ANALISIS SENTIMEN PUBLIK DI TWITTER (X) TERHADAP PASANGAN CALON GUBERNUR DAN WAKIL GUBERNUR JAWA TENGAH 2024 MENGGUNAKAN INDOBERT

SKRIPSI

Diajukan untuk Memenuhi Tugas Akir dan Melengkapi Syarat
Guna Memperoleh Gelar Sarjana Strata Satu (S-1)
dalam Ilmu Teknologi Informasi

Diajukan Oleh:

Bagus Diaz Pratama

NIM.2108096059

PROGRAM STUDI TEKNOLOGI INFORMASI
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI WALISONGO
SEMARANG

TAHUN 2025

PERNYATAAN KEASLIAN

Yang bertandatangan dibawah ini:

Nama : Bagus Diaz Pratama

NIM : 2108096059

Jurusan : Teknologi Informasi

Menyatakan bahwa skripsi yang berjudul :

ANALISIS SENTIMEN PUBLIK DI TWITTER (X) TERHADAP PASANGAN CALON GUBERNUR DAN WAKIL GUBERNUR JAWA TENGAH 2024 MENGGUNAKAN INDOBERT

Secara keseluruhan adalah hasil penelitian/karya saya sendiri, kecuali bagian tertentu yang dirujuk dari sumbernya.

Semarang, 29 April 2025

Pembuat Pernyataan,

Bagus Diaz Pratama

NIM: 2108096059

KEMENTRIAN AGAMA
UNIVERSITAS ISLAM NEGERI WALISONGO
FAKULTAS SAINS DAN TEKNOLOGI
Jl. Prof. Dr. Hamka Ngaliyan Semarang
Telp.024-7601259 Fax. 7615387

LEMBAR PENGESAHAN

Naskah skripsi berikut ini:

Judul : Analisis Sentimen Publik di Twitter (X) Terhadap Pasangan Calon Gubernur dan Wakil Gubernur Jawa Tengah 2024 Menggunakan IndoBERT

Penulis : Bagus Diaz Pratama
NIM : 2108096059
Jurusan : Teknologi Informasi

Telah diujikan dalam sidang *tugas akhir* oleh Dewan Penguji Fakultas Sains dan Teknologi UIN Walisongo dan dapat diterima sebagai salah satu syarat memperoleh gelar sarjana dalam Ilmu Teknologi Informasi.

Semarang, 2 Juni 2025

DEWAN PENGUJI

Ketua Sidang	Sekretaris Sidang
Hery Mustofa, M.Kom.	Adzhal Arwani Mahfudh, M.kom.
NIP. 198703172019031007	NIP. 199107032019031006
Penguji Utama I	Penguji Utama II
Dr. Wenty Dwi Yuniarti, M.Kom.	Nur Cahyo Hendro Wibowo, M.Kom
NIP. 197706222006042005	NIP. 197312222006041001
Pembimbing I	Pembimbing II
Siti Nur'aini, M.Kom	Adzhal Arwani Mahfudh, M.kom.
NIP. 198401312018012001	NIP. 199107032019031006

NOTA PEMBIMBING I

Semarang, 29 April 2025

Yth. Ketua Program Studi Teknologi Informasi
Fakultas Sains dan Teknologi UIN
Walisongo Semarang
Assalamu'alaikum. Wr. Wb.
Dengan ini diberitahukan bahwa saya telah melakukan bimbingan, arahan dan koreksi naskah skripsi dengan :

Judul	Analisis Sentimen Publik di Twitter (X) Terhadap Pasangan Calon Gubernur dan Wakil Gubernur Jawa Tengah 2024 Menggunakan IndoBERT
Nama	**Bagus Diaz Pratama**
NIM	2108096059
Jurusan	Teknologi Informasi

Saya memandang bahwa naskah skripsi tersebut sudah dapat diajukan kepada Fakultas Sains dan Teknologi UIN Walisongo untuk diujikan dalam Sidang Munaqosah.
Wassalamu'alaikum. Wr. Wb.

Pembimbing I,

Siti Nur'aini, M.Kom
NIP. 1984013120180120001

NOTA PEMBIMBING II

Semarang, 29 April 2025

Yth. Ketua Program Studi Teknologi Informasi
Fakultas Sains dan Teknologi UIN
Walisongo Semarang
Assalamu'alaikum. Wr. Wb.
Dengan ini diberitahukan bahwa saya telah melakukan bimbingan, arahan dan koreksi naskah skripsi dengan :

Judul	:	Analisis Sentimen Publik di Twitter (X) Terhadap Pasangan Calon Gubernur dan Wakil Gubernur Jawa Tengah 2024 Menggunakan IndoBERT
Nama	:	**Bagus Diaz Pratama**
NIM	:	2108096059
Jurusan	:	Teknologi Informasi

Saya memandang bahwa naskah skripsi tersebut sudah dapat diajukan kepada Fakultas Sains dan Teknologi UIN Walisongo untuk diujikan dalam Sidang Munaqosah.
Wassalamu'alaikum. Wr. Wb.

Pembimbing II,

Adzhal Arwani Mahfudh,
S.Kom., M.Kom
NIP. 198401312018012001

LEMBAR PERSEMBAHAN

Dengan rasa syukur yang mendalam, dengan telah diselesaikannya skripsi ini, penulis mempersembahkan kepada:

1. Orang tuaku tersayang. Terima kasih kepada Bapak Imam Kusnadi dan Ibu Tri Setyaningrum atas kasih sayang dan cinta yang tak tergantikan, yang tak henti hentinya mendoakan, mendukung, dan membimbing saya sepanjang hidup
2. Santika Anggun Ningtya dan Devi Putri Imanda, saudara-saudara kandung saya yang tidak pernah berhenti mendukung saya dalam kehidupan yang indah ini.
3. Ananda Ummul Ulya yang selalu membersamai dan memberikan semangat penulis pada hari-hari yang tidak mudah selama proses pembuatan skripsi ini
4. Teman teman dekat saya Guntur Syarifuddin, Muhammad Zulfikar, Farid Rafif, Mirza Izzal, dan Varel Aldin yang telah memberikan semangat, dan mendukung.
5. Untuk diri saya sendiri yang telah berjuang sejauh ini, dengan melawan *Mood* yang tidak tentu selama penulisan skripsi ini.

MOTO

”Jika kamu tidak sanggup menahan lelahnya belajar maka kamu harus sanggup menahan perihnya kebodohan.”

Imam Syafi’I

ABSTRAK

Penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap calon gubernur dan wakil gubernur Jawa Tengah 2024 di media sosial Twitter (X) menggunakan model IndoBERT. Penelitian ini didasarkan pada tingginya penggunaan media sosial di Indonesia, khususnya Twitter (X) sebagai *platform* yang bebas mengekspresikan pendapat, serta beberapa pro kontra yang muncul terkait calon gubernur dan wakil gubernur Jawa Tengah 2024. Permasalahan yang dikaji meliputi kecenderungan sentimen masyarakat, penerapan model IndoBERT, dan efektivitas model dalam menganalisis sentimen masyarakat tentang calon gubernur dan wakil gubernur Jawa Tengah 2024. Pengumpulan data dilakukan melalui proses *crawling* menggunakan *Tweet Harvest* dengan kata kunci ”andika perkasa”, ”hendrar prihadi”, ”ahmad luthfi”, ”taj yasin” dari 24 Agustus 2024 dan 31 Oktober 2024, mendapatkan total 2.936 *tweet* yang mana pasangan calon 1 mendapat 1.492 *tweet* dan pasangan calon 2 mendapat 1.444 *tweet*. Data kemudian dilabeli oleh pakar bahasa ke dalam dua kategori yaitu positif dan negatif, dilanjutkan dengan tahap *pre-processing*, serta augmentasi data menggunakan *Easy Data Augmentation* (EDA) untuk menyeimbangkan distribusi kelas. Model IndoBERT diimplementasikan dengan *fine-tuning* menggunakan *learning rate* 3e-5, *batch size* 32, dan *epoch* 3. Hasil penelitian menunjukkan paslon 1 mendapatkan hasil *accuracy* 93%, *precision* 90%, *recall* 84%, *f1-score* 87% sedangkan untuk paslon 2 mendapatkan hasil *accuracy* 95%, *precision* 93%, *recall* 94%, dan *f1-score* 93%. Hasil analisis ini juga mengungkapakan dominasi sentimen positif terhadap kedua pasangan calon dan pasangan calon nomer urut 2 mendapatkan persentase positif yang lebih tinggi dan sentimen negatif yang cukup rendah dibandingkan pasangan calon nomer urut 1.

Kata Kunci: Calon Gubernur dan Wakil Gubernur Jawa Tengah, Analisis Sentimen, IndoBERT

KATA PENGANTAR

Alhamdulillah, segala puji dan syukur senantiasa penulis panjatkan ke hadirat Allah SWT atas limpahan nikmat dan karunia-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul ”Analisis Sentimen Publik di Twitter (X) Terhadap Pasangan Calon Gubernur Jawa Tengah 2024 Menggunakan IndoBERT” dengan baik. Skripsi ini disusun sebagai salah satu syarat kelulusan dalam program sarjana (S1) pada Program Studi Teknologi Informasi di Universitas Islam Negeri Walisongo, Kota Semarang. Selain itu, penelitian ini juga bertujuan untuk memberikan wawasan bagi pembaca mengenai analisis pendapat masyarakat terhadap Calon Gubernur dan Wakil Gubernur Jawa Tengah 2024 di Sosial Media.

Pada kesempatan ini, penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada semua pihak yang telah memberikan dukungan dan bantuan, baik dalam pelaksanaan penelitian maupun dalam penyusunan skripsi ini. Penulis menyadari bahwa tanpa bimbingan, arahan, serta motivasi dari berbagai pihak, proses penyelesaian skripsi ini tidak akan berjalan dengan lancar. Oleh karena itu, penulis mengucapkan penghargaan dan rasa terima kasih yang mendalam kepada:

1. Bapak Prof. Dr. Nizar, M.Ag selaku Rektor Universitas Islam Negeri Walisongo Semarang.

2. Bapak Prof. Dr. H. Musahadi, M.Ag selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Walisongo Semarang.
3. Bapak Dr. Khotibul Umam, M.Kom selaku Ketua Program studi Teknologi Informasi yang telah memberikan bimbingan sejak awal mengikuti perkuliahan.
4. Ibu Siti Nur'aini, M.kom, selaku dosen Pembimbing I sekaligus dosen wali yang telah memberi masukan dan arahan yang membuka pikiran penulis dalam penyusunan skripsi.
5. Bapak Adzhal Arwani Mahfudh, S.Kom., M.Kom selaku dosen Pembimbing II yang telah memberi banyak masukan serta arahan yang membantu dalam penyusunan skripsi.
6. Staff, karyawan dan dosen di lingkungan Universitas Islam Negeri Walisongo Semarang
7. Orang tua tercinta dan keluarga yang selalu menemani dalam membantu penulis dan selalu mendo'akan serta memberikan dukungan kepada penulis.
8. Teman-teman Teknologi Informasi khususnya kelas B aja yang selalu memberi dukungan.

9. Semua pihak yang tidak bisa penulis sebutkan satu persatu yang terlibat dalam pembuatan skripsi ini sehingga dapat terselesaikan dengan baik.

Dalam proses pelaksanaan dan penyusunan skripsi ini, penulis menyadari bahwa masih terdapat berbagai kekurangan dan ketidaksempurnaan. Oleh karena itu, penulis sangat mengharapkan kritik serta saran yang membangun guna meningkatkan kualitas penulisan ini. Semoga skripsi ini dapat memberikan manfaat bagi semua pihak yang membacanya.

Semarang, 29 April 2025

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

A. Latar Belakang

Indonesia merupakan negara yang memiliki jumlah penduduk terbesar keempat di dunia dan menganut sistem demokrasi. Salah satu penerapan demokrasi di Indonesia adalah adanya pemilihan kepala daerah (Pilkada) (Tata Lukmana et al., 2019). Pilkada merupakan pemilihan umum di Indonesia yang dilakukan secara langsung oleh masyarakat setempat yang memenuhi ketentuan peraturan perundang-undangan. Pemungutan suara ini terbuka untuk umum bebas dari intervensi dan dilaksanakan dengan kerahasiaan, integritas, dan keadilan yang dijaga (Satriawan et al., 2024).

Salah satu pemilihan kepala daerah (Pilkada) akan diselenggarakan di Provinsi Jawa Tengah. Sebagai salah satu provinsi dengan jumlah penduduk terbesar di Indonesia, hasil pemilihan di Jawa Tengah memiliki dampak yang signifikan terhadap politik di Indonesia (Retno, 2013). Menurut Badan Pusat Statistik (BPS) tahun 2023, Jawa Tengah memiliki populasi lebih dari 37 juta jiwa, menjadikannnya provinsi kelima terpadat di Indonesia pada tahun 2021.

Pilkada Jawa Tengah 2024 memiliki keunikan tersendiri. Selain menjadi salah satu provinsi dengan jumlah

pemilih terbesar, Jawa Tengah juga dikenal sebagai tolak ukur politik nasional. Isu-isu seperti pembangunan infrastruktur, pengurangan kemiskinan, dan peningkatan kualitas pendidikan menjadi fokus utama dalam kampanye para calon kandidat. Dinamika politik yang terjadi, termasuk koalisi partai dan latar belakang para calon kandidat, menambah kompleksitas dan daya tarik Pilkada saat ini (Annizar, 2024).

Hasil survei terbaru yang dilakukan oleh Poltracking pada September 2024 menunjukkan dinamika persaingan yang menarik antara para calon kandidat. Menurut survei Poltracking, pasangan Andika Perkasa-Hendrar Prihadi memperoleh dukungan sebesar 31,4%, sementara Ahmad Luthfi-Taj Yasin Maimoen memperoleh dukungan sebesar 52,5%. Sebanyak 16,4% responden masih belum menentukan pilihan. Data tersebut menunjukkan bahwa persepsi publik terhadap kandidat masih dinamis dan dapat berubah seiring berjalannya masa kampanye.

Pemilihan kepala daerah (Pilkada) akan diselenggarakan pada Rabu, 27 November 2024 akan menentukan pemilihan gubernur dan wakil gubernur untuk masa jabatan 2024 – 2029. Publik diharapkan menggunakan media sosial untuk memberikan pendapat mengenai calon gubernur dan wakil gubernur, terutama di provinsi Jawa Tengah. Salah satu *platform* media sosial yang sering

digunakan untuk mengungkapkan pendapat publik adalah Twitter. Twitter menjadi sangat penting dalam membentuk sentimen publik dan memengaruhi pilihan pemilih, karena banyak opini publik yang diungkapkan di Twitter memberikan wawasan penting tentang persepsi publik terhadap calon gubernur dan wakil gubernur Jawa Tengah (Endrik & Nugroho Agung, 2023).

Dalam menyampaikan opini dalam media sosial, tentu akan lebih baik jika dilakukan dengan memperhatikan aturan yang ada dalam syariat Islam. Dalam Islam, ada beberapa aturan yang harus diperhatikan ketika ingin menyampaikan pendapat kepada orang lain. Salah satu aturan tersebut adalah harus menginformasikan kebenaran berdasarkan fakta dan juga tidak merekayasa atau memanipulasi fakta. Hal ini dijelaskan dalam Qur'an surat Al-Hajj ayat 30 berikut ini:

ذٰلِكَ وَمَنْ يُّعَظِّمْ حُرُمٰتِ اللّٰهِ فَهُوَ خَيْرٌ لَّهٗ عِنْدَ رَبِّهٖۗ وَاُحِلَّتْ لَكُمُ الْاَنْعَامُ اِلَّا مَا يُتْلٰى عَلَيْكُمْ فَاجْتَنِبُوا الرِّجْسَ مِنَ الْاَوْثَانِ وَاجْتَنِبُوْا قَوْلَ الزُّوْرِۙ ٣٠

Artinya: "*Demikianlah (perintah Allah). Dan barangsiapa mengagungkan apa-apa yang terhormat di sisi Allah maka itu adalah lebih baik baginya di sisi Tuhannya. Dan Telah dihalalkan bagi kamu semua binatang ternak, terkecuali yang diterangkan kepadamu keharamannya, maka jauhilah olehmu berhala-berhala yang najis itu dan jauhilah perkataan-perkataan dusta."(Q.S Al-Hajj : 30).*

Perkembangan penggunaan media sosial di Indonesia, khususnya Twitter telah mengalami peningkatan signifikan dalam beberapa tahun terakhir. Menurut laporan We Are Social tahun 2023, Indonesia memiliki lebih dari 27 juta pengguna aktif Twitter, menjadikannya mendudki posisi ke-4 yang memiliki pengguna aktif di dunia.

Untuk mengumpulkan data dari Twitter, diperlukan teknik *crawling* data yang efektif. Web *crawling* atau *scraping* adalah metode untuk mengekstraksi informasi dari website secara otomatis menggunakan program atau script tertentu (Rahman et al., 2023). Dalam konteks Twitter, *crawling* data dapat dilakukan menggunakan Twitter API (*Application Programming Interface*) yang memungkinkan pengambilan data tweet, informasi pengguna dan metadata terkait secara terstruktur. Beberapa *library* Python seperti *Tweepy* dan *Tweet Harvest* dapat digunakan untuk melakukan proses *crawling* ini. Data yang diambil kemudian dapat disimpan dalam format file seperti CSV atau JSON (Pratama & Hidayat, 2024).

Tujuan dari analisis sentimen ini adalah untuk mendapatkan sentimen publik tentang Andika Perkasa-Hendrar Prihadi dan Ahmad Luthi-Taj Yasin Maimoen, dua calon gubernur dan wakil gubernur yang akan berkompetisi

dalam Pilkada Jawa Tengah 2024. Analisis ini menggunakan sentimen positif, dan negatif untuk menilai kedua kandidat di mata pengguna Twitter dan menentukan sejauh mana meraka diterima atau ditolak oleh publik (Sayarizki & Nurrahmi, 2024). Salah satu metode yang sering digunakan dalam melakukan analisis sentimen yang dikhusukan untuk pemrosesan bahasa Indonesia, yaitu IndoBERT.

