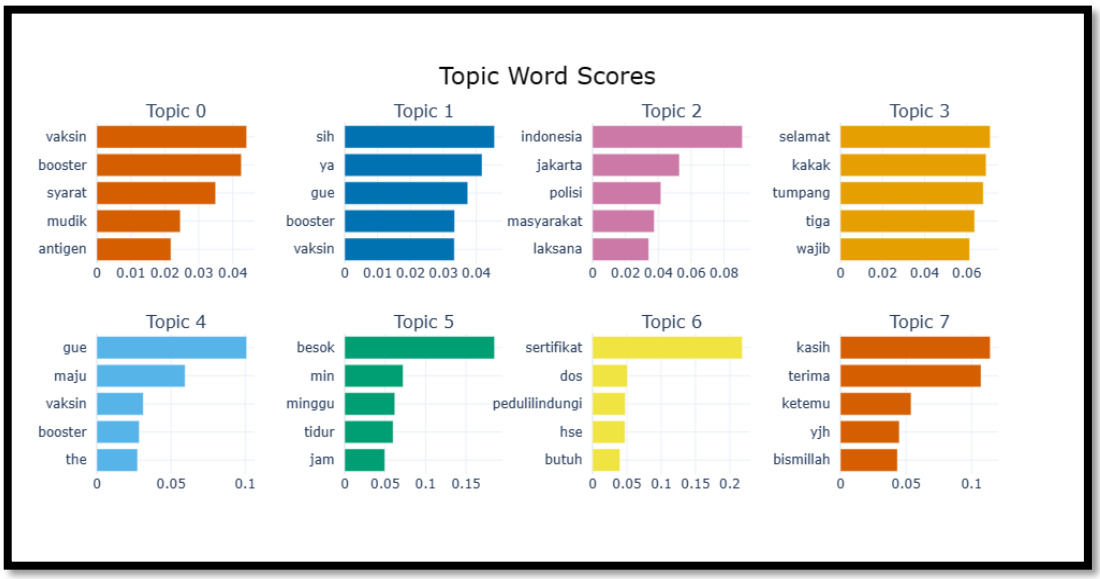
5.1.3 *Transformation* atau *Topic Modeling*

Tahap *topic modeling* untuk mengetahui topik dari sentimen yang paling sering dibicarakan pada *tweet* mengenai vaksin *booster* yang dibagi menjadi dua sentimen yaitu topik pada sentimen positif dan topik pada sentimen negatif. Pemodelan dilakukan dengan menggunakan algoritma BERT. Metode BERT merupakan *unsupervised learning machine* penulis diharuskan menyimpulkan topik dari pola yang sudah dibuat oleh BERT (Dharmawan *et al.*, 2023). Hasil dari *topic modeling* sentimen positif dapat dilihat pada Gambar 27 dan 28. *Source code* dan hasil dari *topic modeling* sentimen negatif dapat dilihat pada Lampiran 4.



Gambar 27. *Map topic modeling* sentimen positif

Map penyebaran topik yang dihasilkan dari pemodelan BERT pada sentimen positif sebanyak 29 topik. Kesamaan antar topik sangat mungkin terjadi karena pemodelan BERT merupakan *unsupervised learning*.



Gambar 28. *Barchart topic modeling* sentimen positif

Penyebaran delapan topik yang divisualisasikan dengan *barchart* sesuai dengan hasil pemodelan yang telah dilakukan oleh BERT.

Dari topik yang didapat oleh BERT di atas penulis menyimpulkan topik yang paling relevan terhadap sentimen positif dan negatif dari data *tweet* vaksin *booster covid-19*. Topik tersebut tersebar sebagai berikut:

5.1.3.1 *Topic Modeling* Positif

Sentimen positif yang dilakukan pemodelan menghasilkan topik-topik yang paling sering dibicarakan sebagai berikut. Hasil *topic modeling* positif dapat dilihat pada Tabel 8.

Tabel 8. *Topic modeling* positif

Tweet	Representasi dokumen	Topik
vaksin *booster* ya pakai masker, habis buka masker wajib vaksin *booster*, ayo vaksin *booster* dosis patuh prokes cuci tangan pakai masker jaga jarak mari jaga prokes pakai masker ruang tempat ramai lupa lengkap dosis vaksin *covid*	masker, habis, tangan, patuh, lepas, jaga, prokes, diam, cuci, pakai	**DISIPLIN (34% atau 443 *tweets*)**
vaksin *booster* tingkat proteksi tubuh *covid*, vaksin *booster* tingkat proteksi tubuh *covid*, vaksin *covid booster*	*covid*, proteksi, infeksi, tingkat, ayo, tubuh, layan, biak, ribet, uni	**EFEKTIVITAS (20% atau 261 *tweets*)**
vaksin *booster* syarat tes *antigen*, vaksin *booster* syarat jalan, syarat wajib vaksin *booster* ya	vaksin, *booster*, syarat, mudik, *antigen*, sehat, lengkap, jarak, selenggara, *fess*	**FASILITAS (27% atau 352 *tweets*)**
satu tugas juta orang vaksin *booster covid*, satu tugas terima vaksin *booster covid* capai juta orang, satu tugas terima vaksin *booster covid* capai juta orang	juta, orang, satu, tugas, capai, *covid*, terima, suntik, ratus, profesi	**PENCAPAIAN (8% atau 104 *tweets*)**

5.1.3.2 *Topic Modeling* Negatif

Sentimen negatif yang dilakukan pemodelan menghasilkan topik-topik yang paling sering dibicarakan sebagai berikut. Hasil *topic modeling* negatif dapat dilihat pada Tabel 9.

Tabel 9. *Topic modeling* negatif

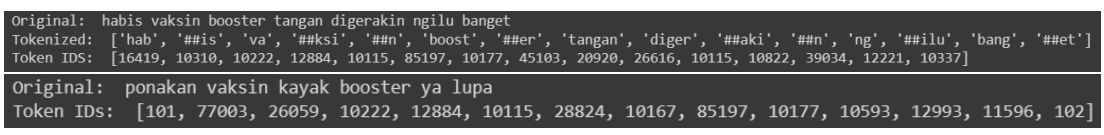
Tweet	Representasi dokumen	Topik
gue sakit habis vaksin *booster*, sakit banget habis vaksin *booster*, habis *booster* siang kipi nya asa kecuali pegal ringan banget bekas	vaksin, *booster*, *covid*, banget, *covid*, pegal, sakit, habis, ya, efek, demam	**EFEK SAMPING (33% atau 103 *tweets*)**

Tweet	Representasi dokumen	Topik
suntik beda banget vaksin primer pakai az demam vaksin priner demam hari *booster* pegal nya lumayan *booster* pegal nya *mild* banget		
takut vaksin *booster*, takut vaksin *booster*, takut vaksin *booster*	takut, vaksin, *booster*, Indonesia, nya, bondong, dana, jarum, ya, gara	**KHAWATIR (8% atau 25 *tweets*)**
mbak februari positif *covid* marah orang percaya *covid* vaksin pakai tes di*covid*kan pas vaksin *booster* marah bikin gumpal darah mandul mandul nular orang, habis vaksin *booster* alhamdulillah pakai lemas demam vaksin jam pagi bangun gigil badan hangat tidur bangun jam an kondisi banget langsung nge mal, kaget bangun kepala kayak pecah badan sakit banget gerak tau efek vaksin *booster* jam *delay* ya *side effect*snya	sakit, bangun, vaksin, banget, kayak, *booster*, jam, *covid*, nenek, marah	**HILANG KEPERCAYAAN (15% atau 47 *tweets*)**