Namun, analisis sentimen bahasa Indonesia memiliki tantangan khusus. Variasi dialek, penggunaan bahasa informal, dan singkatan yang umum digunakan di media sosial dapat mempersulit proses analisis. Selain itu, konteks budaya dan politik lokal juga perlu dipertimbangkan dalam interpretasi hasil. Tantangan – tantangan ini menjadikan penggunaan model khusus bahasa Indonesia seperti IndoBERT semakin penting (Delfian, 2018).

Akibat kemampuan IndoBERT untuk menangkap makna teks bahasa Indonesia dengan lebih baik, masuk akal untuk mengatakan bahwa model IndoBERT dapat lebih baik menangkap sentimen positif, atau negatif dalam data *tweet* Indonesia. Dibandingkan dengan model analisis sentimen konvensional seperti model yang didasarkan pada *lexicon-based models* atau *machine learning based* (Imron et al., 2023).

Bidirectional Encoder Representations from Transformers (BERT) adalah model pra-pelatihan yang melakukan penyematan kata dalam *Natural Language Processing* (NLP), yang mana setiap kata diubah menjadi satu set vektor numerik menggunakan arsitektur *Transformer* (Zempi et al., 2023). Sementara itu, IndoBERT merupakan varian dari model BERT pra-pelatihan yang dikembangkan secara khusus menggunakan korpus bahasa Indonesia. Sepanjang fase pelatihan BERT, ia akan melakukan pelatihan menggunakan teknik *Masked Language Modelling* (MLM) dan *Next Sentence Prediction*. Metode ini memungkinkan BERT untuk meningkatkan pemahaman bahasanya (Jayadianti et al., 2022).

Analisis sentimen menggunakan IndoBERT dapat menunjukkan hasil yang lebih baik dibandingkan dengan BERT biasa, yang mana menggunakan korpus multibahasa dan model pra-pelatihan lainnya (Wilie et al., 2020). Selain itu, IndoBERT telah menunjukkan keefektian yang lebih unggul jika dibandingkan dengan *Support Vector Machines* (SVM), *Naive Bayes*, *K-Nearest Neighbors* (KNN), *decision trees*, dan *random forest models* (Setyo Nugroho et al., 2021).

Penelitian ini akan menggunakan model IndoBERT untuk melakukan analisis sentimen terhadap calon gubernur dan wakil gubernur Jawa Tengah tahun 2024 pada media

sosial Twitter. Kumpulan data yang digunakan terdiri dari *tweet* pengguna Twitter yang dikumpulkan menggunakan pustaka *Python Tweet Harvest*. Evaluasi kinerja model akan dihitung menggunakan metode *confusion matrix*. Dengan adanya penelitian ini, diharapkan memberikan wawasan baru untuk analisis sentimen khususnya konteks pemilihan calon gubernur dan wakil gubernur tahun 2024 menggunakan IndoBERT dapat dikembangkan dan dievaluasi untuk menunjukkan keunggulan metode ini (Sayarizki & Nurrahmi, 2024).

Hasil penelitian ini diharapkan memiliki keterkaitan praktis dan teoritis yang signifikan. Secara praktis, memberikan metodologi baru dalam menganalisis opini publik di sosial media menggunakan IndoBERT dalam konteks Pemilihan Kepala Daerah (Pilkada). Secara teoritis, meningkatkan pemahaman tentang model IndoBERT dalam menganalisis sentimen publik di Twitter terkait isu-isu politik.

B. Rumusan Masalah

Berdasarkan uraian latar belakang, rumusan masalah dari penelitian tugas akhir ini adalah :

1. Bagaimana kecenderungan sentimen masyarakat terhadap calon gubernur dan wakil gubernur Jawa Tengah 2024 di media sosial Twitter?

2. Bagaimana hasil performa model IndoBERT dalam melakukan analisis sentimen tentang calon gubernur dan wakil gubernur Jawa Tengah 2024?

C. Batasan Masalah

Sebagai bentuk antisipasi agar materi yang dikaji dalam penelitian ini tidak terlalu luas, maka diberikan batasan masalah sebagai berikut :

1. Penelitian ini hanya berfokus pada tweet berbahasa Indonesia terkait calon gubernur dan wakil gubernur Jawa Tengah 2024.
2. Metode yang digunakan hanya IndoBERT.
3. Penelitian ini tidak memperhitungkan potensi pengaruh dari bot atau akun palsu dalam analisis sentimen.
4. Data yang digunakan berdasarkan kata kunci masing-masing pasangan calon gubernur dan wakil gubernur 2024 dari tanggal 24 Agustus 2024 – 31 Oktober 2024 .
5. Data akan dikategorikan menjadi 2 class sentimen yaitu negatif, dan positif.

D. Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut :

1. Menganalisis kecenderungan sentimen masyarakat terhadap calon gubernur dan wakil gubernur Jawa Tengah 2024 pada *tweets* di Twitter.
2. Mengetahui performa model IndoBERT dalam melakukan analisis sentimen calon gubernur dan wakil gubernur Jawa Tengah 2024.

E. Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut :

1. Manfaat Praktis

   Secara praktis, penelitian ini memberikan metodologi baru bagi peneliti dalam menganalisis opini publik di media sosial Twitter (X) dengan memanfaatkan model bahasa alami IndoBERT, khususnya dalam konteks Pilkada dan pemilihan kepala daerah lainnya. Hasil penelitian ini dapat menjadi referensi dan panduan bagi peneliti lain yang ingin melakukan analisis sentimen berbasis data media sosial.

2. Manfaat Teoritis

   Secara teoritis, penelitian ini memperkaya literatur akademik tentang peran dan pengaruh media sosial Twitter (X) dalam pembentukan opini

publik pada konteks politik, khususnya dalam Pilkada Jawa Tengah 2024. Selain itu, penelitian ini juga meningkatkan pemahaman tentang penerapan teknologi Natural Language Processing (NLP) dan model IndoBERT dalam menganalisis sentimen publik berbahasa Indonesia di media sosial Twitter, terutama terkait isu-isu politik dan kepemimpinan daerah.

BAB II

LANDASAN PUSTAKA

A. Kajian Teori

1. Analisis Sentimen

Dalam konteks Pemilihan Gubernur Jawa Tengah 2024, pemahaman tentang analisis sentimen menjadi sangat relevan untuk mengukur opini publik di media sosial Twitter (X). Menurut Kamus Bahasa Indonesia (KBBI) online, sentimen adalah:

1. Pendapat atau pandangan yang didasarkan pada perasaan yang berlebih-lebihan terhadap sesuatu (bertentangan dengan pertimbangan pikiran). Contoh: keputusan yang dihasilkan akan tidak adil jika rasa sentimen pribadi.
2. Emosi yang berlebihan. Contoh: rasa sentimen sebagai bangsa Indonesia akan tumbuh kuat jika kita jauh dari negeri ini.
3. Iri hati; dendam; tidak senang.
4. Reaksi yang menguntungkan. Contoh: penurunan harga saham yang disebabkan oleh sentimen pasar.

Menurut Liu, analisis sentimen atau yang sering disebut dengan *opinion mining* adalah titik fokus dari semua kegiatan manusia karena merupakan dampak mendasar pada cara kita berperilaku. Anggapan dan

gagasan pembanding seperti sentimen, evaluasi, sikap, dan emosi menjadi subjek studi yang sangat penting untuk memahami preferensi pemilih (Liu, 2012).

Analisis sentimen adalah topik eksperimen yang aktif dalam pengolahan bahasa alami. Tujuannya adalah membuat strategi yang dapat digunakan untuk mengekstraksi data subjektif, seperti pendapat atau sentimen, dari data teks. Penelitian tentang analisis sentimen berkonsentrasi pada pendapat yang memberikan nilai positif atau negatif untuk suatu sentimen (Liu, 2012). Menjamin terwujudnya hak setiap orang atas kehidupan yang baik jasmani dan rohani, tempat tinggal, dan lingkungan hidup yang sehat.

Hampir semua aspek kehidupan manusia bergantung pada nilai-nilai masyarakat. Produsen dan penyedia jasa dalam dunia bisnis dan industri selalu memperhatikan pendapat publik tentang barang yang mereka pasarkan. Pemerintah akan mempertimbangkan pendapat masyarakat tentang kebijakan yang sudah ada. Sangat mungkin bahwa pendapat masyarakat akan membantu memperbaiki dan menyempurnakan kebijakan atau produk yang telah dibuat (Darwis, 2018).

Pertumbuhan pesat media sosial seperti Twitter (X) telah membuka peluang baru dalam analisis sentimen

publik (Liu, 2012). Dalam konteks Pilgub Jawa Tengah 2024, opini atau sentimen menjadi sumber data yang kaya untuk menganalisis pendapat publik terhadap calon gubernur dan wakil gubernur. Sentimen di media sosial dapat digunakan sebagai alat pemetaan kekuatan calon kepala daerah, di mana analisis terhadap sentimen positif dan negatif dapat mengindikasikan kecenderungan preferensi pemilih dalam menentukan pilihan pada pemilihan kepala daerah.

2. Twitter

Twitter adalah layanan jejaring sosial online yang memungkinkan pengguna untuk mengirim dan membaca pesan singkat yang disebut "microblogs". Jack Dorsey membuat Twiter pada tahun 2006. Panjang *tweet* dibatasi dari 140 karakter hingga 2017. Dan sekarang panjangnya 280 karakter (Aldina Larasati, 2023).

Berdasarkan data dari We Are Social terdapat 556 juta pengguna Twitter di dunia pada Oktober 2023 dan menempati peringkat ke-7 sebagai platform media sosial yang paling aktif di dunia. Besarnya jumlah pengguna ini, termasuk di Provinsi Jawa Tengah, menjadikan Twitter sebagai sumber data yang representatif untuk menganalisis sentimen publik terhadap pasangan calon Gubernur dan Wakil Gubernur Jawa Tengah 2024.

Twitter bersifat publik sehingga status yang dibagikan dapat dilihat oleh orang lain meskipun bukan pengikutnya. Namun, pengiriman *tweet* juga dapat dibagikan hanya kepada temannya atau followers. Karakteristik ini sangat mendukung pengumpulan data untuk analisis sentimen Pilgub Jawa Tengah 2024. Berikut fitur yang ada pada twitter (Adinda Salsabila, 2022):

1. *Trending topic* adalah fitur ini dapat membantu mengidentifikasi momen-momen penting terkait Pilgub Jawa Tengah 2024 melalui *hashtag* yang sedang ramai dibicarakan.
2. *Hashtag* adalah fitur yang dapat mengelompokkan *tweet* atau pesan. Dalam penelitian ini, hashtag seperti #AndikaPerkasa, #Hendrar Prihadi, #Ahmad Luthfi, #Taj Yasin, digunakan untuk mengelompokkan *tweet.*
3. *Retweet* adalah fitur untuk membagikan *tweet* dari pengguna lain. Jumlah *retweet* dapat menjadi indikator seberapa luas persebaran opini tentang pasangan calon.
4. *Following* adalah fitur untuk menghubungkan antarpengguna atau sering disebut teman. Following dapat membantu mengidentifikasi serta

menganalisa jaringan pendukung masing-masing pasangan calon.

Data twitter dapat diambil menggunakan *Application Programming Interface* (API), API ini telah diakui sebagai 10 API teratas dunia. Twitter API banyak digunakan untuk mengumpulkan data dari platform. Sampel dari Twitter API digunakan oleh peneliti ketika ingin menghindari penyaringan *tweet* yang berdasarkan kata kunci atau akun pada saat besarnya *tweet* yang harus diakses atau tren baru yang terdeteksi realtime (Aldina Larasati, 2023). Metode yang sangat populer untuk menganalisis dari kumpulan data tweet adalah analisis sentimen.

3. Crawling Data

Crawling Data adalah proses otomatis pengambilan data dari situs web atau sumber online lainnya dengan menggunakan program komputer, yang sering disebut sebagai "web crawler". Program ini secara otomatis menjelajahi halaman-halaman web, mengunduh dan mengumpulkan informasi yang tersedia di dalamnya. Crawling data di twitter adalah suatu proses untuk mengambil atau mengunduh data dari server twitter dengan bantuan *Application Programming Interface* (API) Twitter baik berupa data user maupun data *tweet* (Oktavian, 2024). *Application Programming Interface*

(API) twitter adalah suatu program atau aplikasi yang disediakan oleh twitter untuk mempermudah *developer* lain untuk mengakses informasi yang ada di website twitter. Pendaftaran sebagai *developer* aplikasi twitter untuk menggunakan API twitter dapat dilakukan di laman *https://dev.twitter.com.* Seteah mendaftar *developer* akan mendapat *consumer key*, *consumer access*, *access token* dan *access token secret* yang akan digunakan sebagai syarat otentifikasi dari data yang akan kita ambil. Tujuan dari otentifikasi adalah untuk hak akses developer dalam mengunduh data yang ada di twitter. (Eka Sembodo et al., 2016)

4. Labelling Data

Labelling merupakan pemberian nilai sentimen terhadap suatu teks yang dapat berupa positif, negatif, atau netral. Pelabelan ini dapat dilakukan dengan berbagai metode, di antaranya adalah dengan cara manual yaitu dengan menginput sendiri nilai sentimen yang terkandung dalam teks biasanya dilakukan oleh seorang ahli pakar, kemudian dengan menggunakan fungsi TF-IDF, dan dengan *Python* dapat juga dilakukan menggunakan *library TextBLob* (Azriyan Arham, 2023).

Setiap metode *labelling* memiliki kelebihan dan kekurangannya masing-masing. Untuk labelling manual,

hasilnya akan lebih akurat tetapi memakan waktu yang cukup banyak apalagi data yang dioalah cukup besar. Jika *labelling* dilakukan secara otomatis seperti menggunakan *library TextBlob*, hasilnya tidak akan seakurat *labelling* manual tetapi lebih efisien dikarenakan hal ini dapat menghemat waktu dan tenaga (Azriyan Arham, 2023).

5. Easy Data Augmentation (EDA)

Teknik augmentasi tingkat kata yang sangat populer adalah (EDA). EDA merupakan salah satu metode augmentasi teks yang dapat digunakan untuk menangani kelas data yang imbalance. Teknik augmentasi ini dikembangkan dan dievaluasi (Wei & Zou, 2019). Penelitian (Liesting et al., 2021) menyebutkan bahwa EDA lebih unggul dibandingkan beberapa metode augmentasi lainnya dan memberikan peningkatan performa model yang digunakan untuk task tersebut.

Tujuan dilakukan EDA untuk meningkatkan variasi data tanpa mengubah makna aslinya. Kumpulan metode EDA menggunakan pendekatan tradisional yang langsung untuk meningkatkan volume data. EDA terdiri dari empat metode berikut: penghapusan acak (RD), penyisipan acak (RI), pertukaran acak (RS), penggantian sinonim (SR).

1. *Synonym Replacement* (SR)

Metode ini membuat sampel baru dari dokumen teks dengan mengganti kata-kata yang cocok dari teks dengan sinonimnya dengan mempertahankan makna teks.

2. *Random Insertation* (RI)

   Metode ini menyisipkan kata-kata baru secara acak ke dalam teks berdasarkan sinonim dari kata-kata yang sudah ada.

3. *Random Swap* (RS)

   Metode ini bertujuan untuk memilih dua kata acak yang bertukar posisi dalam teks untuk menghasilkan variasi baru.

6. *Random Deletion* (RD)

   Metode ini bertujuan secara acak menghapus kata-kata untuk menguji seberapa penting kata-kata tersebut dalam klasifikasi.

Untuk menentukan berapa jumlah kata yang disinonimkan, disisipkan, ditukar, atau dihapus dalam satu datum ditentukan oleh parameter α. Nilai parameter α ini menentukan jumlah perubahan kata sebanyak n kali dalam kalimat I. Secara matematis, jumlah n dapat ditentukan sebagai berikut.

$$n = \alpha \, x \, I$$

Dimana semakin tinggi nilai *α,* maka semakin banyak perubahan terjadi pada suatu kalimat. Penelitian ini menguji beberapa nilai parameter *α* pada keempat teknik EDA. Selain *α*, terdapat juga parameter *naua* (*Number of Augmented Sentences*) yang digunakan untuk menentukan berapa banyak kalimat augmentasi yang akan dihasilkan untuk setiap kalimat asli.

7. Bidirectional *Encoder* Representations from Transformers (BERT)

Bidirectional Encoder Representations from Transformers (BERT) adalah teknik mekanisme yang dikembangkan berdasarkan konsep *deep learning* dan *transformer* untuk mempelajari hubungan kontekstual kata-kata dalam teks. *Deep learning* pertama kali diperkenalkan oleh Geoffrey Hinton pada tahun 2006, menjadi fondasi karena kemampuannya dalam pembelajaran mesin yang menggabungkan jaringan saraf berlapis untuk belajar secara iteratif dengan meniru cara kerja otak manusia. Keunggulan *deep learning* yang dimanfaatkan BERT mencakup kemampuan *feature engineering* otomatis dan hasil yang meningkat seiring penambahan data (Pradana Rachman et al., 2021). Teknologi ini memiliki beberapa kelebihan utama menurut (Suyanto et al., 2019) deep learning memiliki keunggulan

yaitu bersifat *universal* (dapat diimplementasikan di berbagai ranah), *robust* (tahan terhadap variasi data), memiliki kemampuan *generalization* melalui *transfer learning*, dan memiliki skalabilitas tinggi.