5.2 Pembahasan

5.2.1 *Stuck Encoder*

Stuck encoder menjadi bagian dari tahap klasifikasi yang dijalankan oleh BERT. Sebelum memasuki tahap klasifikasi, tiap-tiap kalimat *(data tweet)* dibagi menjadi per suku kata untuk melalui proses *stuck encoder* dan diubah menjadi token-token angka sebelum memasuki proses klasifikasi. Hasil *stuck encoder* dapat dilihat pada Gambar 29.



```
array([  101, 77003, 26059, 10222, 12884, 10115, 28824, 10167, 85197,
       10177, 10593, 12993, 11596,   102,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0,     0,
           0,     0,     0,     0,     0,     0,     0,     0])
```

Gambar 29. Hasil *stuck encoder*

Data *tweet* yang sudah diubah menjadi vektor berupa angka setelah melalui tahapan *stuck encoder*. Angka 101 menunjukan token CLS, angka 102 menunjukan token SEP dan angka 0 menunjukan token PAD.

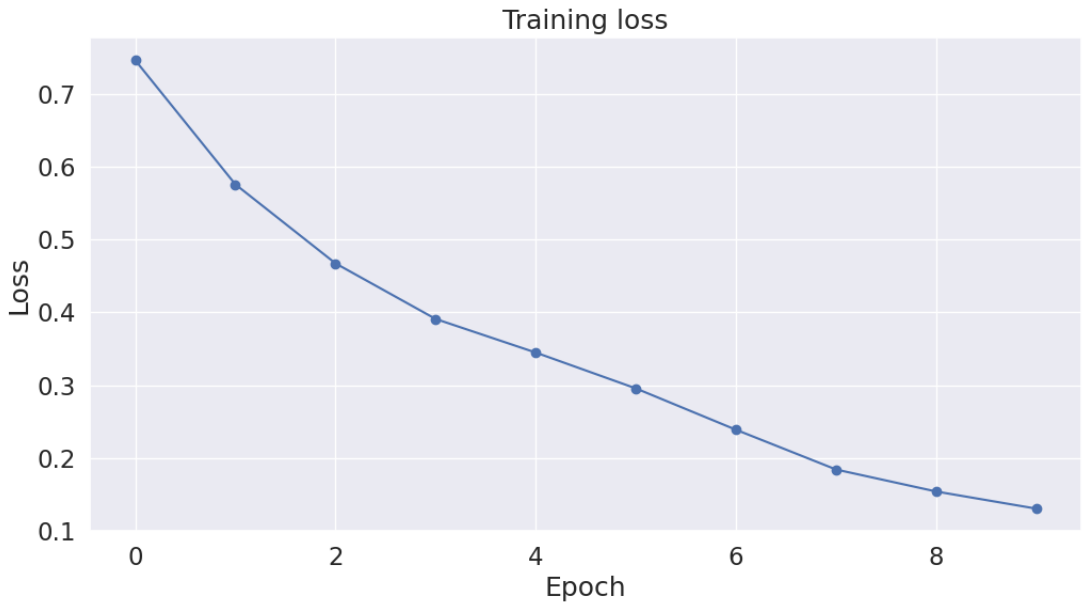
5.2.2 *Graphic Loss*

Pada penelitian ini grafik yang dihasilkan menunjukan grafik *good fit* yaitu dataset menunjukkan hasil yang optimal karena grafik yang dihasilkan melandai ke bawah dapat dilihat pada Gambar 30. Hasil dari *graphic loss* terbagi menjadi tiga yaitu *underfitting, overfitting* dan *good fit*. Masing-masing jenis *graphic loss* memiliki penjelasan berbada terkait dari yang dihasilkan dari klasifikasi yang telah diterapkan dengan metode BERT.

Underfitting dimana sewaktu-waktu, *validation loss* dapat lebih besar daripada *training loss*. Hal ini menunjukkan bahwa model tersebut kurang sesuai. *Underfitting* terjadi ketika model tidak dapat memodelkan data latih secara akurat, sehingga menghasilkan *error* yang signifikan. Selanjutnya hasil pada skenario ini menunjukkan bahwa diperlukan pelatihan lebih lanjut untuk mengurangi kerugian yang terjadi selama data dilatih. Jalan lain yang dapat diambil yaitu, meningkatkan data latih dengan memperoleh lebih banyak sampel atau dengan menambah data.

Overfiitting yaitu secara khusus model ini memiliki performa yang baik pada data latih, namun kurang baik pada data baru di set validasi. Pada titik tertentu, kerugian validasi akan berkurang tetapi akan meningkat lagi. Alasan penting terjadinya hal ini adalah karena model mungkin terlalu rumit untuk datanya atau karena model tersebut dilatih dalam jangka waktu yang lama. Dalam hal tersebut, latihan dapat dihentikan ketika kerugiannya rendah dan stabil, hal ini biasanya disebut penghentian dini. Penghentian dini adalah salah satu dari banyak pendekatan yang digunakan untuk mencegah *overfitting*.

Good fit yaitu menunjukkan kesesuaian yang optimal, yaitu model yang tidak *overfit* atau *underfit*. Menghasilkan penurunan *graphic loss* yang landai dan konsisten menurun. Pada set validasi data menunjukkan kesesuaian dengan data latih, maka dari itu *graphic loss* melandai menurun. Pada penelitian ini penulis mendapatkan grafik *good fit*.



Gambar 30. *Graphic loss* klasifikasi BERT

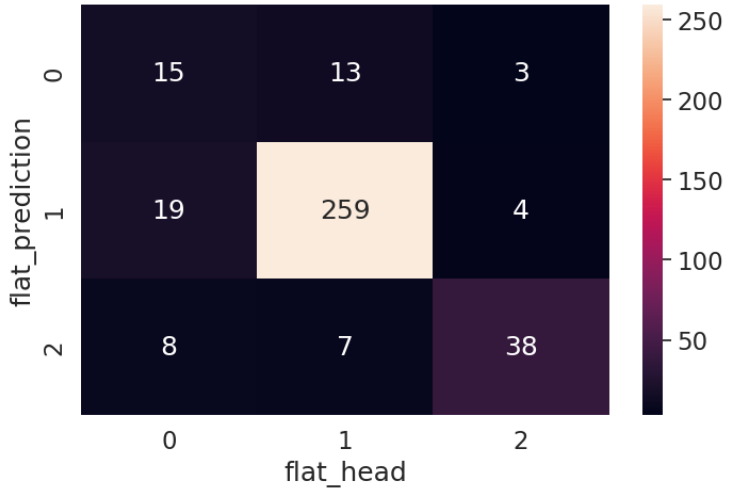
Klasifikasi menghasilkan hasil grafik loss melandai turun (*good fit*) yang berarti semakin sedikit data mengalami error saat pengujian dikarenakan data yang diuji sesuai dengan data latih yang telah melalui proses validasi.

5.3 Evaluasi

Pada analisis sentimen terhadap vaksin *booster covid-19* penulis membagi evaluasi menjadi dua bagian yaitu pada keseluruhan data *tweet* yang membahas tentang vaksin *booster covid-19* yang mencakup sentimen positif, netral dan negatif dan juga pada *topic modeling* terhadap sentimen positif dan negatif dimana nantinya akan diketahui besaran akurasi yang didapat pada *topic modeling* di setiap topik dari keseluruhan sentimen yang sudah ditetapkan. Data latih terbagi menjadi 60%, 70%, 80% dan 90% dan data uji terbagi menjadi 40%, 30%, 20% dan 10% pada masing-masing sektor.

5.3.1 *Confusion Matrix* Keseluruhan Sentimen

Terdapat 1827 data dimana data terbagi menjadi 1305 sentimen positif dan 315 sentimen negatif. Data latih dibagi menjadi 80% dengan pembagian yaitu 1023 sentimen positif dan 262 sentimen negatif. Data uji dibagi menjadi 20% dengan pembagian yaitu 282 sentimen positif dan 53 sentimen negatif. Data latih dan data uji ini memiliki akurasi terbaik. Akurasi yang dihasilkan sebagai berikut. Perbandingan akurasi dapat dilihat pada Lampiran 6. *Heatmap confusion matrix* dapat dilihat pada Gambar 31.



Gambar 31. *Heatmap confusion matrix* klasifikasi BERT

Pembagian perhitungan akurasi dari peta *confusion matrix* di atas didapat sebagai berikut di tiap-tiap sentimen. Dapat dilihat pada Tabel 10.

Tabel 10. *Confusion matrix* klasifikasi BERT

	Precission	*Recall*	*F1-Score*	*Support*
0	0,36	0,48	0,41	31
1	0,93	0,92	0,92	102
2	0,84	0,72	0,78	53
Accuracy			0,85	366
Macro avg	0,71	0,71	0,70	366
Weighted avg	0,87	0,85	0,86	366

Perhitungan manual *confusion matrix* untuk *precission*, *recall* dan *accuracy* dapat dilihat melalui cara di bawah ini.

Sentimen positif (1):
TP: 259
FN: 19 + 4 = 23
FP: 13 + 7 = 20
TN: 15 + 3 + 8 + 38 = 64

Precission: 259 / (259 + 20) = 0,93
Recall: 259 / (259 + 23) = 0,92

Sentimen negatif (2):
TP: 38
FN: 8 + 7 = 15
FP: 3 + 4 = 7
TN: 15 + 13 + 19 + 259 = 306

Precission: 38 / (38 + 7) = 0,84
Recall: 38 / (38 + 15) = 0,72

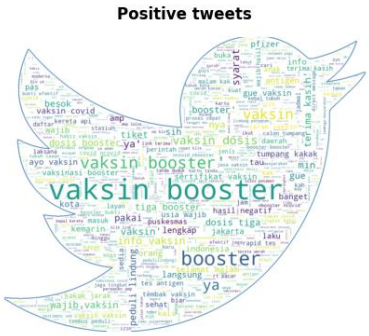
Macro avg:
Total precission: (0,36 + 0,93 + 0,84) / 3 = 0,71
Total recall: (0,48 + 0,92 + 0,72) / 3 = 0,71

5.4 Representasi Pengetahuan

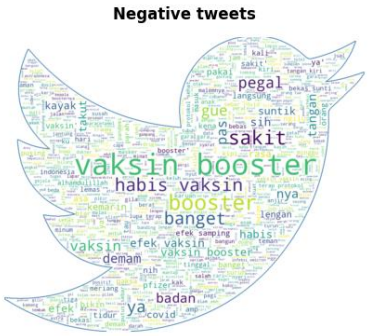
Visualisasi akan ditampilkan dengan media *wordcloud* pada klasifikasi sentimen BERT dan juga *topic modeling* BERT yang sudah dijalankan dengan membagi sentimen ke dalam tiga sentimen yaitu sentimen positif dan negatif. Berikut pembagian visualisasi data melalui media *wordcloud*.

5.4.1 Klasifikasi BERT

Data *tweet* mengenai vaksin *booster covid-19* divisualisasikan melalui *wordcloud* dengan *masking* logo *twitter* agar lebih menarik untuk divisualisasikan sesuai dengan sumber data *tweet* yang diperoleh. Berikut *wordcloud* dari sentimen positif dan negatif dari data *tweet* tentang vaksin *booster covid-19* dengan maksimal 1000 kata yang paling sering dibicarakan di tiap-tiap sentimen. Dapat dilihat pada Gambar 32 dan 33.



Gambar 32. *Wordcloud sentimen* positif



Gambar 33. *Wordcloud sentimen* negatif

5.4.2 *Topic Modeling* BERT

Topic modeling dilakukan untuk merepresentasikan *tweet* pada sentimen positif dan negatif. Hasil dari *topic modeling* sentimen positif dan negatif ditampilkan dalam bentuk *wordcloud* dapat dilihat pada:

1. *Topic modeling* positif
   Visualisasi model dengan *wordcloud* dari topik disiplin, efektivitas, fasilitas dan pencapaian. *Wordcloud* dapat dilihat pada Gambar 34 sampai 37.

Gambar 34. *Wordcloud topic* disiplin

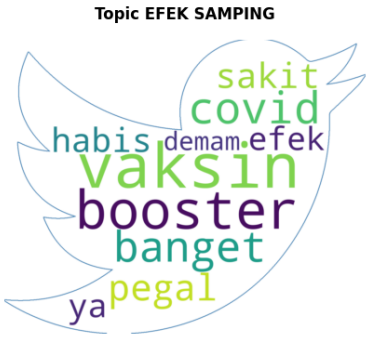
Gambar 35. *Wordcloud topic* fasilitas

Gambar 36. *Wordcloud topic* efektivitas

Gambar 37. *Wordcloud topic* pencapaian

2. *Topic modeling* negatif

Visualisasi model dengan *wordcloud* dari topik efek samping, khawatir, hilang kepercayaan. *Wordcloud* dapat dilihat pada Gambar 38 sampai 40.



Gambar 38. *Wordcloud topic* efek samping

Gambar 39. *Wordcloud topic* khawatir

Gambar 40. *Wordcloud topic* hilang kepercayaan

BAB VI
PENUTUP

6.1 Kesimpulan

Pada penelitian ini penulis dapat menyimpulkan sebagai berikut:

1. Penggunaan metode BERT dipilih karena metode tersebut telah terbukti akurat dalam melakukan klasifikasi teks. Penelitian terdahulu dengan metode yang sama menghasilkan akurasi klasifikasi teks yang baik dimana akurasi mencapai di atas 60%. BERT juga dapat diterapkan untuk melakukan *topic modeling* pada teks, dimana teks akan terbagi menjadi beberapa topik pembicaraan yang paling sering muncul.
2. Hasil yang didapat penulis pada penelitian ini dengan menerapkan metode BERT dapat diaplikasikan pada data *tweet* dengan menghasilkan *output* sesuai dengan algoritma yang sudah direncanakan.
3. Pada penelitian ini penulis mendapatkan total akurasi 85% dari klasifikasi teks yang dilakukan dengan metode BERT dari data *tweet* sebanyak 1827 *tweets* yang terbagi menjadi 1305 sentimen positif dan 315 sentimen negatif. Data latih dan data uji yang diterapkan berskala 80:20% dengan rincian data uji sebagai berikut; 282 sentimen positif 53 sentimen negatif.
4. Hasil dari *topic modeling* pada penelitian ini dibagi ke dalam sentimen positif dan negatif. Topik pada sentimen positif antara lain disiplin sebanuak 443 data *tweets*, efektivitas sebanyak 261 data *tweets*, fasilitas sebanyak 352 data *tweets* dan pencapaian sebanyak 104 data *tweets*. Topik pada sentimen negatif antara lain efek samping sebanyak 103 data *tweets*, khawatir sebanyak 25 data *tweets* dan hilang kepercayaan sebanyak 47 data *tweets*.