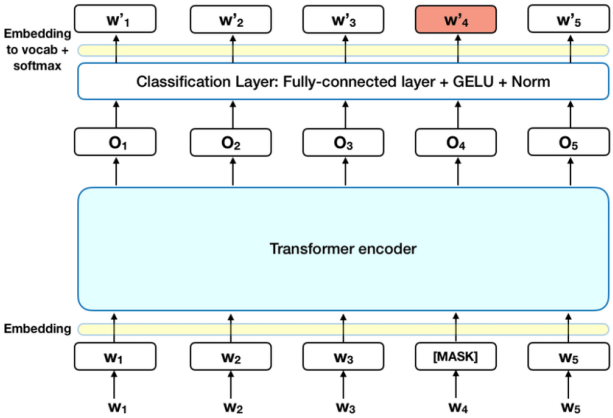
BERT juga mengadopsi arsitektur Transformer yang revolusioner, yang dirancang untuk mengatasi kendala dalam memahami konteks jarak jauh dan memodelkan hubungan antar kata dalam teks yang panjang (Vaswani et al., 2017). Komponen utama yang diadopsi adalah mekanisme *multi-head self-attention*, di mana input dilihat sebagai pasangan kunci-nilai (key dan value) dengan mekanisme encoder-decoder. Encoder digunakan untuk membaca dan memproses data input teks melalui 6 lapisan identik dengan sublapisan *self-attention* dan *feed-forward*, sementara *decoder* menghasilkan prediksi output melalui 6 lapisan serupa dengan tambahan *attention layer* untuk mengakses konten utama.

BERT telah dilatih dengan dua tujuan yaitu *Masked Language Modelling* (MLM) dan *Next Sentence Prediction* (NSP). BERT tidak dapat digunakan untuk pembuatan dan penerjemahan teks karena dilatih dengan MLM, yang hanya memungkinkan model tersebut bersifat dua arah, yang berarti dilatih dari dua arah (kiri-ke-kanan dan kanan-ke-kiri) (Fatyanosa, 2020).

MLM dilakukan untuk memungkinkan BERT memahami konteks dan makna kata-kata dengan belajar mengisi token yang hilang atau tersamar. Selama fase pre-training BERT, persentase tertentu dari token dalam teks input dipilih secara acak dan diganti dengan token [MASK] khusus. Dalam model ini, ketika ada token [MASK] dalam input, model mencoba untuk memprediksi token yang benar tanpa memperhatikan token apa yang ada di input. Untuk mengatasi hal ini, 15% token dipilih untuk dilakukan penyamaran.

a. 80% token sebenarnya diganti token [MASK]
b. 10% dari token waktu diganti token acak
c. 10% dari waktu dibiarkan tidak berubah

Saat melatih loss function BERT hanya mempertimbangkan prediksi token [MASK] dan mengabaikan prediksi yang tidak disamarkan. Ini menghasilkan model yang konvergen jauh lebih lambat dari pada model kiri ke kanan atau kanan ke kiri.



Gambar 2. 1 Proses *Masked Language Modelling* (MLM)

NSP adalah tugas pre-trained yang melibatkan pelatihan BERT untuk memprediksi apakah dua kalimat berurutan atau tidak. Tujuan dari NSP adalah untuk memungkinkan BERT memahami hubungan antar kalimat yang menangkap ketergantungan jangka panjang antar kalimat. Selama pelatihan, model diberi dua kalimat sekaligus sehingga 50% kalimat kedua muncul setelah kalimat pertama dan 50% dari waktu itu adalah kalimat acak dari seluruh *korpus*. Representasi input BERT memiliki tiga *layer embedding* yaitu:

a. *Token Embeddings*

   Lapisan pertama yang dimasuki oleh token, berfungsi sebagai representasi bentuk vektor pada tiap token. Setiap token yang berada dalam urutan

kalimat akan dipetakan ke representasi vektor multidimensi. Selanjutnya token akan diubah menjadi ID berdasarkan pada kosakata model.

b. *Segment Embeddings*

Berfungsi sebagai penanda kalimat pertama atau kalimat kedua dan juga untuk membedakan antara kalimat jika terdapat kalimat lebih dari dua.

c. *Positional Embeddings*

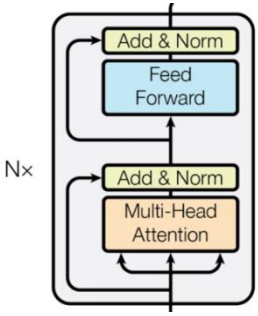
Berfungsi untuk menyimpan informasi terkait posisi kata dalam urutan kalimat.

Algoritma BERT memiliki arsitektur yang berlapis ganda (*multi-layer*) *bidirectional Trasnformers* yang menggunakan lapisan *encoder* untuk membaca *input* (Devlin et al., 2019). *Transformer* dapat belajar serta mengubah pemahaman dari hasil yang didapatkan melalui mekanisme *self-attention*, dimana mekanisme tersebut adalah cara *transformator* memodifikasi suatu kata terkait yang kemudian diubah (Ganesh et al., 2021). *Transformer* memiliki dua mekasnisme yaitu (Alammar, 2018):

1. *Encoder*

Mekanisme ini digunakan untuk membaca data input suatu teks. Terdiri dari tumpukkan = 6 lapisan identik dimana setiap lapisan memiliki 2 sublapisan *self-attention* dan *feed-forward*.

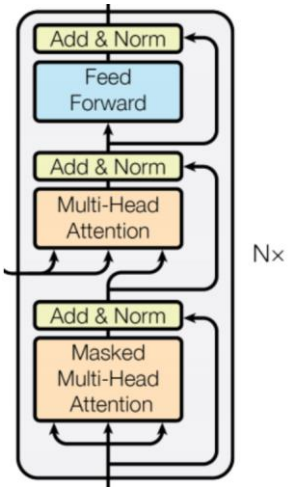
Mekanisme *encoder* pada lapisan *self-attention* dapat membantu node yang tidak hanya fokus pada kata yang divisualisasikan namun juga mendapatkan konteks dari kata tersebut. Pada lapisan *self-attention* tersebut dapat memproses semua posisi sebelumnya berada di *encoder*.



Gambar 2. 2 *Encoder*

2. *Decoder*

Mekanisme ini digunakan untuk menghasilkan urutan keluaran yang diprediksi. Terdiri dari tumpukan N = 6 lapisan yang dapat diidentifikasi dimana setiap lapisannya memiliki 2 sublapisan yang sama dengan lapisan *encoder*. Dengan tambahan attention layer di antara keduanya, dapat membantu *node* untuk mengakses konten utama yang diinginkan dengan melakukan perhatian *multi-head* pada *output encoder*.

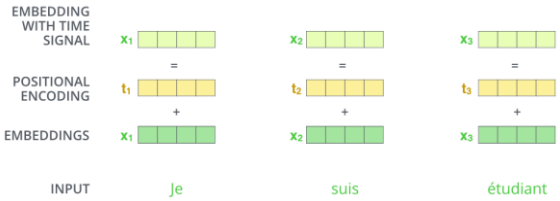


Gambar 2. 3 *Decoder*

Setelah diuraikan mekanisme didalam algoritma BERT, maka diperlihatkan keseluruhan arsitektur BERT itu sendiri. Pertama sumber dan target akan melalui suatu proses dalam lapisan *embedding* untuk menghasilkan data dengan dimensi yang sama yaitu dmodel = 512. Dalam upaya mempertahankan informasi posisi maka *encoding sinusoid-wave-based position* diaplikasikan dan dijumlahkan dengan *embedding* keluaran. Tahap akhirnya adalah menambahkan lapisan *softmax* dan linear di dalam keluaran *decoder* terakhir (Alammar, 2018).

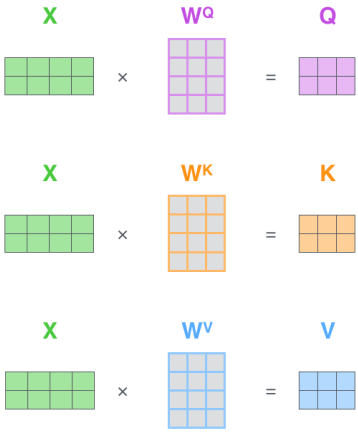
1. Seluruh input yang masuk ke dalam proses *encoder* akan dikonversi ke dalam bentuk *list vector* dengan menggunakan *embedding*.

Kemudian melewati *positional encoding* yang berguna untuk menandai posisi setiap kata. Ukuran yang dimiliki masing-masing vektor *input* adalah 512.



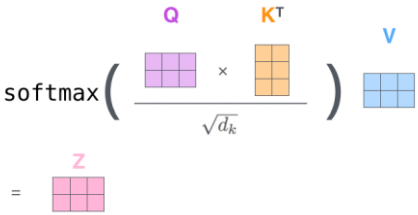
Gambar 2. 4 *Vektor Input*

2. Langkah selanjutnya, vektor *input* akan dilewati dua lapisan dari setiap *encoder*. Lapisan yang akan dilewati adalah lapisan *self-attention* dan *feed-forward neural network*. Pada lapisan *self-attention* akan menghasilkan tiga matriks dari masing-masing matriks *input* yang ada yaitu *query, key,* dan *value* dimana tiga matriks tersebut merupakan hasil dari perkalian pada proses *embedding*.



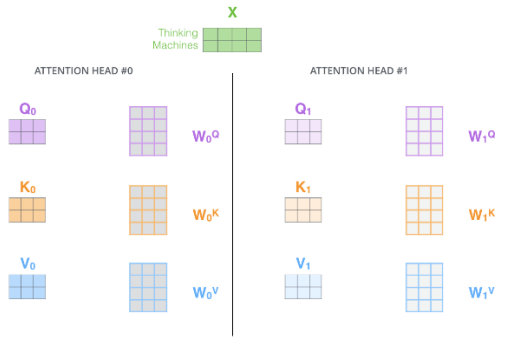
Gambar 2. 5 Proses Perhitungan Matriks pada *Self-Attention*

Selanjutnya menghitung skor pada proses *self-attention*. Skor diperoleh dengan mengalikan secara *dot-product* matriks kueri dengan matriks kunci, kemudian hasilnya akan dilanjutkan kedalam operasi *softmax* yang berfungsi untuk menormalisasi skor supaya bernilai positif dan jumlah seluruh skor sama dengan satu. Setelah proses tersebut dilakukan, maka hasilnya akan dikalikan dengan matriks nilai dengan tujuan menjaga keaslian *values* dari tiap kata yang ingin difokuskan serta menghapus kata yang tidak relevan.

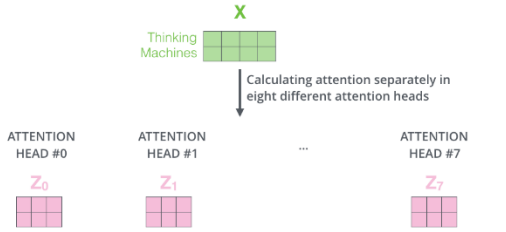


Gambar 2. 6 Proses Perhitungan Skor *Input*

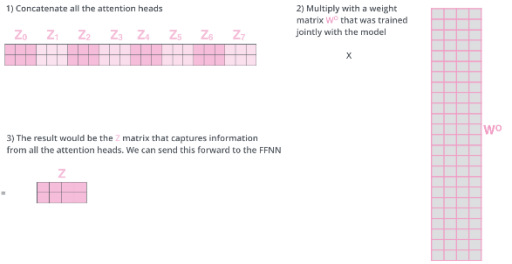
Dikarenakan arsitektur *Transformer* memiliki delapan *attention head* maka perhitungan *self-attention* dilakukan sebanyak delapan kali. Pada *multi-headed attention*, matriks bobot yang digunakan pada perhitungan diatas berbeda pada setiap *attention head* yang berakibat pada berbedanya matriks kueri, kunci dan nilai. Setelah dilakukan penghitungan selama delapan kali maka matriks-matriks tersebut akan digabung dan dikalikan dengan matriks bobot (W^o) sehingga hasilnya adalah sebuah matriks. Hal ini dilakukan karena *feed-forward layer* hanya menerima input sebuah matriks saja.



Gambar 2. 7 Ilustrasi *Attention Head* 0 dan 1

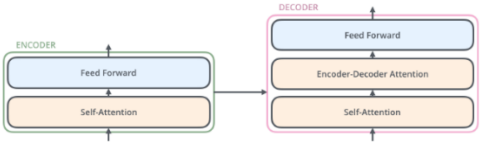


Gambar 2. 8 Ilustrasi Hasil dari 8 *Attention Head*



Gambar 2. 9 proses Penggabungan *Attention Head*

3. *Output* yang dihasilkan dari *encoder* merupakan vektor *key* dan vektor *value* yang akan menuju ke *decoder*. Setiap *input* dan *output* dari lapisan *self-attention* dan *feed-forward neural network* di *encoder* akan di proses di *decoder* yang akan menjadi lapisan *add* dan *norm*. Proses yang terjadi pada tahap *decoder* sama seperti yang terjadi pada tahap *encoder*, namun dibedakan dengan lapisan *attention* yang membantu *decoder* fokus pada bagian kata yang relevan. *Output* dari setiap tahap akan masuk ke dalam *decoder* dan memberikan hasil yang serupa sepreti tahap *encoder*. Tumpukan yang terjadi pada *decoder* akan menghasilkan vektor nilai *float*, dimana vektor nilai *float* akan dirubah menjadi kata-kata dengan bantuan lapisan tambahan yaitu lapisan *fully connected* dan lapisan *softmax.*



Gambar 2. 10 Proses *encoder* dan *decoder*

8. IndoBERT

IndoBERT adalah metode modifikasi dari BERT-base yang sudah ada dengan mengikuti konfigurasi dari BERT-base (*uncased*) yang memiliki 12 *hidden layers* dengan masing-masing memiliki 768 dimensi, 12 *attention heads*, dan 2.072 dimensi *feed-forward hidden layers*. Secara total, IndoBERT sudah dilatih lebih dari 220 juta kata Bahasa Indonesia dengan 3 sumber utama yaitu Indonesia Wikipedia (74 juta kata), artikel seperti Liputan6, Kompas, Tempo (55 juta kata), dan Indonesia Web Corpus (90 juta kata) (Koto et al., 2020). IndoBERT dilatih dengan dataset Indo4B dan TPUv3-8 dalam dua fase (Wilie et al., 2020).

Terdapat empat model pada IndoBERT yaitu IndoBERTBASE, IndoBERTLARGE, IndoBERT-liteBASE, IndoBERT-liteLARGE. Semua model IndoBERT dilatih menggunakan TPU v3-8 dalam dua fase. Pada fase pertama, dilakukan dengan input data sebesar 128 dan waktu pelatihannya selama 35,38,89 dan 134 jam untuk model pada IndoBERT. Pada fase kedua IndoBERTBASE memiliki *batch size* sebesar 256 dan *learning rate* sebesar 2e-5 sedangkan IndoBERTLARGE pada fase pertama memiliki *batch size* sebesar 256 dan *learning rate* sebesar 1e-4 dan pada fase kedua diturunkan

menjadi *batch size* sebesar 128 dan *learning rate* sebesar 8e-5.

Model	Maximum Sequence Length = 128				Maximum Sequence Length = 512			
	Batch Size	Learning Rate	Steps	Duration (Hr.)	Batch Size	Learning Rate	Steps	Duration (Hr.)
IndoBERT-lite$_{BASE}$	4096	0.00176	112.5 K	38	1024	0.00088	50 K	23
IndoBERT$_{BASE}$	256	0.00002	1 M	35	256	0.00002	68 K	9
IndoBERT-lite$_{LARGE}$	1024	0.00044	500 K	134	256	0.00044	129 K	45
IndoBERT$_{LARGE}$	256	0.0001	1 M	89	128	0.00008	120 K	32

Gambar 2. 11 Summary Pelatihan *IndoBERT*

9. Representasi Input dan Output BERT

Sebelum melakukan pemodelan dengan algoritma IndoBERT, data yang berupa kata perlu diubah menjadi input yang bisa diterima oleh model. Terdapat dua proses yang harus dilakukan supaya data dapat diterima oleh model yaitu tokenisasi dan penyesuaian label. Berikut adalah langkah-langkah yang dilakukan untuk merepresentasikan input pada model IndoBERT (Pradana, 2024):

1. Tahap pertama data akan di tokenisasi berdasarkan token menggunakan WordPiece. Kata akan diperikasa berdasarkan *vocabulary* yang tersedia pada IndoBERT. Kata yang tidak termuat dalam *vocabulary* akan diubah menjadi token *unknown* [UNK]. Kata dipecah menjadi dua buah sub-kata yang mana sub-kata pertama adalah kata yang ada di dalam *vocabulary* dan sub-kata kedua adalah kata yang memuat sufiks dari sub-kata pertama ditandai dengan ## di depan sub-kata kedua.

Gambar 2. 12 Tokenisasi dengan *WordPiece*

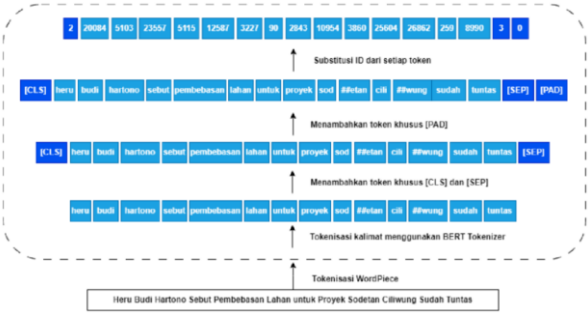
2. Setiap kalimat input akan ditambahkan token khusus yaitu token [CLS] pada awal kalimat dan token [SEP] pada akhir kalimat. Token [CLS] diartikan sebagai penanda awal kalimat dan juga sebagai pengumpulan rata-rata pada token kata. Token [SEP] dimaknai sebagai pemisah antar kalimat. Tahap ini disebut dengan *token embedding*.

Gambar 2. 13 *Token Embedding*

3. Setiap kalimat input memiliki panjang yang tidaklah sesuai satu sama lain, maka perlu dilakukan proses penyeragaman panjang kata. Hal ini dilakukan dengan menambahkan token [PAD].

Gambar 2. 14 Menambahkan *token* [PAD]

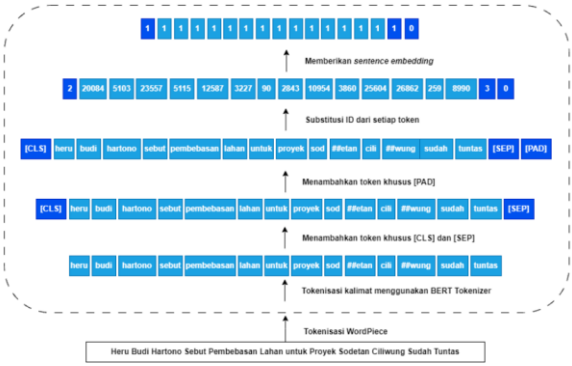
4. Selanjutnya adalah proses mensubstitusikan ID dari setiap token. ID merupakan nomor unik yang didapat dari urutan indeks kata berdasarkan kemunculannya. Setiap kata dan sub-kata harus diubah dalam bentuk ID karena pada saat fase *pre-trained* input yang diterima hanyalah ID.