6.2 Saran

Pada penelitian lanjutan yang akan datang diharapkan ada pengembangan tentang *topic modeling* dengan algoritma BERT. Dilakukan analisis terhadap sistem untuk pengklasifikasian dan pemodelan oleh BERT agar mendapatkan efisiensi waktu, penambahan VRAM sangat memengaruhi *loading time* dari proses klasifikasi dan *modeling*. Penambahan metode lain untuk perbandingan terhadap BERT terutama pada akurasi dan pemodelan.

DAFTAR PUSTAKA

Ainapure, B. S., Pise, R. N., Reddy, P., Appasani, B., Srinivasulu, A., Khan, M. S., & Bizon, N. (2023). Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches menampilkan lexicon dan deep learning-based approaches. *Sustainability*, *15*(3), 1–21.

Albadani, B., Shi, R., & Dong, J. (2022). A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM. *Applied System Innovation*, *5*(13), 1–16.

Alfanzar, A. I., Khalid, & Rozas, I. S. (2020). Topic Modeling Skripsi Menggunakan Metode Latent Dirichlet Allocation. *Jurnal Sistem Informasi*, *7*(1), 7–13.

Alkatiri, A. B. M., Nadiah, Z., & Nasution, A. N. S. (2020). Opini Publik Terhadap Penerapa New Normal di Media Sosial Twitter. *Journal of Strategic Communication*, *11*(1), 19–26.

Amalia, H. (2021). Omicron penyebab COVID-19 sebagai variant of concern. *Jurnal Biomedika Dan Kesehatan*, *4*(4), 139–141.

Askaria, A. O. (2019). Pengaruh Promosi Melalui Media Social Twitter Pada Online Shop Shopee @Shopeeid Terhadap Keptusan Pembelian (Survei Terhadap Followers Akun Twitter @Shopeeid). *Jurnal Studi Manajemen Dan Bisnis*, *6*(2), 71–77.

Ashraf, M. R., Jana, Y., Umer, Q., Jaffar, M. A., Chung, S., & Ramay, W. Y. (2023). BERT-Based Sentiment Analysis for Low-Resourced Languages: A Case Study of Urdu Language. *IEEE Access*, *11*(1), 110245–110259.

Balaputra, I. (2022). Mewujudkan Masyarakat Sehat dan Produktif dengan Vaksinasi Covid-19 Dosis Lanjutan (Booster). *Jurnal Pengabdian Masyarakat Al-Qodiri (JPMA)*, *1*(1), 9–14.

Buana, I. K. S. (2018). Aplikasi untuk Pengoperasian Komputer dengan Mendeteksi Gerakan Menggunakan OpenCV Python. *Prosiding SINTAK*, *2*, 189–194.

Budiarto, J. (2021). Identifikasi Kebutuhan Masyarakat Nusa Tenggara Barat pada Pandemi Covid-19 di Media Sosial dengan Metode Crawling. *Jurnal Teknologi Informasi Dan Multimedia*, *2*(4), 244–250.

Buntoro, G. A. (2017). Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. *Integer Journal*, *2*(1), 32–41.

Darwis, D., Pratiwi, E. S., & Pasaribu, A. F. O. (2020). Penerapan Algoritma SVM untuk Analisis Sentimen Pada Twitter KPK RI. *Jurnal Ilmiah Edutic*, *7*(1), 1–11.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805v2*, *2*, 1–16.

Dharmawan, S., Mawardi, V. C., & Perdana, N. J. (2023). Klasifikasi Ujaran Kebencian Menggunakan Metode FeedForward Neural Network (IndoBERT). *Jurnal Ilmu Komputer Dan Sistem Informasi*, *11*(1), 1–6.

Faisal, D. R., & Mahendra, R. (2022). Two-Stage Classifier for COVID-19 Misinformation Detection Using BERT: a Study on Indonesian Tweets. *ArXiv:2206.15359v1*, *1*, 1–29.

Guswandri, A., & Cahyono, R. P. (2023). Penerapan Sentimen Analis Menggunakan Metode Naive Bayes dan SVM. *Jurnal Ilmu Data*, *2*(12), 1–16.

Irmanda, H. N., & Astriratma, R. (2020). Klasifikasi Jenis Pantun dengan Metode Support Vector Machine (SVM). *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *4*(5), 915–922.

Kulsumarwati, A., Purnamasari, I., & Dermawan, B. A. (2021). Penerapan SVM dan Information Gain Pada Analisis Sentimen Pelaksanaan Pilkada saat Pandemi. *Jurnal Teknlogi Informatika Dan Komputer MH. Thamrin*, *7*(2), 101–109.

Logan, T. (2019). A Practical, Iterative Framework for Secondary Data Analysis in Educational Research. *The Australian Educational Researcher*, *10*, 1–20.

Melinda, R. N., Ningrum, L. M., Suryabrata, I. B., Dwipa, G. S. B. A., & Sukoco, T. P. (2021). Program Perhitungan RAB Pekerjaan Struktur Baja (WF BEAM) Menggunakan Bahasa Python. *TIERS Information Technology Journal*, *2*(1), 31–38.

Narulita, L. F. (2019). Analisa Sentimen Pada Tinjauan Buku dengan Algoritma K-Nearest Neighbour. *KONVERGENSI*, *13*(2), 76–81.

Naseem, U., Razzak, I., Khushi, M., Eklund, P. W., & Kim, J. (2021). COVIDSenti A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis. *IEEE Transactions on Computational Social Systems*, *8*(4), 976–988.

Normawati, D., & Prayogi, S. A. (2021). Implementasi Naïve Bayes Classifier dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter. *Jurnal Sains Komputer & Informatika*, *5*(2), 697–711.

Openg, J. B. J. R., H, M. E., & Hamzah. (2022). Klasifikasi Unggas Ordo Anseriformes Berdasarkan Citra Menggunakan Metode Deep Learning Dengan Algoritma Convolutional Neural Network (CNN). *Seminar Nasional Teknik Elektro, Informatika Dan Sistem Informasi (SINTaKS)*, *1*(1), 1–6.

Pratama, F. A., & Romadhony, A. (2020). Identifikasi Komentar Toksik dengan BERT. *E-Proceeding of Engineering*, *7*(2), 1–18.

Pratiwi, B. P., Handayani, A. S., & Sarjana. (2020). Pengukuran Kinerja Sistem Kualitas Udara dengan Teknologi WSN Menggunakan Confusion Matrix. *JURNAL INFORMATIKA UPGRIS*, *6*(2), 66–75.

Putri, C. A., Adiwijaya, & al Faraby, S. (2020). Analisis Sentimen Review Film Berbahasa Inggris Dengan Pendekatan Bidirectional Encoder Representations

from Transformers. *Jurnal Teknik Informatika Dan Sistem Informasi*, *6*(2), 181–193.