Gambar 2. 15 Substitusi ID Token

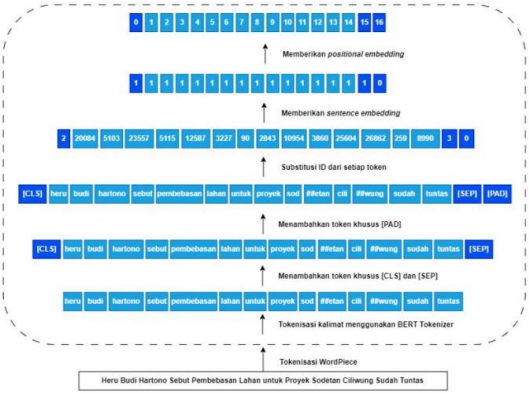
5. Proses berikutnya adalah *sentence embedding* adalah untuk memisahkan atau membedakan satu kalimat lain menggunakan token [PAD]. Semua

token selain [PAD] akan diubah menjadi angka 1 dan [PAD] diubah menjadi angka 0.



Gambar 2. 16 *Sentence Embedding*

6. Proses berikutnya adalah *positional embedding* untuk mengetahui konteks dari kalimat, karena jika ditemukan kata yang sama pada kalimat maka akan dimaknai berbeda karena posisinya berbeda. Setiap sub-kata memiliki urutan yang sama dengan kata sebelum proses pemisahan.

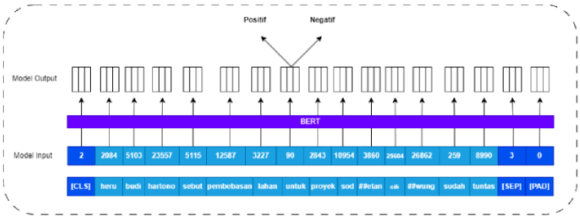


Gambar 2. 17 *Positional Embedding*

7. Proses diatas adalah proses representasi input pada model BERT. Input tersebut akan diterima oleh model IndoBERT yang kemudian akan diteruskan melalui tumpukan *encoder* yang di dalamnya ada *self-attention* dan *feed-forward network*, proses ini akan diulangi sebanyak 12 kali.

Gambar 2. 18 Representasi *Input* BERT

8. Setelah melakukan *encoder* setiap token akan menghasilkan output berupa vektor *hidden layer* berukuran 768. Pada model klasifikasi token khusus [CLS] akan digunakan sebagai input.



Gambar 2. 19 Proses Klasifikasi BERT

10. Hyperparameter

Hyperparameter adalah parameter yang tidak ditentukan oleh mesin. *Hyperparameter* diatur secara manual sesuai dengan kebutuhan untuk membantu memperkirakan parameter. Pemilihan parameter yang tepat dibutuhkan agar model dapat bekerja lebih optimal. Oleh sebab itu, *hyperparameter tuning* merupakan tahapan yang penting dalam *machine learning.*

a. *Max sentence length*, merupakan batas terpanjang kalimat yang digunakan dalam pembentukan model.
b. *Batch size* pada *machine learning* mengacu pada jumlah contoh pelatihan yang digunakan dalam satu iterasi (Rochmawati et al., 2021). Semakin besar *batch size* maka waktu yang dibutuhkan untuk proses pelatihan menjadi semakin lama.
c. *Learning Rate*, merupakan salah satu parameter yang ditetapkan untuk menghitung nilai koreksi

bobot selama proses pelatihan (Rochmawati et al., 2021).

d. *Drop out* adalah teknik yang digunakan untuk mencegah *overfitting* ketika proses pelatihan pada *machine learning*.
e. *Epoch* adalah *hyperparameter* yang menentukan berapa kali algoritma mempelajari dataset pelatihan. Satu *epoch* berarti setiap sampel dari dataset mendapatkan kesempatan untuk memperbarui internal model parameter. Epoch terdiri dari beberapa batch.

11. Metrik Evaluasi

Untuk mengukur kinerja model yang telah dilatih untuk pemrosesan bahasa alami dapat dilakukan dengan menggunakan metrik evaluasi. Metrik evaluasi adalah alat penting untuk mengukur seberapa baik model IndoBERT mampu menyelesaikan tugas pemrosesan bahasa yang diberikan. *Confusion matrix* yang biasa digunakan adalah akurasi, *recall*, presisi dan *f1-score* untuk mengukur metrik evaluasi biasa menggunakan metode yang bernama *confusion matrix*. *Confusion matrix* adalah tabel yang merangkum kinerja model kemudian digunakan untuk mengidentifikasi kelemahan dalam operasi algoritma (Jayaswal, 2020).

Jika data tersebut aktualnya adalah data positif dan diprediksi sebagai data positif juga, maka akan dihitung sebagai *true positve*. Tetapi jika data tersebut diprediksi positif, maka akan dihitung *false negative*. Dan jika data tersebut aktualnya adalah data negatif dan diprediksi sebgai data negatif juga, maka akan dihitung sebagai *true negative*. Tetapi jika data tersebut diprediksi positif maka akan dihitung *false positive*. Hasil klasifikasi biner pada suatu dataset dapat direpresentasikan dengan matriks 2 x 2 yang disebut *confusion matrix*. (Fauzan, 2022)

Class		*Actual Values*	
		Positif	Negatif
Predicted Values	Positif	TP	FP
	Negatif	FN	TN

Gambar 2. 20 *Confusion Matrix*

Confusion matrix sangat berguna untuk menganalisis kualitas *classifier* dalam mengenali data-data dari kelas yang ada. TP dan TN menyatakan bahwa kualitas *classifier* mengenali data dengan benar, artinya data positif dikenali sebagai *positive* dan data negatif dikenali sebagai *negative*. Sebaliknya, FP dan FN menyatakan bahwa *classifier* salah dalam mengenali data, data *negative* dikenali sebagai positif dan data *positive* dikenali sebagai negatif. Terdapat

beberapa rumus umum yang dapat digunakan untuk menghitung performa klasifikasi. (Praptiwi, 2018)

a. *Accuracy* adalah metrik yang mengukur sejauh mana model memprediksi dengan benar

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

b. *Precision* adalah metrik yang mengukur sejauh mana model memprediksi kelas positif dengan benar

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

c. *Recall* adalah metrik yang mengukur sensitivitas model terhadap kejadian positif

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

d. *F1-Score* adalah metrik perbandingan antara *recall* dan presisi yang dibobotkan

$$F1\ Score = 2 * \frac{precision + recall}{precision + recall}$$

B. Kajian Penelitian yang Relevan

Penelitian ini membutuhkan referensi sebagai sumber informasi lain yang bertujuan untuk mendukung penelitian dan mendapatkan perbandingan antara penelitian satu dengan penelitian-penelitian lainya. Beberapa penelitian yang relevan dengan penelitian ini adalah :

Tabel 2. 1 Kajian Penelitian yang Relevan

Judul Penelitian	Hasil	Penulis	Tempat & Tahun Publikasi
Analisis Sentimen Pemilu 2024 dengan *Naive Bayes* Berbasis *Particle Swarm Optimization*	Penelitian ini bertujuan untuk menganalisis sentimen publik di twitter yang berkaitan dengan pemilu 2024 menggunakan metode *Naive Bayes* berbasis *Particle Swarm Optimization* (PSO). Data yang digunakan adalah 1000 *tweet*. Hasil penelitian menunjukkan jika menggunakan	Tommy Dwi Putra, Ema Utami, Mei P. Kurniawan	Jurnal Informatika & Komputer (EXPLORE), tahun 2023

	Naïve Bayes tanpa *Particle Swarm Optimization* (PSO) memiliki nilai akurasi 73,67 sedangkan jika menggunakan *Particle Swarm Optimization* (PSO) memiliki akurasi 78,33. Hal ini menunjukkan tingkat akurasi dari algoritma *Naïve Bayes* berbasis *Particle Swarm Optimization* (PSO) lebih unggul.		

Analisis Sentimen Calon Presiden 2024 di Media Sosial X Menggunakan *Naïve Bayes* dan *Smote*	Penelitian ini bertujuan untuk menganalisis sentimen publik di twitter dalam konteks pemilihan presiden. Metode yang digunankan *naive bayes* dan teknik *Synthetic Minority Oversampling* (SMOTE). Hasil penelitian menunjukkan jika dengan SMOTE akurasi meningkat menjadi 88% pada dataset Ganjar-Mahfud, sementara tanpa SMOTE akurasi	Muhammad Hafidz Ardian Sunata, Faldy Irwiensyah, Firman Noor Hasan	Jurnal Media Informatika Budidarma, tahun 2024

	rendah 69% pada dataset Anies-Imin. Dari total 1589 tweet dengan sentimen positif, sekitar 27,7% mengarah ke Anies-Imin, 28,7% ke Prabowo-Gibran, dan 45,3% ke Ganjar-Mahfud.		
Analisis Sentimen Masyarakat Media Sosial Twitter Terhadap Kinerja Pejabat Gubernur DKI Jakarta Menggunaka	Penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap kinerja Gubernur DKI Jakarta yaitu Heru Budi Hartono. Penelitian menggunakan	Lambang Surya Pradana	Skripsi Program Studi Matematika Fakultas Sains dan Teknologi UIN Syarif Hidayatullah, tahun 2024

n Model IndoBERT	algoritma IndoBERT. Hasil penelitian menunjukkan performa yang baik dengan akurasi sebesar 90,5% dan *f1-score* 90,49%.		
Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models	Penelitian ini bertujuan untuk membantu pemerintah dan institusi terkait untuk memahami persepsi publik sebelum diselenggarakan pemilu 2024. Penelitian ini membandingkan beberapa model yaitu: *TextBlob* dengan akurasi	Lenggo Geni, Evi Yulianti, Dana Indra Sensuse	Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)

	35%, *Multinomial Naive Bayes* 81%, *SVM* 80%, *IndoBERT base-p1* 83,5%, *IndoBERT base-p2* 82%, *IndoBERT large-p1* 81,5%, *IndoBERT large-p2* 82%. Berdasarkan hasil tersebut maka didapatkan Kesimpulan bahwa *IndoBERT base-p1* memiliki tingkat akurasi yang lebih tinggi dibandingkan model lainnya.		

IndoBERT For Indonesian Fake News Detection	Penelitian ini bertujuan untuk mengidentifikasi berita palsu yang mana dapat menyesatkan masyarakat. Sistem pendeteksi berita palsu diperlukan karena dapat mengurangi penyebaran informasi yang menyesatkan. Penelitian ini mendapatkan data sebanyak 3.465 berita palsu dan 766 berita asli data dikumpulkan dari *turnbackhoax.id.*	Sani Muhamad Isa, Gary Nico, Mikhael Permana	ICIC Express Letters. Volume 16, Number 3, Maret 2022

	Penelitian ini menggunakan model IndoBERTBase dan didapatkan akurasi sebesar 94.66%. Penelitian ini juga membandingakan dengan model lain seperti *TF-IDF + SVM* menghasilkan akurasi 90% dan *TF-IDF + Naive Bayes* menghasilkan akurasi 83%. Hal ini menunjukkan bahwa model IndoBERTBase menunjukkan model yang		

	paling unggul dalam melakukan analisis sentimen dibandingakan model *machine learning* yang lain.		
Easy Data Augmentation untuk Data yang *Imbalance* pada Konsultasi Kesehatan Daring	Pada penelitian ini *Easy Data Augmentation* (EDA) meningkatkan konsultasi kesehatan daring, dengan model *Random Forest* mencapai akurasi 88,86%. Dari empat teknik yang diuji (SR, RI, RS, RD), *Random Deletion* (RD)	Anisa nur Azizah, Misbachul Falach Asy'ari, Ifnu Wisma Dwi Prastya, Diana Purwitasari	JTIK: Jurnal Teknologi Informasi dan Ilmu Komputer Vol 10 No 5: Oktober 2023

	dengan parameter 0,4 memberikan hasil terbaik. Augmentasi juga meningkatkan prediksi pada kelas minoritas, terutama dengan model AdaBoost yang naik 3,78%. Secara keseluruhan, EDA efektif menangani data tidak seimbang, dengan RD sebagai teknik EDA paling optimal.		

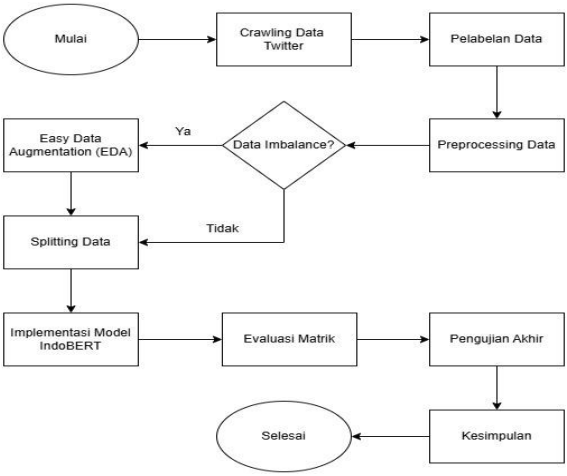
Berdasarkan informasi yang tersedia dalam tabel kajian penelitian, analisis sentimen menggunakan model IndoBERT sudah banyak diteliti namun tentang topik Pilkada jarang

ditemui. Penelitian ini berusaha mengisi gap tersebut dengan lebih menuju ke suatu daerah yaitu Jawa Tengah. Model IndoBERT digunakan karena dapat memahami teks bahasa Indonesia lebih baik dibandingkan dengan metode lainnya. Selain itu, penelitian ini juga menerapkan *Easy Data Augmentation* (EDA) untuk menangani data yang tidak seimbang dalam melakukan analisis sentimen. Dengan teknik seperti *Synonym Replacement* (SR), *Random Insertion* (RI), *Random Swap* (RS), dan *Random Deletion* (RD), EDA dapat meningkatkan kualitas dataset dan kinerja model.

BAB III
METODOLOGI PENELITIAN

Penelitian memerlukan alur proses yang sistematis dan terperinci pada setiap tahapannya. Metode penelitian berfungsi sebagai pedoman komprehensif yang memberikan struktur dan urutan yang jelas dalam menjalankan penelitian. Setiap proses dalam alur ini berperan sebagai panduan dari tahap awal hingga tahap akhir penelitian untuk memastikan bahwa setiap langkah dilaksanakan dengan baik. Tujuan metodologi penelitian adalah untuk memastikan bahwa tahapan-tahapan penelitian dilakukan secara terstruktur.

A. Tahapan Penelitian



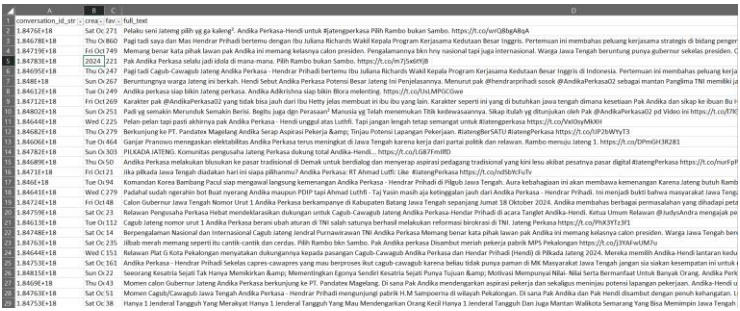
Gambar 3. 1 Tahapan Penelitian

B. Uraian Metodelogi

1. Crawling data Twitter

Pengambilan data dimulai dengan mengekstraksi data dari *platform* Twitter menggunakan teknik *crawling* dengan *library Tweet Harvest*. Pengumpulan data dilakukan dengan menggunakan nama kedua calon gubernur dan wakil gubernur Jawa Tengah 2024 yaitu: ”Andika Perkasa-Hendrar Prihadi” dan ”Ahmad Luthi-Taj Yasin Maimoen”. Data dikumpulkan dari tanggal 24 Agustus 2024 hingga 31 Oktober 2024. Dengan dilakukanya pendekatan ini memastikan bahwa data mentah yang dikumpulkan dari *tweet* mencakup beragam perspektif pengguna dan tanggapan terhadap kedua calon gubernur dan wakil gubernur.

Hasil *crawling* data mendapatkan total 2936 *tweets* tentang kedua pasangan calon gubernur Jawa Tengah 2024. Setelah proses *crawling* selesai data disimpan ke dalam file yang berbentuk *Comma Separated Value* (CSV). Proses ini memastikan bahwa kumpulan data mencerminkan berbagai pendapat dan tanggapan dari berbagai sumber mengenai kedua calon gubernur dan wakil gubernur Jawa Tengah 2024.