Rachman, F. F., & Pramana, S. (2020). Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter. *Indonesian of Health Information Management Journal*, *8*(2), 100–109.

Razaq, E. R. M., Jacob, D. W., & Hamami, F. (2021). Analisis Sentimen Kepuasan Mahasiswa Terhadap Pembelajaran Online Selama Pandemi Covid-19 Pada Media Sosial Twitter Menggunakan Perbandingan Algoritma Klasifikasi. *E-Proceedings of Engineering*, *8*(5), 1–7.

Rohman, A. N., Utami, E., & Raharjo. Suwanto. (2019). Deteksi Kondisi Emosi pada Media Sosial Menggunakan Pendekatan Leksikon dan Natural Language Processing. *Jurnal Eksplora Informatika*, *9*(1), 70–76.

Rosalina, R., Auzar, & Hermandra. (2020). Penggunaan Bahasa Slang di Media Sosial Twitter. *Jurnal Tuah Pendidikan Dan Pengajaran Bahasa*, *2*(1), 77–84.

Sari, A. C., Hartina, R., Awalia, R., Irianti, H., & Ainun, N. (2018). Komunikasi dan Media Sosial. In *https://www.researchgate.net/profile/Astari-Clara-Sari/publication/329998890_KOMUNIKASI_DAN_MEDIA_SOSIAL/links/5c2f3d83299bf12be3ab90d2/KOMUNIKASI-DAN-MEDIA-SOSIAL.pdf* (pp. 1–9).

Sari, I. P., & Sriwidodo. (2020). Perkembangan Teknologi Terkini dalam Mempercepat Produksi Vaksin COVID-19. *Majalah Farmasetika*, *5*(5), 204–217.

Utama, H. S., Rosiyadi, D., Prakoso, B. S., & Ariadarma, D. (2019). Analisis Sentimen Ganjil Genap di Tol Bekasi Menggunakan Algoritma Support Vector Machine. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *3*(2), 243–250.

Wang, T., Lu, K., Chow, K. P., & Zhu, K. (2020). COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model. *IEEE Access*, *8*, 132162–138169.

Wijayanti, R., Khodra, M. L., & Widyantoro, D. H. (2021). Indonesian Abstractive Summarization using Pre-trained Model. *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT). IEEE*, *1*(1), 79–84.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. *ArXiv:2009.05387v3*, *1*, 1–15.

Yousefinaghani, S., Dara, R., Mubareka, S., Papadopoulos, A., & Sharif, S. (2021). An Analysis of COVID-19 Vaccine Sentiments and Opinions on Twitter. *International Journal of Infectious Diseases*, *108*, 256–262.

Yulita, W., Nugroho, E. D., & Algifari, M. H. (2021). Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma Naive Bayes Classifier. *Jurnal Data Mining Dan Sistem Informasi*, *2*(2), 1–9.

Yunefri, Y., Fadrial, Y. E., & Sutejo. (2021). Chatbot Pada Smart Cooperative Oriented Problem Menggunakan Natural Language Processing dan Naive Bayes Classifier. *Journal of Information Technology and Computer Science (INTECOMS)*, *4*(2), 131–141.

LAMPIRAN

Lampiran 1. Hasil *Preprocessing*

no	text	label
1	program vaksin covid booster jalan indonesia aman hamil booster vaksin covid jelas	1
2	vaksin booster efek anjir mboh	-1
3	tiket vaksin booster ya prancis paris is such cantik city	1
4	tiket muncul vaksin booster	1
5	bentar vaksin booster	1
6	sukses program perintah pegawai lembaga masyarakat narkotika nusakambangan ikut vaksin booster	1
7	kayak gue begadang besok vaksin booster	0
8	info arek ngagel bratang pucang nek atene vaksin booster pt	1
9	twit mas vaksin booster moderna faskes ya	1
10	oh my god asli excited banget skshskdh maju cerita bangsat gue gatakut gue vaksin booster	0
11	vaksin booster anjir	0
12	maksimal jarak vaksin booster habis lebaran jadwal	1
13	tonton konser umur vaksin booster bingung nih	1
14	kipi hoaks ambil booster vaksin aman efektif cegah mati	1
15	ya tahun vaksin booster ya	1
16	cek tas pvc ubek buka printilan vaksin periksa ya booster biar aman	1
17	min kereta tuju jakarta st lamongan vaksin booster ya	1
18	butuh sertif booster dada perlu jalan umrah tonton konser pakai jasa joki vaksin ya vaksin langsung full dosis tunggu order via direct message orang	1
19	kakak jarak vaksin booster jarak ya vaksin	1
20	ready tembak vaksin vaksin hapus data benar data sertif salah negatif positif proses muncul sertif bayar	-1
21	masuk akal amp waras negara hormat hak warga hak laku hak dibooster jadi dagang vaksin syarat apa syarat jalan fungsi vaksin booster	1
22	upaya benteng kuat imunitas pandemi covid klinik pratama menteri ekonomi himpun mahasiswa katolik dinas sehat daerah khusus ibukota jakarta gelar vaksinasi booster dosis pegawai jenis vaksin booster pfizer	1
23	karuan blak an pandemi jual vaxin enak pakai paksa tinggal sabet pakai golok coba jual plus pakai mandat biar chaos biar mbersihin jabat usaha model serta jongos nya	-1
24	yjh bismillah last day banget ketemu ji vaksin booster beli outfit ji tonton uang kumpul moga menang ketemu jiii depann	1
25	sih bantu badan selenggara jamin sosial acara vaksin gratis booster sifat wajib perlu ribut ky sih	0
26	vaksin booster kali sakit banget astaga badan pnas dinginn	-1

no	text	label
27	humas polisi sektor johar polisi resor metro jakarta pusat bhayangkara bina aman tertib masyarakat lurah johar aiptu sunarto forum waspada masyarakat keluarga johar pantau giat vaksinasi covid vaksin booster puskesmas camat johar lobi kantor lurah johar	1
28	vaksin vaksin vaksin booster vaksin suasana hati booster kali tangan gue suntik mon maaf nih anjur urgensi pakai banget iya	1
29	gais vaksin booster graha masuk coi sertifnya peduli lindung vaksin booster aman	1
30	gerai vaksin booster polresta bandara soetta layan calon tumpang	1
31	tangan pegellllll aaaaaaaa efek vaksin booster	-1
32	selamat malam kak rencana jalan kakak rajabasa wajib vaksin dosis tiga booster tumpang usia informasi lengkap syarat tentu kakak cek taut terima kasih	1
33	booster vaksin kayak sakit garuk ketek susah	-1
34	yjh terima kasih kak ada insyaallah labil banget tonton treasure kira uang pp tinggal jakarta full vaksin and booster	1
35	nder sehat vaksin booster vaksin gih chip lengan baru versi	1
36	habis vaksin booster tangan langsung sakit auto diminumin paracetamol biar tambah sakit	0
37	wajib vaksin dosis tiga booster kait bijak tentu sila koordinasi kantor sehat labuh kkp maskapai lokasi bandara berangkat terima kasih kau	1
38	warga ambon vaksin booster dosis	1
39	vaksin booster gkada efek samping booster astra teman ku meriang	-1
40	vaksin booster efek sih lapar	0
41	vaksin booster efek samping kakak strong banget ya	1
42	efek samping kak pas vaksin pas booster merek boosternya keras	-1
43	malam admin mohon info baru syarat terbang domestik kait vaksin tes covid terbang cgk plm butuh surat tes negatif vaksin booster ya	0
44	urus booster vaksin	1
45	masyarakat domisili jakarta selatan sila vaksin booster upk menteri sehat isi link registrasi ya jadwal vaksin isi link	1
46	vaksin booster bayar mlsno weh	1
47	dok booster vaksin vaksin kah terima kasih	1
48	hi cari vaksin amp booster amp domosil sumatra utara wa ya pulau sumatra utara jamin percaya aman	1
49	hemat sehat mari vaksin booster ueue	1
50	vaksin booster booster jab taun boost ulang wajib vaksin	0
51	vaksin booster alhamdulillah bebas demam no keluh apa pas malemnya pegeeel tangan habis suntik sekian review booster juned tim pfizer	-1
52	pagi habis vaksin booster keras tangan sakit banget ya gera	-1
53	pasca booster obat kutu pakai minggu telah ya kucing garuk pakai obat kutu vaksin	0
54	sertifikat vaksin masalah jalan via pesawat booster booster sertifikat chatt contact center apes	0

no	text	label
55	selamat malam mohon maaf tumpang kakak kota usia wajib vaksin dosis tiga booster	1
56	butuh sertifikat vaksin booster booster saya dos data peduli lindung	1
57	wajib vaksin booster kakak	1
58	tau gue vaksin booster gue percaya beli	1
59	kampung leone is back satriaus iya iya chan jemie ulama mayoritas ri mayoritas nus sebut ulama maksud	0
60	guyss tolong infoo vaksin booster mesti ulang dosis	0
61	sahabat cari vaksin booster dompet dhuafa buka sempat sahabat proteksi tubuh vaksin booster	1
62	by the way booster pakai sinopharm vaksin booster pakai pfizer moderna	1
63	untuk vaksin booster geser ya langkah ikhtiar jaga sakit covid syarat administrasi jalan apa apa	0
64	iqram sian uji nasional akun kena beku iya booster ambil kj pesan pi cucuk otak vaksin beli expired	-1
65	malam admin besok surabaya jakarta vaksin tiga booster solusi min	1
66	target vaksin booster ya biar laku rencana perintah jual vaksin covid dajjal nikah pangeran bungsu natal covid aman jelang puasa lebaran covid muncul bangsat	-1
67	jbrfess ngereta jember lmj vaksin booster ya probowangi terima kasih	1
68	booster vaksin suntik	1
69	wajib vaksin dosis wajib booster kakak	1
70	informasi second booster vaksin daerah	1
71	jam vaksin booster asa sumeng pusing	-1
72	ready adopt kak jantan warna kuning langsat ukur tiny aktif amp lincah vaksin booster bebas kutu non stambum	0
73	nih negara duit alias bangkrut obrol tumbuh ekonomi roket jabat otak bisnis amp nurani main	-1
74	usia masyarakat booster puskesmas rumah sakit yogyakarta laksana vaksin booster jabat lingkung perintah kota yogyakarta terang emma	1
75	min vaksin booster sih kayak	1
76	wajib vaksin booster ih fasilitas	1
77	habis vaksin booster nyut nyut tan	-1
78	barusan banget antre obat kemarin habis vaksin booster eh sakit	-1
79	alhamdulillah vaksin booster moga sehat	1
80	kakak buku nikah kartu vaksin kalah booster bolh beli	0
81	info vaksin booster mana terima kasih	1
82	januari menteri sehat keluar bijak baru booster masyarakat	1
83	vaksin booster siang tanda meriang surat pemberitahuan tahun info senior	0
84	gue vaksin booster nih	1
85	perintah tetap vaksin booster bayar rupiah	1
86	adek fitrie kirana ajak masyarakat vaksin booster ih covid lindung	1

no	text	label
87	guys mumpung dyandra syarat vaksin booster vaksin booster ya booster	1
88	teman teman lupa vaksin booster ya tiket booster cek dinkes daerah ya	1
89	solusi peduli lindung warna kuning tiket vaksin booster mohon bantu	1
90	yjh orang tua ngizinin labil banget tonton treasure uang full vaksin booster akomodasi aman pp moga rezeki ketemu junghwan kakak ya terima kasih kak ada	1
91	permisi min ulang jumat besok pergi bandung kereta surabaya syarat kereta usia vaksin booster tiket vaksin booster peduli lindung vaksin booster	1
92	vaksin booster bayar rupiah menteri sehat angka make sense	1
93	masyarakat usia terima sunti vaksin covid booster syarat lewat sunti booster yuk cek aplikasi peduli lindung	1
94	ayo vaksinasi booster vaksin booster fungsi tingkat imun bentuk dosis vaksin dosis vaksin terima anggap dosis utama vaksin	1
95	mjb kak usia wajib vaksin dosis booster barangkali minat direct message kak	1
96	vaksinasi covid dosis booster masyarakat usia januari tunda vaksin vaksinasi baik vaksinasi	1
97	dinkes biak numfor tunggu datang vaksin covid booster	1
98	takut vaksin booster timbul overdosis takut booster over dosis hitung jamin emergency use badan awas obat makan jamin aman mutu khasiat	1
99	individu vaksin daerah negara herd immunity capai manfaat vaksinasi lengkap tambah booster mudah akses layan fasilitas publik	1
100	vaksinasi booster fungsi tingkat imun bentuk dosis vaksin tiga vaksin terima anggap dosis utama vaksin ayo vaksinasi booster	1

Lampiran 2. *Topic modeling* positif

index	topic	representation	representative_docs
0	-1	booster, vaksin, pfizer, ya, dosis, info, syarat, vaksinasi, covid, perintah	jasa joki vaksin only dosis available dosis booster aman tembus peduli lindung bayar akhir sertifikat muncul akun peduli lindung order direct message orang,selamat malam vaksin primer az dasar rekomendasi kemenkes booster astrazeneca pfizer moderna tanggal mei sedia booster primer az sinovac pfizer jenis booster pfizer ya,vaksin sedia mei sinovac dosis usia dosis pfizer dosis usia dosis dosis booster primer sinovac az pfizer moderna dosis dosis booster primer moderna nakes covovax dosis drop out
1	0	sih, ya, booster, kali, iya, vaksin, gue, nya, teman, biar	vaksin booster sih,vaksin booster sih,oh booster ya teman vaksin suruh ulang vaksin nya