	A	B	C	D
1	conversation_id_str	crea	fav	full_text
2	1.8476E+18	Sat Oc	271	Pelaku seni Jateng pilih yg ga kaleng². Andika Perkasa-Hendi untuk #JatengberkasaPilih Rambo bukan Sambo. https://t.co/wrQBbpA8qA
3	1.8468E+18	Thu Oc	860	Pagi tadi saya dan Mas Hendrar Prihadi bertemu dengan Ibu Juliana Richards Wakil Kepala Program Kerjasama Kedutaan Besar Inggris. Pertemuan ini membahas peluang kerjasama strategis di bidang pengen
4	1.8473E+18	Fri Oct	749	Memang benar kata pihak lawan pak Andika ini memang kelasnya calon presiden. Pengalamannya bkn hny nasional tapi juga internasional. Warga Jawa Tengah beruntung punya gubernur sekelas presiden. C
5	1.8478E+18	2024	221	Pak Andika Perkasa selalu jadi idola di mana-mana. Pilih Rambo bukan Sambo. https://t.co/m7jSx6Yrj8
6	1.8469E+18	Thu Oc	247	Pagi tadi Cagub-Cawagub Jateng Andika Perkasa - Hendrar Prihadi bertemu Ibu Juliana Richards Wakil Kepala Program Kerjasama Kedutaan Besar Inggris di Indonesia. Pertemuan ini membahas peluang kerja
7	1.848E+18	Sun Oc	267	Beruntungnya warga Jateng ini berkah. Hendi Sebut Andika Perkasa Potensi Besar Jateng Ini Penjelasannya. Menurut pak @hendrarprihadi sosok @AndikaPerkasa02 sebagai mantan Panglima TNI memiliki ja
8	1.8461E+18	Tue Oc	249	Andika perkasa siap bikin Jateng perkasa. Andika Adikrishna siap bikin Blora melenting. https://t.co/UsLMP6CGwe
9	1.8471E+18	Fri Oct	269	Karakter pak @AndikaPerkasa02 yang tidak bisa jauh dari Ibu Hetty jelas membuat ini ibu ibu yang lain. Karakter seperti ini yang di butuhkan jawa tengah dimana kesetiaan Pak Andika dan sikap ke ibuan Ibu H
10	1.8480E+18	Sun Oc	251	Padi yg semakin Merunduk Semakin Berisi. Begitu juga dgn Perasaan³ Manusia yg Telah menemukan Titik kedewasaannya. Sikap itulah yg ditunjukan oleh Pak @AndikaPerkasa02 pd Video ini https://t.co/f7K
11	1.8464E+18	Wed C	225	Pelan pelan tapi pasti akhirnya pak Andika Perkasa - Hendi unggul atas Luthfi. Tapi jangan lengah tetap semangat untuk #JatengperkasaHttps://t.co/VxiOsyMkRH
12	1.8468E+18	Thu Oc	279	Berkunjung ke PT. Pandatex Magelang Andika Serap Aspirasi Pekerja & Tinjau Potensi Lapangan Pekerjaan. #JatengBerSATU #JatengPerkasa https://t.co/IJP2bWYyT3
13	1.8460E+18	Tue Oc	464	Ganjar Pranowo menegaskan elektabilitas Andika Perkasa terus meningkat di Jawa Tengah karena kerja dari partai politik dan relawan. Rambo menuju Jateng 1. https://t.co/DPmiGH3R2B1
14	1.8478E+18	Sun Oc	303	PILKADA JATENG. Komunitas pengusaha Jateng Perkasa dukung total Andika-Hendi... https://t.co/LiG87FmIfD
15	1.8468E+18	Thu Oc	50	Andika Perkasa melakukan blusukan ke pasar tradisional di Demak untuk berdialog dan menyerap aspirasi pedagang tradisional yang kini lesu akibat pesatnya pasar digital #JatengPerkasa https://t.co/nurFpP
16	1.8471E+18	Fri Oct	21	Jika pilkada Jawa Tengah diadakan hari ini siapa pilihanmu? Andika Perkasa: RT Ahmad Luthfi: Like #JatengPerkasa https://t.co/hd5JKYuTv
17	1.846E+18	Tue Oc	94	Komandan Korea Bambang Pacul siap mengawal langsung kemenangan Andika Perkasa - Hendrar Prihadi di Pilgub Jawa Tengah. Aura kebahagiaan ini akan membawa kemenangan Karena Jateng butuh Ramb
18	1.8464E+18	Wed C	279	Padahal sudah ngerahin bot Buat nyerang Andika maupun PDIP tapi Ahmad Luthfi - Taj Yasin masih aja ketinggalan jauh dari Andika Perkasa - Hendrar Prihadi. Ini menjadi bukti bahwa masyarakat Jawa Teng
19	1.8472E+18	Fri Oct	48	Calon Gubernur Jawa Tengah Nomor Urut 1 Andika Perkasa berkampanye di Kabupaten Batang Jawa Tengah sepanjang Jumat 18 Oktober 2024. Andika membahas berbagai permasalahan yang dihadapi peta
20	1.8475E+18	Sat Oc	23	Relawan Pengusaha Perkasa Hebat mendeklarasikan dukungan untuk Cagub-Cawagub Jateng Andika Perkasa-Hendar Prihadi di acara Tangket Andika-Hendi. Ketua Umum Relawan @JudyAndra mengajak pe
21	1.8461E+18	Tue Oc	112	Cagub Jateng nomor urut 1 Andika Perkasa berani ubah aturan di TNI salah satunya berhasil melakukan reformasi birokrasi di TNI. Jateng Perkasa https://t.co/PhX3YTz3f1
22	1.8474E+18	Sat Oc	14	Berpengalaman Nasional dan Internasional Cagub Jateng Jendral Purnawirawan TNI Andika Perkasa Memang benar kata pihak lawan pak Andika ini memang kelasnya calon presiden. Warga Jawa Tengah ber
23	1.8476E+18	Sat Oc	235	@bab merah memang seperti ibu cantik-cantik dan cerdas. Pilih Rambo bkn Sambo. Pak Andika perkasa Disambut meriah pekerja pabrik MPS Pekalongan https://t.co/JYKkFwUMJ1u
24	1.8464E+18	Wed C	153	Relawan Plat G Kota Pekalongan menyatakan dukungannya kepada pasangan Cagub-Cawagub Andika Perkasa dan Hendar Prihadi (Hendi) di Pilkada Jateng 2024. Mereka memilih Andika-Hendi lantaran kedu
25	1.8475E+18	Sat Oc	161	Andika Perkasa - Hendrar Prihadi Sekelas capres-cawapres yang mau berproses ikut cagub-cawagub karena beliau tidak punya paman di MK Masyarakat Jawa Tengah jangan sia siakan kesempatan ini untuk
26	1.8481E+18	Sun Oc	22	Seorang Kesatria Sejati Tak Hanya Memikirkan & Mementingkan Egonya Sendiri Kesatria Sejati Punya Tujuan & Motivasi Mempunyai Nilai- Nilai Serta Bermanfaat Untuk Banyak Orang. Andika Perk
27	1.8469E+18	Thu Oc	43	Momen calon Gubernur Jateng Andika Perkasa berkunjung ke PT. Pandatex Magelang. Di sana Pak Andika mendengarkan aspirasi pekerja dan sekaligus meninjau potensi lapangan pekerjaan. Andika-Hendi u
28	1.8476E+18	Sat Oc	51	Momen Cagub/Cawagub Jawa Tengah Andika Perkasa - Hendrar Prihadi mengunjungi pabrik H.M Sampoerna di wilayah Pekalongan. Di sana Pak Andika dan Pak Hendi disambut dengan penuh kehangatan. L
29	1.8475E+18	Sat Oc	38	Hanya 1 Jenderal Tangguh Yang Merakyat Hanya 1 Jenderal Tangguh Yang Mau Mendengarkan Orang Kecil Hanya 1 Jenderal Tangguh Dan Juga Mantan Walikota Semarang Yang Bisa Memimpin Jawa Tengah

Gambar 3. 2 Hasil *Crawling Data* Twitter

2. Labelling Data

Setelah data terkumpul maka akan dilakukan proses pelabelan yang bertujuan agar model dapat memahami pola dan karakteristik dari sentimen setiap teks. Proses pelabelan dilakukan oleh seorang ahli pakar bahasa Indonesia yang memahami berbagai aspek kata tentang bahasa Indonesia. Hal ini dilakukan agar setiap data *tweet* di klasifikasikan dengan benar menurut ahli pakar bahasa Indonesia. Data di klasifikasikan menjadi 2 kelas yaitu negatif, dan positif.

3. Pre-Processing Data

Pada tahap ini akan dijelaskan terkait pengolahan teks beserta hasilnya. Pengolahan teks diperlukan sebagai proses pertama pengolahan data, hal ini harus dilakukan untuk membersihkan data yang diolah agar tidak mengandung *noise*, sehingga memiliki dimensi yang lebih kecil dan lebih terstruktur. Data akan diolah dengan lebih baik dan

menginterpretasi hasil pengamatan. Berikut pengolahan awal teks yang akan dilakukan dalam penelitian ini:

a. *Cleaning Text*

Pada proses ini akan dilakukan penghilangan karakter tidak valid seperti tanda baca, angka serta simbol dilakukan pada *tweet* yang terdapat didalamnya. Tahapan ini biasa disebut dengan *text cleaning*.

Tabel 3. 1 Hasil Cleaning text

Tweet	***Tweet* Bersih**
Ganjar Pranowo menegaskan elektabilitas Andika Perkasa trs meningkat di Jawa Tengah krn kerja dari partai politik dan relawan. Rambo menuju Jateng 1. https://t.co/DPmGH3R281	Ganjar Pranowo menegaskan elektabilitas Andika Perkasa trs meningkat di Jawa Tengah krn kerja dari partai politik dan relawan Rambo menuju Jateng 1
Pengalaman Jd Pemimpin di Jateng Prabowo Subianto Menyebut Ahmad Luthfi dan Taj Yasin Orang Tepat https://t.co/vnoFT0C3cL	Pengalaman Jd Pemimpin di Jateng Prabowo Subianto Menyebut Ahmad Luthfi dan Taj Yasin Orang Tepat

b. *Case Folding*

Pada proses ini, melakukan penyeragaman karakter pada teks seperti mengubah teks menjadi huruf kecil (*lowercase*).

Tabel 3. 2 Hasil *Case Folding*

Tweet	***Tweet* Bersih**
Ganjar Pranowo menegaskan elektabilitas Andika Perkasa trs meningkat di Jawa Tengah krn kerja dari partai politik dan relawan Rambo menuju Jateng 1	ganjar pranowo menegaskan elektabilitas andika perkasa trs meningkat di jawa tengah krn kerja dari partai politik dan relawan. rambo menuju jateng 1
Pengalaman jd Pemimpin di Jateng Prabowo Subianto Menyebut Ahmad Luthfi dan Taj Yasin Orang Tepat	pengalaman jd pemimpin di jateng prabowo subianto menyebut ahmad luthfi dan taj yasin orang tepat

c. *Normalization*

Proses mengubah kata tidak baku atau singkatan menjadi bentuk bakunya. Normalisasi digunakan untuk menangani bahasa tidak formal, slang, ejaan yang tidak standar. Dalam melakukan normalisasi dapat menggunakan kamus yang sering digunakan yaitu '*colloquial-indonesian-lexicon.csv'*, '*kbba.txt*', '*slangword.txt'*, '*slang.csv'* yang bisa diakses melalui *GitHub*.

Tabel 3. 3 Hasil *Normalization*

Tweet	***Tweet* Bersih**
ganjar pranowo menegaskan elektabilitas andika perkasa trs meningkat di jawa tengah krn kerja dari partai politik dan relawan. rambo menuju jateng 1	ganjar pranowo menegaskan elektabilitas andika perkasa terus meningkat di jawa tengah karena kerja dari partai politik dan relawan. rambo menuju jateng 1.
pengalaman jd pemimpin di jateng prabowo subianto menyebut ahmad luthfi	pengalaman jadi pemimpin di jateng prabowo subianto menyebut ahmad luthfi

dan taj yasin orang tepat	dan taj yasin orang tepat

4. Easy Data Augmentation

Easy Data Augmentation (EDA) bekerja dengan melakukan modifikasi kecil pada kalimat asli untuk menghasilkan kalimat baru yang tetap mempertahankan makna dari teks tersebut. Teknik ini berguna ketika dataset yang tersedia tidak seimbang, karena augmentasi data dapat membantu model generalisasi lebih baik dan mengurangi overfitting. EDA terdiri dari empat teknik utama yaitu *Synonym Replacement* (SR), *Random Insertion* (RI), *Random Swap* (RS), dan *Random Deletion* (RD).

a. Synonym Replacement (SR)

Teknik ini membuat variasi kalimat dengan mengganti bebrapa kata dengan sinonimnya. Sinonim diperoleh dari kamus Tesaurus Bahasa Indonesia. Berikut adalah contoh penggunaan *Synonym Replacement.*

Tabel 3. 4 Contoh Synonym Replacement

Kalimat Asli	Kalimat Hasil SR
ahmad luthfi belum terpikir gabung partai golkar meski	ahmad luthfi belum berfikir gabung partai golkar meski

diusung di pilgub jateng	diusung di pilgub jateng

b. Random Insertion (RI)

Metode ini memasikkan kata yang sudah ada ke dalam kalimat dalam posisi yang tidak ditemukan. Dengan cara ini dapat meningkatkan variasi kalimat tanpa mengubah maknanya. Berikut adalah contoh penggunaan *Random Insertion.*

Tabel 3. 5 Contoh Random Insertion

Kalimat Asli	Kalimat Hasil RI
ahmad luthfi belum terpikir gabung partai golkar meski diusung di pilgub jateng	ahmad luthfi belum pernah terpikir gabung partai golkar meski diusung di pilgub jateng

c. Random Swap (RS)

Teknik ini memungkinkan dua kata dalam satu kalimat untuk mengubah posisinya secara acak. Ini berguna untuk membuat variasi dalam struktur kalimat. Berikut adalah contoh penggunaan *Random Swap*.

Tabel 3. 6 Contoh Random Swap

Kalimat Asli	Kalimat Hasil RI
ahmad luthfi belum terpikir gabung partai golkar meski diusung di pilgub jateng	meski belum diusung gabung partai golkar ahmad lutfi terpikir di pilgub jateng

d. Random Deletion (RD)

Taknik ini menghapus kata-kata tertentu dalam kalimat. Jika kalimat terlalu pendek penghapusan tidak dilakukan. Operasi ini membantu meningkatkan kekuatan model. Berikut adalah contoh *penggunaan Random Deletion.*

Tabel 3. 7 Contoh Random Deletion

Kalimat Asli	Kalimat Hasil RI
ahmad luthfi belum terpikir gabung partai golkar meski diusung di pilgub jateng	ahmad luthfi belum gabung partai golkar meski diusung di pilgub jateng

Hanya data berkelas minoritas yang akan diproses melalui proses EDA. Tujuannya adalah untuk menyeimbangkan distribusi kelas dalam dataset sehingga model klasidikasi tidak bias terhadap kelas mayoritas.

Diharapkan bahwa dengan meningkatnya jumlah sampel untuk kelas minoritas, model akan belajar lebih baik dan membuat prediksi yang lebih akurat untuk semua kelas.

Untuk setiap kalimat yang akan ditambahkan, satu dari empat teknik EDA akan dipilih secara acak dalam proses augmentasi. Tujuan dari pemilihan acak ini adalah untuk menciptakan variasi yang lebih beragam dalam data augmentasi, sehingga model dapat belajar dari berbagai bentuk kalimat dengan mempertahankan makna dan label kalimat aslinya.

Setiap kalimat kelas minoritas akan dibuat menjadi dua versi augmentasi dengan teknik EDA yang dipilih secara acak, dengan parameter *naug* diatur ke 2. Tujuan pemilihan nilai ini adalah untuk mengimbangi penambahan jumlah data dengan kualitas hasil augmentasi. Diharapkan kalimat yang dibuat tetap relevan dan tidak terlalu menyimpang dari makna aslinya dengan jumlah augmentasi yang tidak berlebihan. Selain itu, menetapkan *naug* = 2 juga membantu mencegah overfitting, yang dapat terjadi ketika kalimat augmentasi terlalu banyak dibuat.

5. Splitting Data

Setelah melewati proses *labelling* dan *pre-processing* data *splitting* atau pembagian dataset adalah proses penting untuk kinerja yang optimal pada sebuah model mencegah *overfitting* dan memvalidasi kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya (Wildan Amru Hidayat & Nastiti, 2024). Penelitian ini membagi data menjadi tiga proporsi, yaitu data *training* untuk pelatihan dan *fine-tuning* model, data *validation* digunakan untuk mengevaluasi kinerja model selama proses pelatihan dan mencegah *overfitting*, dan data testing untuk mengevaluasi kinerja akhir model setelah proses pelatihan selesai. Rasio data dibagi dengan 70% data *testing*, 15% sebagai data *validation*, dan 15% sebagai data *testing*.

6. Tokenizing

Untuk membangun model dengan menggunakan pre-trained IndoBERT yang dirancang khusus untuk pemrosesan bahasa Indonesia, diperlukan Teknik tokenisasi khusus. Tujuan dilakukan *tokenization* adalah text tersebut dapat di proses dan diolah oleh model IndoBERT.

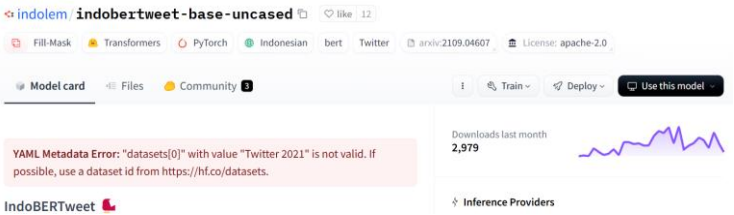
Tabel 3. 8 Ilustrasi Tokenizing IndoBERTweet

Text	ganjar pranowo tegas elektabilitas andika perkasa terus tingkat di jawa tengah karena kerja dari partai politik dan rawan rambo tuju jateng 1
Tokenized	['[CLS]', 'ganja', '##r', 'pran', '##owo', 'tegas', 'ele', '##kt', '##abilitas', 'andika', 'perkasa', 'terus', 'tingkat', 'di', 'jawa', 'tengah', 'karena', 'kerja', 'dari', 'partai', 'politik', 'dan', 'rawan', 'ramb', '##o', 'tuju', 'jateng', '1', '[SEP]']
Token ID	[2, 18006, 30359, 7979, 6378, 5127, 22483, 6661, 8234, 30268, 12917, 944, 1181, 26, 1069, 1172, 211, 494, 98, 2304, 1681, 41, 10588, 13378, 30370, 23641, 9183, 111, 3]

7. Implementasi Model IndoBERT

Dalam proses ini, metode pembuatan model akan melibatkan implementasi metode *transfer learning* pada model IndoBERT untuk melakukan analisis sentimen pada data hasil *crawling* dari media sosial Twitter. Model yang

akan digunakan dalam penelitian ini adalah *IndoBERTweet*. Model ini diimpor melalui repositori *indolem* yang tersedia di platform *HuggingFace* seperti pada gambar dibawah.



Gambar 3. 3 Repositori Model IndoBERT

Tahap pemodelan akan dilakukan menggunakan bahasa pemrograman *Python* dengan memanfaatkan *Google Colab* sebagai *notebook* untuk menjalankan program. Model *pre-trained* yang akan digunakan adalah *indobertweet-base-uncased*, dan akan diimpor langsung dari platform *HuggingFace* menggunakan API *Transformer*. Hal ini dapat dilakukan dengan menginstall dan mengimpor *library* tersebut menggunakan *Python*.

6.1 Grafik dan Diagram

Hasil dari proses model IndoBERT akan disajikan dalam bentuk grafik dan diagram visual. Grafik akan memvisualisasikan tingkat akurasi training dan validasi model selama proses pelatihan.

6.2 Matrik Evaluasi

Setelah melakukan *training*, kinerja model juga diukur menggunakan berbagai matrik evaluasi yaitu accuracy, precision, recall, dan f1-score. Untuk menganalisis kemampuan model dalam mengklasifikasikan sentimen secara lebih mendalam. Hal ini dilakukan agar mendapatkan penilaian yang komprehensif terhadap keakuratan dan performa IndoBERT dalam melaksanakan analisis sentimen.

8. Evaluasi Model

Pada tahap ini, model IndoBERT yang telah melalui proses pelatihan dan fine-tuning, akan diuji untuk memastikan kinerjanya dalam mengelompokkan sentimen teks. Evaluasi ini akan melibatkan teks acak yang tidak memiliki label. Tujuan dari pengujian ini adalah untuk menentukan kapasitas model dalam mengklasifikasikan teks.

BAB IV
HASIL DAN PEMBAHASAN

A. Crawling Data Twitter

Pada penelitian ini untuk mengumpulkan data adalah *Crawling* data dengan mengguanakan bahasa permrograman *Python* untuk mengambil data dari *Twitter*. Hasil *Crawling* disebut *dataset*, *dataset* yang diperoleh dari *crawling* masih berupa data mentah atau kotor karena masih terdapat atribut yang tidak relevan. Proses *crawling* dilakukan dengan menggunakan *Tweet Harvest* yang diambil dari *Twitter* dengan bantuan Google Colab. Data yang dikumpulkan berasal dari masyarakat Indonesia terkait para peserta calon gubernur dan wakil gubernur Jawa Tengah tahun 2024. Dalam mengumpulkan data tersebut peneliti menggunakan kata kunci "Andika Perkasa", "Hendrar Prihadi", "Ahmad Lutfi", "Taj Yasin" selama rentang waktu 24 Agustus hingga 31 Oktober 2024.

```
data = "andikaperkasa_Oktober.csv"
search_keyword = "andika perkasa lang:id until:2024-09-01 since:2024-08-01"
limit = 1000

!npx --yes tweet-harvest@2.6.1 -o "{data}" -s "{search_keyword}" -l {limit} --token '
```

Gambar 4. 1 Source Code Crawling Data

Sebelum melakukan proses *crawling* data dari *Twitter*, langkah pertama yang dilakukan adalah menginstall Node.js, karena *tweet harvest* berjalan di atas Node.js. Setelah Node.js

terinstall langkah berikutnya adalah menggunakan *tweet harvest* sebuah alat berbasis Node.js yang memungkinkan pengambilan data tweet berdasarkan kata kunci tertentu.

Dalam kode yang digunakan, variabel ‘data’ digunakan untuk menyimpan data hasil *crawling* yang akan disimpan dalam file .csv. ‘search_keyword’ digunakan untuk memasukkan kata kunci, bahasa dan rentang waktu yang akan dicari dan limit untuk membatasi jumlah maksimal tweet yang akan diambil.

Selanjutnya untuk proses *crawling* dilakukan dengan perintah !npx --yes tweet-harvest@2.6.1 -o "{data}" -s "{search_keyword}" -l {limit} --token "", yang menjalankan *tweet harvest* versi 2.6.1 tanpa perlu menginstalnya. Perintah ini akan menyimpan data dalam file .csv yang ditentukan, menggunakan kata kunci pencarian “andika perkasa" dengan rentang waktu yang sudah ditentukan, serta membatasi jumlah *tweet* sebesar “1000". Dari proses tersebut terkumpul data sebanyak 2.936 data, untuk pasangan calon 1 mendapat 1.492 *tweet* sedangkan pasangan calon 2 mendapat 1.444 *tweet*.

Data yang didapat merupakan *tweet* yang terdiri dari 15 atribut ‘*conversation_id_str’*, ‘*created_at’*, ‘*favorite_count*’, ‘*full_text*’, ‘*id_str*’, ‘*image_url*’, ‘*in_reply_to_screen_name*’, ‘*lang’*, ‘*location’*, ‘*quote_count*’, ‘*reply_count*’, ‘*retweet_count*’, ‘*tweet_url*’, ‘*user_id_str*’, ‘*username’*. Agar

data lebih efisien data yang digunakan hanya '*full_text*'. Berikut merupakan contoh data yang terkumpul.

Tabel 4. 1 Contoh Dataset yang Digunakan

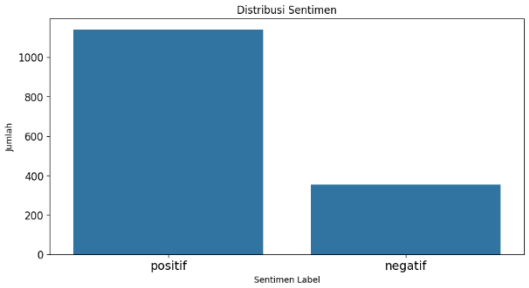
No	full_text
1	Pelaku seni Jateng pilih yg ga kaleng kaleng. Andika Perkasa-Hendi untuk #jatengperkasa Pilih Rambo bukan Sambo. https://t.co/wrQ8bgA8qA
2	Pagi tadi saya dan Mas Hendrar Prihadi bertemu dengan Ibu Juliana Richards Wakil Kepala Program Kerjasama Kedutaan Besar Inggris. Pertemuan ini membahas peluang kerjasama strategis di bidang pengentasan kemiskinan kesehatan dan pendidikan. Kedutaan Besar Inggris menyatakan https://t.co/Kl19v03Ddj
3	Memang benar kata pihak lawan pak Andika ini memang kelasnya calon presiden. Pengalamannya bkn hny nasional tapi juga internasional. Warga Jawa Tengah beruntung punya gubernur sekelas presiden. Calon Gubernur dan Wakil Gubernur Jawa Tengah Andika Perkasa dan Hendrar Prihadi https://t.co/Vv7IZmFYtB
.....	.....
1490	Gak sederhana tapi calon lain gak bisa . Pak Hendi kata gue lah. https://t.co/SQKaT7XM7K

1491	@jawafess biyen omahku ng smg utara banjire ngeriii udan sediluk banjir. bar p hendi dadi walikota ng smg utara wes gak ono banjir. contone wes biyen kota lama ki rak kopen bar di revitalisasi karo p hendi saiki dadi uapikk poll. pokoe 01
1492	Luar biasa nih Jateng Pokoknya Andika yg punya ya Jawa Tengah Yakin makin kuat dan perkasa karena ANDIKA - HENDI yg TERBAIK untuk JATENG https://t.co/8cYfp9oKkR,

B. Labelling Data

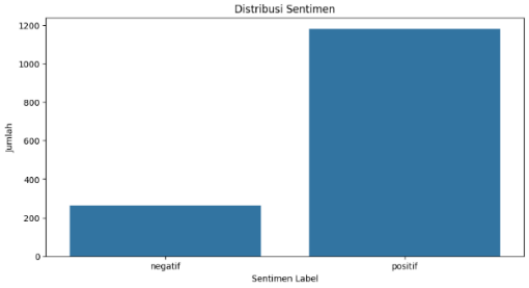
Setelah proses *crawling* data selesai, langkah selanjutnya adalah melakukan pelabelan data. Proses pelabelan dilakukan secara manual oleh seorang ahli bahasa, Ibu Febriana Wulan Suci, yang memiliki latar belakang akademik dalam bidang Sastra Indonesia dan saat ini berprofesi sebagai guru mata pelajaran Bahasa Indonesia.

Dari hasil *crawling*, total data yang dikumpulkan dengan total 2.936 tweet pasangan calon 1 mendapat 1.492 *tweet* sedangkan pasangan calon 2 mendapat 1.444 *tweet*. Setelah itu data akan dikategorikan menjadi 2 kategori yaitu positif, dan negatif.



Gambar 4. 2 Grafik Distribusi Label Paslon 1

Berdasarkan gambar diagram di atas menunjukkan distribusi sentimen publik terhadap pasangan calon 1 yaitu Andika Perkasa dan Hendrar Prihadi berdasarkan dari Twitter. Dari hasil pelabelan manual diperoleh sebanyak 1144 *tweet* positif, dan 347 *tweet* negatif untuk pasangan calon nomer urut 1.



Gambar 4. 3 Grafik Distribusi Label Paslon 2

Berdasarkan gambar diagram di atas menunjukkan distribusi sentimen publik terhadap pasangan calon 2 yaitu Ahmad Luthfi dan Taj Yasin berdasarkan dari Twitter. Dari

hasil pelabelan manual diperoleh sebanyak 1180 *tweet* positif, dan 263 *tweet* negatif untuk pasangan calon nomer urut 2.

Hasil Pelabelan menunjukkan bahwa mayoritas sentimen publik terhadap pasangan calon bersifat positif. Untuk pasangan calon 1, terdapat 1.144 *tweet* positif dan 347 *tweet* negatif. Sementara itu, pasangan calon 2 memperoleh 1.180 *tweet* positif dan 263 *tweet* negatif. Hal ini mengindikasikan bahwa kedua pasangan calon mendapatkan respons yang cenderung positif dari warganet di Twitter, dengan pasangan calon 2 memiliki proporsi sentimen positif yang sedikit lebih tinggi dibanding pasangan calon 1.

Dari grafik ini, terlihat adanya ketidakseimbangan distribusi kelas dalam dataset, yang ditandai dengan *tweet* sentimen positif yang jauh lebih tinggi dibandingkan dengan sentimen negatif. Ketidakseimbangan ini dapat berdampak pada analisis lebih lanjut, terutama dalam penerapan model pembelajaran mesin, karena model cenderung lebih terpengaruh oleh kelas yang dominan, dalam hal ini sentimen positif. Tetapi sebelum data masuk dalam proses EDA, harus memasuki tahap *pre-processing* dahulu untuk memastikan data bersih, terstruktur dan siap untuk diproses.

C. Pre-Processing Data

Langkah selanjutnya adalah *preprocessing* data, pada tahap ini terdiri dari 3 proses yang berurutan yaitu *cleaning text*,

case folding, *normalization*. Ketiga proses ini dilakukan untuk memastikan data yang akan diolah oleh model agar menjadi lebih terstruktur dan dan bersih dari *noise*. Berikut merupakan tahapan dari *preprocessing* yang dilakukan pada penelitian ini.

1. Cleaning Text

Tahap ini adalah tahap dataset dilakukan pembersihan dengan tujuan menjadikan dataset terstruktur agar dapat dibaca oleh sistem. Pada tahap ini peneliti menghapus atribut dataset yang tidak diperlukan.

Langkah pertama pada tahap *cleaning* data yaitu menghapus atribut data yang tidak diperlukan dalam dataset dan hanya menggunakan atribut *text*.

```
df['text'] = df['text'].astype(str)
df
```

Gambar 4. 4 Source Code Menghapus Atribut yang Tidak Perlu

Setelah dilakukan proses tersebut maka didapatkan hasil dataset yang hanya terdapat atribut text.

Setelah menghapus atribut pada dataset langkah selanjutnya adalah menghapus atribut pada teks seperti menghapus *URL*, menghapus *mention*, menghapus *hastag*, menghapus karakter spesial dan menghapus spasi yang berlebih.

```
def clean_text(text):
  if isinstance(text, str):
      # URLs
      text = re.sub(r'http\S+', '', text)
      # mentions
      text = re.sub(r'@\S+', '', text)
      # hashtags
      text = re.sub(r'#\S+', '', text)
      # special characters
      text = re.sub(r'[^\w\s]', '', text)
      # extra spaces
      text = re.sub(r'\s+', ' ', text).strip()

  else:
    # non-string values
    text = ''
  return text

df['cleaned_text'] = df['text'].apply(clean_text)
```

Gambar 4. 5 Source Code Proses Cleaned Text

Berdasarkan *source code* diatas fungsi '*clean_text*' digunakan untuk membersihkan teks dalam dataset. Fungsi ini mengecek apakah *input* berupa *string*. Jika benar teks akan masuk dalam 5 tahapan proses pembersihan, yaitu menghapus elemen yang tidak relevan seperti URL, *mention*, *hastag*, karakter khusus dan spasi yang berlebih. Setelah teks dibersihkan hasilnya disimpan dalam kolom baru *cleaned_text* dengan menerapkan fungsi pada setiap baris dalam kolom *text* menggunakan df['cleaned_text'] = df['text'].apply(clean_text). Berikut merupakan contoh hasil dari proses *cleaning text.*

Tabel 4. 2 Contoh Hasil Proses Cleaned Text

No	text	cleaned_text
1	Pelaku seni Jateng pilih yang ga kaleng. Andika Perkasa-Hendi untuk Jawa Tengah Pilih Rambo bukan Sambo.	Pelaku seni Jateng pilih yang ga kaleng Andika PerkasaHendi untuk Jawa Tengah Pilih Rambo bukan Sambo
2	Pagi tadi saya dan Mas Hendrar Prihadi bertemu dengan Ibu Juliana Richards Wakil Kepala Program Kerjasama Kedutaan Besar Inggris. Pertemuan ini membahas peluang kerjasama strategis di bidang pengentasan kemiskinan kesehatan dan pendidikan. Kedutaan Besar Inggris menyatakan mendukung	Pagi tadi saya dan Mas Hendrar Prihadi bertemu dengan Ibu Juliana Richards Wakil Kepala Program Kerjasama Kedutaan Besar Inggris Pertemuan ini membahas peluang kerjasama strategis di bidang pengentasan kemiskinan kesehatan dan pendidikan Kedutaan Besar Inggris menyatakan mendukung

	https://t.co/Kl19v03Ddj	
3	Memang benar kata pihak lawan pak Andika ini memang kelasnya calon presiden. Pengalamannya bukan hanya nasional tapi juga internasional. Warga Jawa Tengah beruntung punya gubernur sekelas presiden. Calon Gubernur dan Wakil Gubernur Jawa Tengah Andika Perkasa dan Hendrar Prihadi https://t.co/Vv7IZmFYtB	Memang benar kata pihak lawan pak Andika ini memang kelasnya calon presiden Pengalamannya bukan hanya nasional tapi juga internasional Warga Jawa Tengah beruntung punya gubernur sekelas presiden Calon Gubernur dan Wakil Gubernur Jawa Tengah Andika Perkasa dan Hendrar Prihadi
.....	.....	.....

1490	Gak sederhana tapi calon lain gak bisa . Pak Hendi kata gue lah. https://t.co/SQKaT7XM7KAdikrishna. #Blora #JatengPerkasa #JatengBerSATU https://t.co/PXW7yTKdp6	gak sederhana tapi calon lain gak bisa pak hendi kata gue lah
1491	@jawafess biyen omahku ng smg utara banjire ngeriii udan sediluk banjir. bar p hendi dadi walikota ng smg utara wes gak ono banjir. contone wes biyen kota lama ki rak kopen bar di revitalisasi karo p hendi saiki dadi uapikk poll. pokoe 01	biyen omahku ng smg utara banjire ngeriii udan sediluk banjir bar p hendi dadi walikota ng smg utara wes gak ono banjir contone wes biyen kota lama ki rak kopen bar di revitalisasi karo p hendi saiki dadi uapikk poll pokoe
1492	Luar biasa nih Jateng Pokoknya Andika yg punya ya Jawa Tengah	luar biasa nih jateng pokoknya andika yg punya ya jawa tengah

	Yakin makin kuat dan perkasa karena ANDIKA - HENDI yg TERBAIK untuk JATENG https://t.co/8cYfp9oKkR,	yakin makin kuat dan perkasa karena andika hendi yg terbaik untuk jateng

2. Case Folding

Langkah selanjutnya adalah *case folding*, yaitu proses mengubah seluruh teks menjadi huruf kecil yang bertujuan untuk menyamakan format teks sehingga kata-kata dengan huruf besar dan huruf kecil dianggap sama. Berikut *source code* yang digunakan untuk melakukan proses *case folding*.

```
df['case_folding'] = df['cleaned_text'].str.lower()
df
```

Gambar 4. 6 Source Code Case Folding

Berdasarkan *source code* diatas menunjukkan bahwa atribut *cleaned_text* yang digunakan dalam proses ini. Proses ini dilakukan dengan menerapkan fungsi .str.lower() pada kolom *cleaned_text*, kemudian hasilnya disimpan dalam kolom baru bernama *case_folding*.