index	topic	representation	representative_docs
2	1	selamat, kakak, tumpang, tiga, wajib, tes, dosis, rapid, malam, pacar	selamat malam kak kakak antarkota vaksin dosis tiga booster wajib rapid tes antigen rt pacar ya informasi lengkap syarat jalan kakak sila cek link terima kasih,selamat malam kak calon tumpang kakak jarak terima vaksin tiga booster wajib hasil negatif rapid tes antigen rt pacar ya info lengkap sila cek link terima kasih,selamat malam kak tumpang kakak jarak vaksin dosis tiga booster wajib hasil negatif tes rt pacar antigen
3	2	gue, maju, vaksin, booster, yjh, ya, akomodasi, puskesmas, the, moga	maju gue vaksin booster,maju gue vaksin booster,maju gue vaksin booster
4	3	indonesia, jakarta, negara, covid, warga, republik, masyarakat, sehat, thailand, booster	ikut vaksin dosis lengkap vaksin booster indonesia polisi republik indonesia restuban,warga indonesia dosis tiga booster vaksin covid ahad indonesia barat,sobat bina masyarakat januari vaksinasi covid dosis booster masyarakat usia kombinasi vaksin booster polisi republik indonesia polisi republik indonesia
5	4	besok, min, minggu, jadwal, pagi, stasiun, bandung, booster, jam, info	besok vaksin booster,besok booster vaksin,besok vaksin booster
6	5	lengkap, booster, vaksin, prokes, alias, empat, deh, nalar, upid, tahub	vaksin booster prokes lengkap,vaksin booster prokes lengkap,vaksin booster prokes lengkap
7	6	sertifikat, data, dos, hse, pedulilindungi, butuh, beda, muncul, booster, sgac	butuh sertifikat vaksin booster booster dos,booster vaksin pas vaksin iseng cek sertifikat muncul sertifikat nya booster,sertifikat vaksin lengkap booster
8	7	sehat, hak, syarat, tubuh, otomatis, vaksin, event, hidup, mudah, ayo	masuk akal amp waras negara hormat hak warga hak laku hak dibooster jadi dagang vaksin syarat apa syarat jalan fungsi vaksin booster,biasa hidup sehat taat protokol sehat lupa ikut vaksin booster tambah daya imunitas tubuh imun kuat kunci hidup sehat karsa byakta karya lang bangga,yuk tingkat imun tubuh vaksin booster lupa jaga sehat ya jaga

index	topic	representation	representative_docs
			kondisi tubuh ya vaksin jaga diri risiko sehat fwd hospital care proteksi
9	8	tiket, kartu, pesan, tanda, duduk, muncul, beli, tunjuk, sa, surat	atuh baru tiket vaksin booster tiket baru kemarin bareng antre booster buanyaaak banget tiket vaksin nya muncul peduli lindung pesan tugas cek ya bu,tiket muncul vaksin booster,pesan tiket wajib pakai kartu tanda duduk sistem sinkron peduli lindung pesan tiket masuk no nik kartu tanda duduk vaksin kali booster stadion geger gedhen tiket kebel masuk

Lampiran 3. *Topic modeling* negatif

index	topic	representation	representative_docs
0	-1	vaksin, booster, sakit, covid, banget, pegal, habis, ya, efek, badan	gue sakit habis vaksin booster,efek vaksin booster tidur badan pegal banget,sakit banget habis vaksin booster
1	0	habis, banget, vaksin, booster, ya, sih, allah, gue, sakit, maaf	rest cape banget kemarin tidur makan atur drop habis vaksin booster wkwk rusuh ya pas gue tinggal videnist ngereog,halo kak bismillah ya allah moga banget skincare skintific kbtulan moistnya serum acne nya sih susah banget rep bismillah hadiah habis vaksin booster suasana hati terima kasih kak,ya allah habis vaksin booster ngilu banget
2	1	efek, booster, vaksin, samping, nya, ya, badan, bikin, az, pas	kira booster efek samping kayak vaksin,gue manusia ya vaksin asa efek apa booster moderna parah tusuk suntik nya asa nya rasa efek apa,booster vaksin nya moderna booster vaksin pfizer berat efek samping dosis
3	2	tangan, lengan, sakit, booster, vaksin, nyeri, kiri, bekas, pegal, banget	tangan bekas vaksin booster pegal banget,tangan kiri pegal banget habis booster vaksin,tangan sakit habis vaksin booster
4	3	booster, habis, vaksin, gue, pfizer, awas, varian, anjing, kemarin, diam	habis vaksin booster meriang,habis vaksin booster rsud kasih pfizer kipi tangan kiri kebas kemarin vaksin,fasilitas vaksin sinopharm dosis amp kantor pas booster dapetnya pfizer sertifikat boosternya ribet pakai jasa calo tembak mahal banget bye
5	4	takut, vaksin, booster, indonesia, nya,	takut vaksin booster,takut vaksin booster,takut vaksin booster

index	topic	representation	representative_docs
		ya, tuntut, jarum, sertif, bondong	
6	5	demam, vaksin, booster, habis, pusing, ngilu, teman, minggu, kayak, pegal	vaksin booster bikin demam badan pegal pusing,habis vaksin booster lemas pusing demam,habis vaksin booster langsung demam
7	6	drugs, gue, kena, iya, booster, alias, vaksin, sakit, kipi, pegal	gue ulas efek samping drugs ya liver hancur vaksin drugs drugs meriang akibat vaksin booster drugs booster drugs kena mati kemarin antek receh kasebul dieharder benci ulama mati kena sirosis efek vaksin,kampung is back leone iya iya satriaus chan jemie kipi diam dok tugas pusekesmas satu tugas coped lumpuh orang habis vak real anak muda kena strok vaksin tiga alias booster,kampung leone iya iya satriaus chan is back jemie amat tahan tubuh tahan sakit kena efek vaksin booster pura benar vaksin dibooster alias bohong tubuh aroma necrosis
8	7	sakit, minggu, banget, covid, mandul, marah, vaksin, asa, positif, booster	jumat minggu booster vaksin pfizer alhamdulillah nya keluh kayak vaksin malemnya lengan suntik pegal banget asa subuh,goees januari positif covid kmd sep oktober vaksin amp kipi rasa minggu suruh booster trauma ktk vaksin amp kipi lebih kena covid paksa vaksin,mbak februari positif covid marah orang percaya covid vaksin pakai tes dicovidkan pas vaksin booster marah bikin gumpal darah mandul mandul nular orang

Lampiran 4. *Source code* dan *topic modeling* sentimen negatif

```
df['label'] = df['label'].replace(-1, 2)

docs = list(df.loc[:, "text"].values)
```

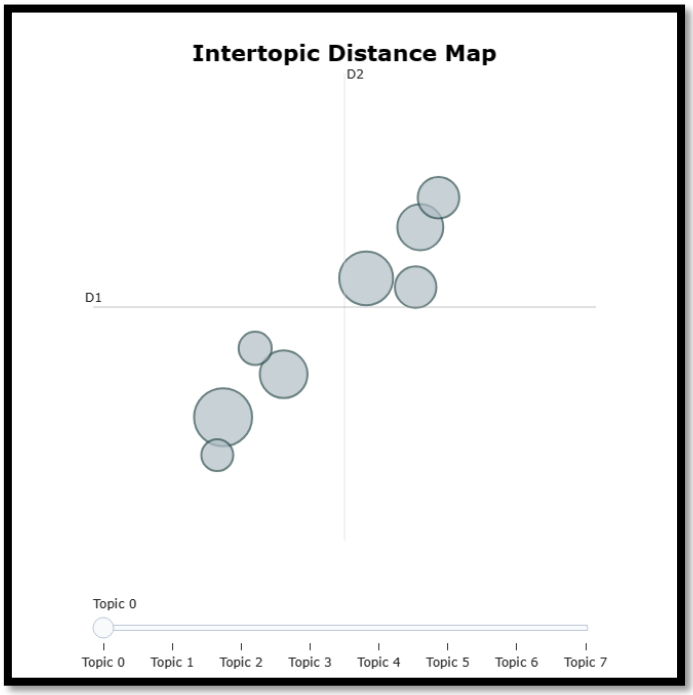
Untuk mengganti label sentimen negatif dari -1 menjadi 2 karena BERT tidak dapat membaca angka kurang dari 0 (-).

```
model.visualize_topics()
```