Tabel 4. 3 Contoh Hasil Proses Case Folding

No	cleaned_text	case_folding
1	Pelaku seni Jateng pilih yang ga kaleng Andika PerkasaHendi untuk Jawa Tengah Pilih Rambo bukan Sambo	pelaku seni jateng pilih yang ga kaleng andika perkasahendi untuk jawa tengah pilih rambo bukan sambo
2	Pagi tadi saya dan Mas Hendrar Prihadi bertemu dengan Ibu Juliana Richards Wakil Kepala Program Kerjasama Kedutaan Besar Inggris Pertemuan ini membahas peluang kerjasama strategis di bidang pengentasan kemiskinan kesehatan dan pendidikan Kedutaan Besar Inggris menyatakan mendukung	pagi tadi saya dan mas hendrar prihadi bertemu dengan ibu juliana richards wakil kepala program kerjasama kedutaan besar inggris pertemuan ini membahas peluang kerjasama strategis di bidang pengentasan kemiskinan kesehatan dan pendidikan kedutaan besar inggris

		menyatakan mendukung
3	Memang benar kata pihak lawan pak Andika ini memang kelasnya calon presiden Pengalamannya bukan hanya nasional tapi juga internasional Warga Jawa Tengah beruntung punya gubernur sekelas presiden Calon Gubernur dan Wakil Gubernur Jawa Tengah Andika Perkasa dan Hendrar Prihadi	memang benar kata pihak lawan pak andika ini memang kelasnya calon presiden pengalamannya bukan hanya nasional tapi juga internasional warga jawa tengah beruntung punya gubernur sekelas presiden calon gubernur dan wakil gubernur jawa tengah andika perkasa dan hendrar prihadi
.....	....	....
1490	gak sederhana tapi calon lain gak bisa pak hendi kata gue lah	gak sederhana tapi calon lain gak bisa

		pak hendi kata gue lah
1491	biyen omahku ng smg utara banjire ngeriii udan sediluk banjir bar p hendi dadi walikota ng smg utara wes gak ono banjir contone wes biyen kota lama ki rak kopen bar di revitalisasi karo p hendi saiki dadi uapikk poll pokoe	biyen omahku ng smg utara banjire ngeriii udan sediluk banjir bar p hendi dadi walikota ng smg utara wes gak ono banjir contone wes biyen kota lama ki rak kopen bar di revitalisasi karo p hendi saiki dadi uapikk poll pokoe
1492	luar biasa nih jateng pokoknya andika yg punya ya jawa tengah yakin makin kuat dan perkasa karena andika hendi yg terbaik untuk jateng	luar biasa nih jateng pokoknya andika yg punya ya jawa tengah yakin makin kuat dan perkasa karena andika hendi yg terbaik untuk jateng

3. Normalization

Proses selanjutnya adalah normalisasi atau *normalize*. Proses ini untuk mengubah teks menjadi bentuk yang lebih

standar atau konsisten. Tujuan utamanya adalah mengubah teks dengan penulisan kata yang disingkat atau kata tidak baku menjadi baku. Sumber kamus yang digunakan untuk melakukan normalisasi tersedia secara publik di *GitHub*, yaitu *colloquial-indonesian-lexicon.csv* terdiri dari 15.397 baris kata, *kbba.txt* terdiri dari 1.319 baris kata, *slangword.txt* terdiri dari 691 baris kata dan *kamus_slang.csv* terdiri dari 15.007 baris kata. Dengan menggunakan sumber ini, diharapkan penelitian ini dapat memastikan bahwa teks yang dianalisis telah memenuhi standar bahasa baku, sehingga hasil analisis menjadi lebih akurat dan relevan. Berikut merupakan contoh isi dari kamus kamus yang digunakan.

Tabel 4. 4 Contoh isi 'colloquial-indonesian-lexicon.csv'

No	Slang	Formal
1	woww	wow
2	netaas	menetas
.....	.....	.....
15397	gaharus	enggak harus

Tabel 4. 5 Contoh isi 'kbba.txt'

No	Slang	Formal
1	7an	tujuan

2	abis	habis
.....	.....	.....
1319	mgkin	mungkin

Tabel 4. 6 Contoh isi 'slangword.txt'

No	Slang	Formal
1	alay	aneh
2	ancur	hancur
.....	.....	.....
691	akherat	akhirat

Tabel 4. 7 Contoh isi 'kamus_slang.csv'

No	Slang	Formal
1	met	selamat
2	aminn	amin
.....	.....	.....
15007	fans2	Fan-fan

Berikut source code yang digunakan untuk melakukan proses normalisasi.

```
# Load kamus alay
kamus_alay = pd.read_csv("https://raw.githubusercontent.com/nasalsabila/kamus-alay/master/colloquial-indonesian-lexicon.csv")
alay_dict = pd.Series(kamus_alay['formal'].values, index=kamus_alay['slang']).to_dict()

# Load kamus kbba.txt
kamus_kbba = pd.read_csv('kbba.txt', sep='\t')
kbba_dict = pd.Series(kamus_kbba['tujuan'].values, index=kamus_kbba['?an']).to_dict()

# Load kamus slang.csv
kamus_slang = pd.read_csv('kamus_slang.csv')
slang_dict = pd.Series(kamus_slang['formal'].values, index=kamus_slang['slang']).to_dict()

# Load kamus slangword.txt
kamus_sw = pd.read_csv('slangword.txt', sep=':', engine='python', header=None, names=['slang', 'formal'])
sw_dict = pd.Series(kamus_sw['formal'].values, index=kamus_sw['slang']).to_dict()

# Gabungkan semua kamus
all_dicts = [alay_dict, kbba_dict, slang_dict, sw_dict]
combined_dict = {}
for d in all_dicts:
    combined_dict.update(d)
```

Gambar 4. 7 Source Code Kamus Alay

```
# Fungsi untuk normalisasi teks
def normalize_text(text):
    normalized_text = []
    for word in text.split():
        if word in combined_dict:
            normalized_text.append(combined_dict[word])
        else:
            normalized_text.append(word)
    return ' '.join(normalized_text)

df['normalized_text'] = df['case_folding'].apply(normalize_text)
df
def normalize_text(text):
    normalized_text = []
    normalized_count = 0  # Initialize a counter for normalized words
    for word in text.split():
        if word in combined_dict:
            normalized_text.append(combined_dict[word])
            normalized_count += 1  # Increment the counter if word is normalized
        else:
            normalized_text.append(word)
    return ' '.join(normalized_text), normalized_count  # Return both normalized text and count

df['normalized_text'], df['normalized_count'] = zip(*df['case_folding'].apply(normalize_text))
df
```

Gambar 4. 8 Souce Code Spell Normalized

Souce code di atas berfungsi untuk melakukan normalisasi teks, yaitu mengonversi kata-kata dalam teks menjadi bentuk standarnya. Hal pertama yang dilakukan adalah pembacaan file kamus-kamus yang diperlukan seperti *colloquial-indonesian-lexicon.csv*, *kbba.txt, slangword.txt*, dan *kamus_slang.csv*. Kemudian semua kamus digabung menjadi satu *dictionary* (combined_dict). Selanjutnya membuat fungsi normalize_text(text) jika kata ditemukan dalam kamus *combined_dict* maka kata tersebut

akan diganti dengan bentuk normalnya. Jika tidak, kata akan tetap dipertahankan. Hasil normalisasi kemudian diterapkan ke *DataFrame* menggunakan kolom *case_folding* dan disimpan dalam kolom baru bernama *normalized_text*. Berikut merupakan contoh hasil dari proses *normalized_text*.

Tabel 4. 8 Contoh Hasil Spell Normalized

No	case_folding	normalized_text
1	pelaku seni jateng pilih yang ga kaleng andika perkasahendi untuk jawa tengah pilih rambo bukan sambo	pelaku seni jateng pilih yang enggak kaleng andika perkasahendi untuk jawa tengah pilih rambo bukan sambo
2	pagi tadi saya dan mas hendrar prihadi bertemu dengan ibu juliana richards wakil kepala program kerjasama kedutaan besar inggris pertemuan ini membahas peluang kerjasama strategis di bidang pengentasan kemiskinan kesehatan	pagi tadi saya dan mas hendrar prihadi bertemu dengan ibu juliana richards wakil kepala program kerjasama kedutaan besar inggris pertemuan ini membahas peluang kerjasama strategis di bidang pengentasan

	dan pendidikan kedutaan besar inggris menyatakan mendukung	kemiskinan kesehatan dan pendidikan kedutaan besar inggris menyatakan mendukung
3	memang benar kata pihak lawan pak andika ini memang kelasnya calon presiden pengalamannya bukan hanya nasional tapi juga internasional warga jawa tengah beruntung punya gubernur sekelas presiden calon gubernur dan wakil gubernur jawa tengah andika perkasa dan hendrar prihadi	memang benar kata pihak lawan pak andika ini memang kelasnya calon presiden pengalamannya bukan hanya nasional tapi juga internasional warga jawa tengah beruntung punya gubernur sekelas presiden calon gubernur dan wakil gubernur jawa tengah andika perkasa dan hendrar prihadi
.....	....	....

1490	gak sederhana tapi calon lain gak bisa pak hendi kata gue lah	enggak sederhana tapi calon lain enggak bisa pak hendi kata saya lah
1491	biyen omahku ng smg utara banjire ngeriii udan sediluk banjir bar p hendi dadi walikota ng smg utara wes gak ono banjir contone wes biyen kota lama ki rak kopen bar di revitalisasi karo p hendi saiki dadi uapikk poll pokoe	biyen omahku ng semoga utara banjire ngeriii udan sediluk banjir bar p hendi dadi walikota ng semoga utara wes enggak sono banjir contone wes biyen kota lama ki rak kopen bar di revitalisasi karo p hendi saiki dadi uapikk poll pokoe
1492	luar biasa nih jateng pokoknya andika yg punya ya jawa tengah yakin makin kuat dan perkasa karena andika hendi yg terbaik untuk jateng	luar biasa ini jateng pokoknya andika yang punya iya jawa tengah yakin makin kuat dan perkasa karena andika hendi yang terbaik untuk jateng

D. Easy Data Augmentation (EDA)

Pada penelitian ini, Teknik *Easy Data Augmentation* (EDA) diterapkan untuk meningkatkan variasi dan kuantitas data teks yang digunakan dalam pelatihan model. EDA merupakan teknik augmentasi sederhana yang melibatkan empat operasi dasar, yaitu *Synonym Replacement*, *Random Insertion*, *Random Swap*, dan *Random Deletion*.

Implementasi EDA dimulai dengan membaca dataset dari file excel yang berisi kolom normalized_text. Hasil augmentasi dari setiap teknik dievaluasi berdasarkan jumlah perubahan kata yang terjadi. Kriteria ini digunakan untuk mengukur seberapa signifikan modifikasi yang dilakukan terhadap kalimat asli. Hasil augmentasi yang memiliki perubahan kata terbanyak dipilih sebagai hasil terbaik, karena dianggap memberikan variasi yang lebig besar tanpa mengorbankan makna kalimat. Berikut source code dari keempat teknik EDA yang digunakan dalam penelitian ini.

```python
# 1. Synonym Replacement (SR)
def synonym_replacement(text, alpha=alpha):
    words = text.split()
    n = max(1, int(alpha * len(words)))
    indices = random.sample(range(len(words)), n)
    for i in indices:
        words[i] = get_synonym(words[i])
    return " ".join(words)

# 2. Random Insertion (RI)
def random_insertion(text, alpha=alpha):
    words = text.split()
    n = max(1, int(alpha * len(words)))
    for _ in range(n):
        word = random.choice(words)
        synonym = get_synonym(word)
        position = random.randint(0, len(words))
        words.insert(position, synonym)
    return " ".join(words)

# 3. Random Swap (RS)
def random_swap(text, alpha=alpha):
    words = text.split()
    n = max(1, int(alpha * len(words)))
    for _ in range(n):
        if len(words) < 2:
            break
        idx1, idx2 = random.sample(range(len(words)), 2)
        words[idx1], words[idx2] = words[idx2], words[idx1]
    return " ".join(words)

# 4. Random Deletion (RD)
def random_deletion(text, alpha=alpha):
    words = text.split()
    if len(words) == 1:
        return text
    words = [word for word in words if random.random() > alpha]
    return " ".join(words) if words else random.choice(text.split())
```

Gambar 4. 9 Source Code Empat Teknik EDA

Metode pertama, *Synonym Replacement* (SR) yaitu menganti kata dengan sinonimnya. Tesaurus Bahasa Indonesia yang dibuat oleh Pusat Bahasa Departmen Pendidikan Nasional pada tahun 2008, memiliki kata sinonim dalam format json. *Random Insertion* (RI) adalah metode kedua yang secara acak dimasukkan kata sinonim dari salah satu kata yang ada dalam data ke dalam kalimat. *Random Swap* (RS) adalah metode yang secara acak mengubah urutan kata dalam suatu kalimat. *Random Deletion* (RD) adalah metode yang secara acak menghapus suatu kalimat.

Untuk menerapkan augmentasi pada teks, masing-masing dari keempat metode memerlukan parameter α. Dalam teknik *Synonym Replacement* (SR), *Random Insertion* (RI), dan

Random Swap (RS), parameter α digunakan untuk menentukan proporsi kata yang diubah. Di sisi lain, dalam *Random Deletion* (RD) parameter α digunakan untuk menentukan kemungkinan setiap kata dalam kalimat akan dihilangkan. Sesuai dengan saran (Wei & Zou, 2019), nilai α dalam penelitian ini ditetapkan sebesar 0,1. *Source code* yang digunakan untuk inisialisasi parameter sebagai berikut:

```
# Inisialisasi parameter
alpha = 0.1
n_aug = 2
```

Gambar 4. 10 Inisialisasi Parameter

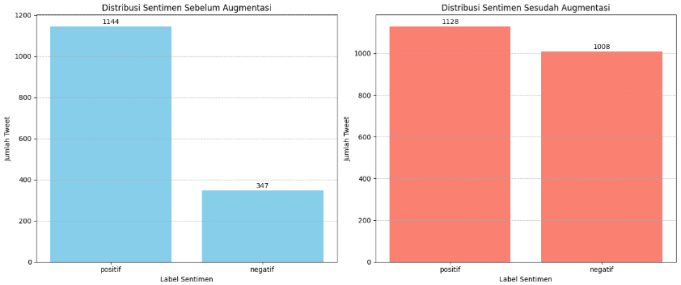
Dalam penelitian ini, nilai parameter naug ditetapkan sebesar 2. Naug adalah parameter tambahan yang digunakan untuk menentukan banyak kalimat baru yang dibuat dari satu data asli. Oleh karena itu, setiap data yang melewati proses EDA akan diubah menjadi dua versi peningkatan. Nilai-nilai ini diterapkan berdaarkan kebutuhan dataset untuk menghindari *overfitting*, yang apat terjadi jika terlalu banyak kalimat peningkatan dibuat. Berikut adalah hasil proses dari EDA.

Tabel 4. 9 Contoh Hasil Proses EDA

normalized_text	Augmentasi 1	Augmentasi 2
andika perkasa berpeluang tantang ahmad	andika perkasa ahmad lutfi berpeluang	andika perkasa tantang ahmad lutfi di pilgub jateng

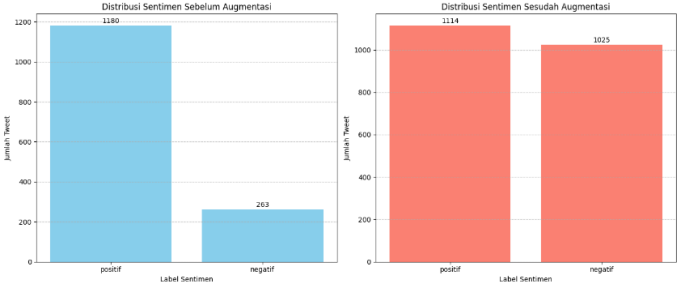
luthfi di pilgub jateng	tantang di pilgub jateng	
adik gus dur jadi ketua tim pemenangan andika perkasa hendrar prihadi	jadi ketua tim pemenangan andika perkasa hendrar prihadi adik gus dur	adik gus dur tim pemenangan andika perkasa hendrar prihadi
gerindra usung ahmad lutfi di pilgub jateng	di pilgub jateng gerindra usung ahmad lutfi	gerindra dukung ahmad lutfi di pilgub jateng
ahmad lutfi sosok pemimpin yang selalu ngopeni nglakoni	sosok pemimpin yang selalu ngopeni ngelakoni ahmad lutfi	ahmad lutfi sosok pemimpin yang ngopeni nglakoni

Setelah proses ini didapati jumlah data meningkat untuk pasangan calon gubernur dan wakil gubernur nomer urut 1 Andika Perkasa dan Hendrar Prihadi dari 1.492 data menjadi 2.137 data dengan distribusi data sebelum dilakukan EDA yaitu 1.144 data berlabel positif, dan 347 data berlabel negatif sedangkan setelah dilakukan EDA distribusi label bertambah menjadi 1.128 data berlabel positif, dan 1.008 data berlabel negatif.



Gambar 4. 11 Perbandingan Distribusi Label Sebelum dan Sesudah Proses EDA pada Paslon 1

Untuk pasangan calon gubernur dan wakil gubernur nomer urut 2 Ahmad Lutfi dan Taj Yasin dari 1.443 data menjadi 2.139 data dengan distribusi data sebelum dilakukan EDA yaitu 1180 data berlabel positif, dan 263 data berlabel negatif. Setelah dilakukan EDA distribusi data bertambah menjadi 1114 data berlabel positif, dan 1025 data berlabel negatif.



Gambar 4. 12 Perbandingan Distribusi Label Sebelum dan Sesudah Proses EDA pada Paslon 2

E. Splitting Data

Proses *splitting* data adalah langkah penting dalam *machine learning* untuk membagi dataset menjadi tiga bagian yaitu data latih (*training set*) sebesar 70% , data validasi (*validation set*) sebesar 15%, dan data uji (*test set*) sebesar 15%. Data yang digunakan untuk splitting data adalah kolom *cleaned_text* dan label. Proses pembagian data dilakukan secara acak (*random sampling*) untuk memastikan distribusi data pada masing-masing kelompok tetap mewakili setiap kata-kata. Berikut *source code* yang digunakan untuk melakukan *splitting* data.

```
# Membagi dataset menjadi train, validation, dan test
train_df, temp_df = train_test_split(df, test_size=0.3, stratify=df['sentimen'], random_state=42)
val_df, test_df = train_test_split(temp_df, test_size=0.5, stratify=temp_df['sentimen'], random_state=42)

print(f"\nTrain set size: {len(train_df)}")
print(f"Validation set size: {len(val_df)}")
print(f"Test set size: {len(test_df)}")
```

Gambar 4. 13 source Code Splitting Data

Source code di atas bertujuan untuk membagi dataset menjadi tiga bagian: *training set*, *validation set*, dan *test set* dengan menggunakan stratified sampling agar distribusi kelas pada kolom ‘sentimen’ tetap seimbang pada setiap subset.

Pertama, df.dropna(subset=['sentimen'], inplace=True) kode ini menghapus semua baris dalam dataset yang memiliki nilai NaN (kosong) pada kolom ‘sentimen’. Dengan begitu, hanya data yang memiliki label sentimen yang akan digunakan. Dataset dibagi menjadi *training set* (70%) dan *validation* + *test*

set (30%) menggunakan fungsi train_test_split(). Parameter test_size=0.3 menentukan bahwa 30% data akan digunakan untuk validasi dan pengujian, sementara 70% lainnya menjadi data latih. Parameter *random_state=42* untuk memastikan hasil pembagian data konsisten.