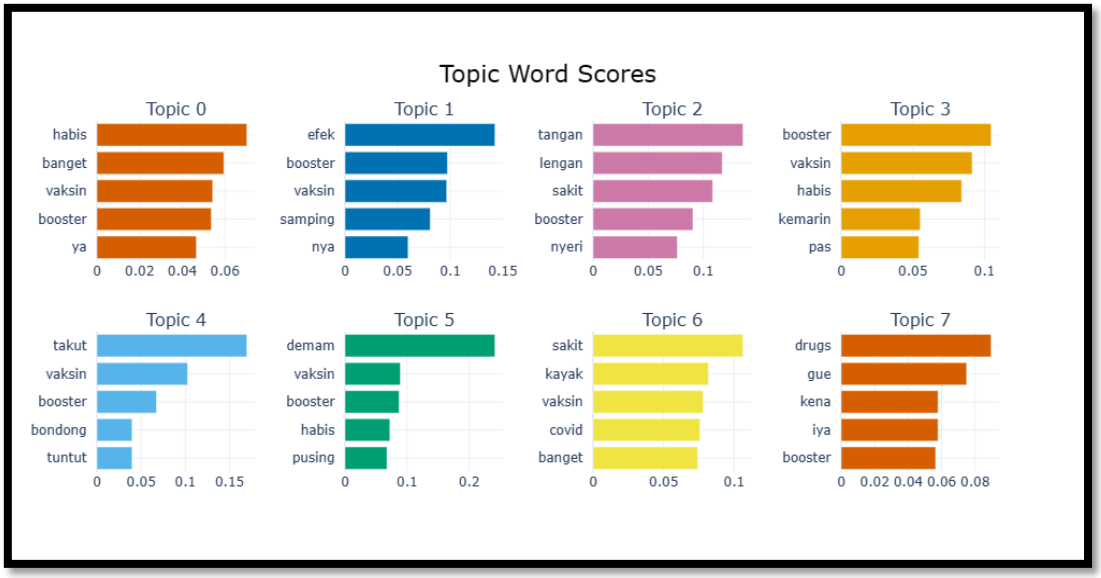
Visualisasi data *tweet* dengan *mapping* dari topik yang sudah dilakukan pemodelan oleh BERT.

```
model.get_representative_docs(0)
model.visualize_barchart()
```

Visualisasi data *tweet* teratas dengan *barchart* dari topik yang sudah dilakukan pemodelan oleh BERT.



Map penyebaran topik yang dihasilkan dari pemodelan BERT pada sentimen positif sebanyak delapan topik. Kesamaan antar topik sangat mungkin terjadi karena pemodelan BERT merupakan *unsupervised learning*.



Penyebaran delapan topik yang divisualisasikan dengan *barchart* sesuai dengan hasil pemodelan yang telah dilakukan oleh BERT.

Lampiran 5. *Source code* klasifikasi BERT dan akurasi pada epoch

```
from transformers import BertTokenizer

print("Loading BERT Tokenizer")
tokenizer = BertTokenizer.from_pretrained('bert-base-multilingual-uncased', do_lower_case=True)
```

Memanggil model BERT *tokenizer* untuk melatih data *tweet*. BERT sudah mengenal banyak sekali bahasa termasuk bahasa Indonesia.

```
from tensorflow.keras.preprocessing.sequence import pad_sequences

MAX_LEN = 68

print("Padding/truncating all sentences to %d values" % MAX_LEN)
print('Padding token: "{:}", ID: {:}'.format(tokenizer.pad_token, tokenizer.pad_token_id))

input_ids = pad_sequences(input_ids, maxlen=MAX_LEN, dtype='long', value=0, truncating='post', padding='post')

print("Done")
```

Mengubah seluruh kalimat yang sudah mendapatkan token menjadi sebuah vektor yang kemudian akan memasuki tahap klasifikasi dengan pembagian seperti pada gambar.

```
from torch.utils.data import TensorDataset, DataLoader, RandomSampler, SequentialSampler

batch_size = 32

train_data = TensorDataset(train_input, train_mask, train_labels)
train_sampler = RandomSampler(train_data)
train_dataloader = DataLoader(train_data, sampler=train_sampler, batch_size=batch_size)

validation_data = TensorDataset(validation_input, validation_mask, validation_labels)
validation_sampler = SequentialSampler(validation_data)
validation_dataloader = DataLoader(validation_data, sampler=validation_sampler, batch_size=batch_size)

test_data = TensorDataset(test_input, test_mask, test_labels)
test_sampler = SequentialSampler(test_data)
test_dataloader = DataLoader(test_data, sampler=test_sampler, batch_size=batch_size)
```

Melatih, memvalidasi dan menguji data *tweet* oleh BERT.

```
======= Epoch 1 / 10 =======
Training...
   Average training loss: 0.75
   Training epoch took: 0:00:16
Running Validation...
   Accuracy: 0.76
   Validation took: 0:00:01
======= Epoch 2 / 10 =======
Training...
   Average training loss: 0.58
   Training epoch took: 0:00:14
Running Validation...
   Accuracy: 0.80
   Validation took: 0:00:01
======= Epoch 3 / 10 =======
Training...
   Average training loss: 0.47
   Training epoch took: 0:00:14
Running Validation...
   Accuracy: 0.80
   Validation took: 0:00:01
======= Epoch 4 / 10 =======
Training...
   Average training loss: 0.39
   Training epoch took: 0:00:14
Running Validation...
   Accuracy: 0.80
   Validation took: 0:00:01
```

```
======= Epoch 5 / 10 =======
Training...
   Average training loss: 0.34
   Training epoch took: 0:00:15
Running Validation...
   Accuracy: 0.73
   Validation took: 0:00:01
======= Epoch 6 / 10 =======
Training...
   Average training loss: 0.30
   Training epoch took: 0:00:15
Running Validation...
   Accuracy: 0.79
   Validation took: 0:00:01
======= Epoch 7 / 10 =======
Training...
   Average training loss: 0.24
   Training epoch took: 0:00:14
Running Validation...
   Accuracy: 0.80
   Validation took: 0:00:01
======= Epoch 8 / 10 =======
Training...
   Average training loss: 0.18
   Training epoch took: 0:00:14
Running Validation...
   Accuracy: 0.79
   Validation took: 0:00:01
```

```
======= Epoch 9 / 10 =======
Training...
   Average training loss: 0.15
   Training epoch took: 0:00:14
Running Validation...
   Accuracy: 0.79
   Validation took: 0:00:01
======= Epoch 10 / 10 =======
Training...
   Average training loss: 0.13
   Training epoch took: 0:00:14
Running Validation...
   Accuracy: 0.78
   Validation took: 0:00:01
Training complete!
```

Data loss pada tiap *epoch* yang semakin mengecil hingga *epoch* ke-10 membuat akurasi semakin meningkat dan menghasilkan akurasi yang baik.

Lampiran 6. Perbandingan akurasi pada data latih dan data uji

Data latih	Data uji	Akurasi
60	40	79%
70	30	81%
80	**20**	**85%**
90	10	81%