Selanjutnya, subset validasi + test (30% dari data asli) kembali dibagi menjadi *validation set* (67%) dan *test set* (33%). Parameter test_size=0.3 berarti bahwa sepertiga dari subset ini sekitar (15% dari total dataset awal) digunakan sebagai test set, sementara 15% dari total dataset awal digunakan untuk validasi.

Langkah terakhir yaitu mencetak jumlah data dalam *training set*, *validation set*, dan *test set* sehingga dapat mengecek pembagian data. Pada proses pembagian data tersebut paslon 1 memperoleh data uji sebanyak 321 data, data validasi sebanyak 320 data, data latih sebanyak 1.495 data. Sedangkan paslon 2 memperoleh data uji 321 data, data validasi 321 data, data latih 1497 data.

F. Tokenizing

Sebelum dataset digunakan dataset harus sesuai dengan format input yang diterima oleh IndoBERT. Setelah didapatkan dataset yang sudah diolah melalui proses *pre-processing*, tahapan selanjutnya menambahkan token khusus yaitu [CLS] di awal kalimat, dan [SEP] di akhir kalimat yang berfungsi sebagai

pemisah antar kalimat. Setelah itu akan dilakukan proses *encoding* yang dilakukan menggunakan *tokenizer* yang sesuai dengan indeks *vocabulary* yang telah didefinisikan IndoBERT yang telah dilatih sebelumnya.

Untuk tokenisasi, penelitian ini menggunakan BERT Tokenizer dari model *indobertweet-base-uncased* yang telah dilatih sebelumnya. Proses ini menghasilkan representasi teks dalam bentuk token numerik yang sesuai dengan kosakata model. Teknik tokenisasi *WordPiece* memecah teks menjadi bagian kecil.

Setelah proses tokenisasi, setiap token diubah menjadi indeks numerik yang disesuaikan dengan kosakata model, yang terdiri dari 32.000 token. Dua vektor utama kemudian dibentuk menggunakan representasi numerik ini. Yang pertama adalah *Input Ids* yang digunakan untuk menampilkan token teks, termasuk token khusus [CLS] dan [SEP]; yang kedua adalah *attention mask*, yang merupakan vektor biner yang menunjukkan token mana yang harus diperhatikan oleh model (1 untuk token yang relevan, 0 untuk *padding*).

Model BERT dapat menerima input teks dengan panjang maksimal 512 kata. Jika kalimat terlalu pendek, maka akan ditambahkan token [PAD] untuk memenuhi panjang maksimal. Sebaliknya, jika kalimat terlalu panjang maka akan dipangkas menjadi beberapa kata (*truncate*). Kata yang tidak

terdapat dalam *vocabulary* akan menjadi sub kata (*subword*) dengan menggunakan simbol ##. Berikut *source code* pada proses tokenizing.

```python
# Membuat custom dataset dengan nama kolom yang benar
class SentimentDataset(Dataset):
    def __init__(self, dataframe, tokenizer, max_len=128, text_column='normalized_text', label_column='sentimen'):
        self.data = dataframe
        self.tokenizer = tokenizer
        self.max_len = max_len
        self.text_column = text_column
        self.label_column = label_column

        # Pastikan kolom yang diperlukan ada
        if self.text_column not in self.data.columns:
            raise ValueError(f"Kolom '{self.text_column}' tidak ditemukan. Kolom yang tersedia: {self.data.columns.tolist()}")
        if self.label_column not in self.data.columns:
            raise ValueError(f"Kolom '{self.label_column}' tidak ditemukan. Kolom yang tersedia: {self.data.columns.tolist()}")

    def __len__(self):
        return len(self.data)

    def __getitem__(self, index):
        text = self.data.iloc[index][self.text_column]
        label = self.data.iloc[index][self.label_column]

        # Ensure text is a string
        if not isinstance(text, str):
            text = str(text)

        encoding = self.tokenizer(
            text,
            add_special_tokens=True,
            max_length=self.max_len,
            padding='max_length',
            truncation=True,
            return_attention_mask=True,
            return_tensors='pt'
        )
```

```python
# Membaca file TSV dan membuat dataset
import pandas as pd

def read_tsv(file_path):
    df = pd.read_csv(file_path, sep='\t')
    # Mengubah label sentimen menjadi numerik
    df['sentimen'] = df['sentimen'].map({'negatif': 0, 'positif': 1})
    return df

# Membaca file TSV dan membuat dataset
train_data = read_tsv('train.tsv')
val_data = read_tsv('val.tsv')
test_data = read_tsv('test.tsv')

# Membuat dataset menggunakan custom class dengan nama kolom yang benar
train_dataset = SentimentDataset(train_data, tokenizer, text_column='normalized_text', label_column='sentimen')
val_dataset = SentimentDataset(val_data, tokenizer, text_column='normalized_text', label_column='sentimen')
test_dataset = SentimentDataset(test_data, tokenizer, text_column='normalized_text', label_column='sentimen')
```

Gambar 4. 14 Source Code Proses Tokenisasi

Source code di atas digunakan untuk memuat dataset sentimen, mengolahnya dengan *tokenizer* serta menyiapkan data *loader* yang akan digunakan dalam proses pelatihan model. Pertama, membuat sebuah kelas SentimentDataset yang

merupakan turunan dari torch.utils.data.Dataset yang bertujuan untuk menangani data teks dan label secara terstruktur.

Di dalam kelas ini, fungsi __init__ akan menyimpan data, tokenizer, dan parameter panjang maksimum token, sertta memastikan kolom teks dan label ada di *dataframe*. Fungsi __len__ digunakan untuk mengetahui jumlah total data, sedangkan __getitem__ bertugas mengambil satu data berdasrkan indeks, melakukan tokenisasi terhadap teks dengan tokenizer IndoBERT dan mengembalikannya dalam bentuk tensor untuk digunakan oleh model.

Selanjutnya, terdapat fungsi read_tsv yang digunakan untuk membaca file berformat .tsv (*tab-separated values*) dan mengubah label sentimen dari bentuk teks seperti ”positif” dan “negatif” menjadi angka 1 dan 0, sehingga dapat diproses oleh model. Setelah itu, data dari file train.tsv, val.tsv, dan test.tsv dibaca menggunakan fungsi tersebut. Terakhir, setiap *dataframe* tersebut diubah menjadi dataset Pytorch menggunakan kelas SentimenDataset, agar digunakan dalam pelatihan, validasi, dan pengujian model. Berikut merupakan contoh hasil dari tokenizing.

Tabel 4. 10 Contoh Hasil Proses Tokenisasi

Original	memang benar kata pihak lawan pak andika ini memang kelas calon presiden alam bukan hanya nasional tapi juga

	internasional warga jawa tengah untung punya gubernur kelas presiden calon gubernur dan wakil gubernur jawa tengah andika perkasa dan hendrar prihadi
Tokenized	['[CLS]', 'memang', 'benar', 'kata', 'pihak', 'lawan', 'pak', 'andika', 'ini', 'memang', 'kelas', 'calon', 'presiden', 'alam', 'bukan', 'hanya', 'nasional', 'tapi', 'juga', 'internasional', 'warga', 'jawa', 'tengah', 'untung', 'punya', 'gubernur', 'kelas', 'presiden', 'calon', 'gubernur', 'dan', 'wakil', 'gubernur', 'jawa', 'tengah', 'andika', 'perkasa', 'dan', 'hendra', '##r', 'pri', '##had', '##i', '[SEP]']
Token IDs	[2, 731, 839, 661, 1241, 4031, 556, 30268, 92, 731, 1194, 1874, 1871, 668, 531, 344, 1287, 469, 186, 2506, 1470, 1069, 1172, 5133, 1121, 3256, 1194, 1871, 1874, 3256, 41, 3022, 3256, 1069, 1172, 30268, 12917, 41, 18326, 30359, 4538, 480, 30356, 3]
Attention Mask	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

G. Implementasi Model IndoBERT

Setelah proses tokenisasi selesai, data kemudian dimasukkan ke dalam jaringan IndoBERT. Proses *fine-tuning* IndoBERT untuk tugas klasifikasi melibatkan penambahan lapisan klasifikasi pada model. Pustaka *Tranformers* menyediakan kelas *BertForSequenceClassification*, yang dimaksudkan untuk menyelesaikan tugas klasifikasi teks. Nilai *logits* dihitung oleh model dengan keluaran dari *pooler*, yang kemudian diproses menggunakan fungsi *softmax* untuk menghasilkan prediksi akhir. Penulis menggunakan model indobertweet-base-uncased untuk *fine-tuning* penelitian. *Source code* untuk proses *fine-tuning* dalam fase pelatihan dan validasi sebagai berikut:

```
from torch.utils.data import DataLoader

# Membuat DataLoader
batch_size = 32
train_dataloader = DataLoader(train_dataset, batch_size=batch_size, shuffle=True)
val_dataloader = DataLoader(val_dataset, batch_size=batch_size)
test_dataloader = DataLoader(test_dataset, batch_size=batch_size)

# Menyiapkan model untuk fine-tuning
model = AutoModelForSequenceClassification.from_pretrained(model_name, config=config)
model = model.to(device)

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

# Mendapatkan nilai learning rate
num_epochs = 4
total_steps = len(train_dataloader) * num_epochs
scheduler = get_linear_schedule_with_warmup(
    optimizer,
    num_warmup_steps=0,
    num_training_steps=total_steps
)
```

Gambar 4. 15 Souce Code Membuat DataLoader

```python
# Training function
from sklearn.metrics import accuracy_score

# Training loop
train_losses = []
train_accuracies = []
val_losses = []
val_accuracies = []

print(f"\nMulai training untuk {num_epochs} epochs...")

for epoch in range(num_epochs):
    print(f"\nEpoch {epoch+1}/{num_epochs}")

    # Training
    train_loss, train_acc = train_epoch(model, train_dataloader, optimizer, scheduler, device)
    train_losses.append(train_loss)
    train_accuracies.append(train_acc)

    # Validation
    val_loss, val_acc, val_report = evaluate(model, val_dataloader, device)
    val_losses.append(val_loss)
    val_accuracies.append(val_acc)

    print(f"Train Loss: {train_loss:.4f}, Train Accuracy: {train_acc:.4f}")
    print(f"Val Loss: {val_loss:.4f}, Val Accuracy: {val_acc:.4f}")
    print("Validation Classification Report:")
    print(val_report)
```

Gambar 4. 15 Source Code Fase Pelatihan

Dalam proses *fine-tuning* untuk pelatihan dan validasi, membutuhkan beberapa *hyperparameter* seperti *batch size*, *learning rate*, dan *epoch*. Berdasarkan referensi dari penelitian Devlin 2020, *hyperparameter* yang direkomendasikan untuk *fine-tuning* agar mencapai performa optimal mencakup *batch size* sebesar 16 atau 32, serta *learning rate* dalam rentang 5e-5, dan 3e-5. Selain itu model dilatih menggunakan *optimizer Adam* dengan jumlah *epoch* sebanyak 2, 3, atau 4. Namun dalam penelitian ini, jumlah *epoch* ditetapkan 3 dan 4, *batch size* 32 dan *learning rate* 5e-5 dan 3e-5 untuk memperoleh gambaran yang luas terkait proses iterasi selama pelatihan. Skenario kombinasi epoch, dan *learning rate* untuk paslon 1 dan paslon 2 dapat dilihat pada tabel berikut.

Tabel 4. 11 Skenario Hyperparameter

Skenario	Epoch	Learning Rate
Skenario 1	3	3e-5
Skenario 2	3	5e-5
Skenario 3	4	3e-5
Skenario 4	4	5e-5

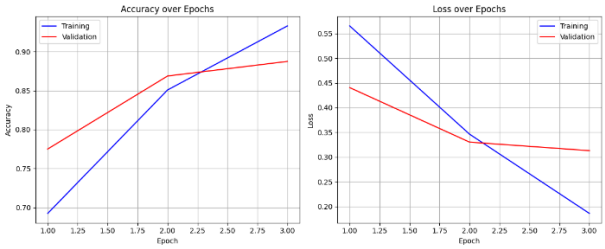
Dari skenario-skenario yang telah dilakukan untuk menguji paslon 1 didapatkan hasil sebagai berikut.

Tabel 4. 12 Hasil Skenario Hyperparameter Paslon 1

Skenario	Data	Accuracy	Precision	Recall	F1 score
Skenario 1	Training	0,93	0,90	0,84	0,87
	Validation	0,90	0,87	0,92	0,89
Skenario 2	Training	0,93	0,89	0,84	0,87
	Validation	0,88	0,87	0,91	0,89
Skenario 3	Training	0,94	0,92	0,78	0,95
	Validation	0,86	0,83	0,94	0,88
Skenario 4	Training	0,98	0,91	0,89	0,90
	Validation	0,90	0,90	0,92	0,91

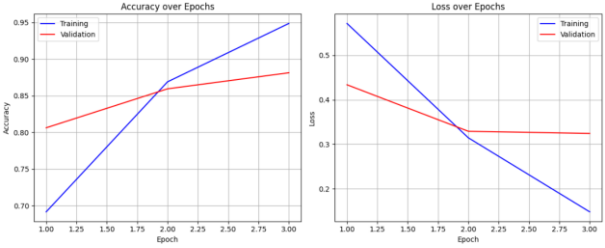
Berdasarkan hasil tabel skenario diatas, berdasarkan grafik skenario 1 tidak menunjukkan *overfitting*. Akurasi pada data validasi terus meningkat seiring dengan data latih. *Loss* pada data validasi juga terus menurun, meskipun sedikit lebih lambat dari data latih. Ini menandakan model masih belajar dengan baik dan mampu melakukan generalisasi ke data baru.

Berikut grafik akurasi *training* dan *validation* serta *training loss* dan *validation loss* pada skenario 1.



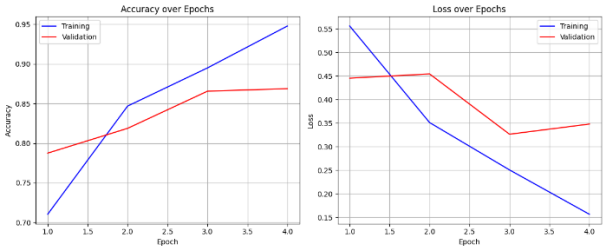
Gambar 4. 16 Grafik Akurasi Training dan Validation serta Grafik Training Loss dan Validation Loss Paslon 1 Skenario 1

Pada skenario 2, performanya masih mirip dengan skenario 1 tetapi terdapat sedikit penurunan pada *accuracy* dan *precision* di data *validation*. Hal ini menunjukkan bahwa penggunaan *learning rate* yang lebih besar dapat mempengaruhi kemampuan model, sehingga generalisasi model sedikit menurun dibandingkan dengan *learning rate* yang lebih kecil. Berikut grafik akurasi *training* dan *validation* serta *training loss* dan *validation loss* pada skenario 2.



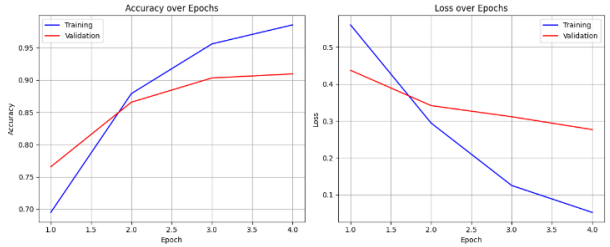
Gambar 4. 17 Grafik Akurasi Training dan Validation serta Grafik Training Loss dan Validation Loss Paslon 1 Skenario 2

Pada skenario 3 menunjukkan indikasi sedikit *overfitting*. Meskipun *learning rate* yang kecil membantu dalam proses pembelajaran yang stabil, pelatihan selama 4 *epoch* tampaknya model mulai menghafal data latih, yang mengakibatkan penurunan kemampuan generalisasi pada data validasi. Berikut grafik akurasi *training* dan *validation* serta *training loss* dan *validation loss* pada skenario 3.



Gambar 4. 18 Grafik Akurasi Training dan Validation serta Grafik Training Loss dan Validation Loss Paslon 1 Skenario 3

Pada skenario 4 grafik menunjukkan overfitting. Akurasi latih terus meningkat, namun akurasi validasi stagnan setelah *epoch* 3. *Loss* latih terus menurun, sementara *loss* validasi meningkat setelah *epoch* 3, mengindikasikan model menghafal data latih dan menurunkan kemampuan generalisasi. Berikut grafik akurasi *training* dan *validation* serta *training loss* dan *validation loss* pada skenario 4.



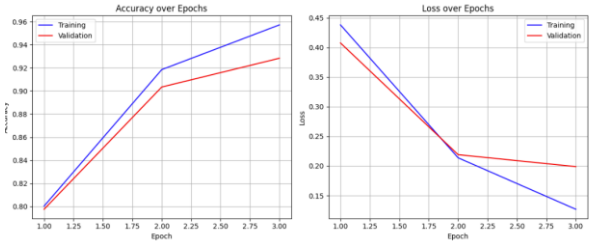
Gambar 4. 19 Grafik Akurasi Training dan Validation serta Grafik Training Loss dan Validation Loss Paslon 1 Skenario 4

Secara keseluruhan dari keempat skenario yang diuji, skenario 1 menjadi skenario terbaik karena memberikan keseimbangan yang optimal antara akurasi *training* dan *validation* serta memiliki performa yang lebih stabil tanpa indikasi *overfitting*.

Tabel 4. 13 Hasil Skenario Hyperparameter Paslon 2

Skenario	**Data**	**Accuracy**	**Precision**	**Recall**	**F1 score**
Skenario 1	Training	0,95	0,93	0,94	0,93
	Validation	0,93	0,92	0,92	0,92
Skenario 2	Training	0,98	0,92	0,94	0,93
	Validation	0,93	0,94	0,93	0,93
Skenario 3	Training	0,98	0,90	0,94	0,92
	Validation	0,92	0,94	0,90	0,92
Skenario 4	Training	0,98	0,92	0,96	0,94
	Validation	0,94	0,96	0,92	0,94

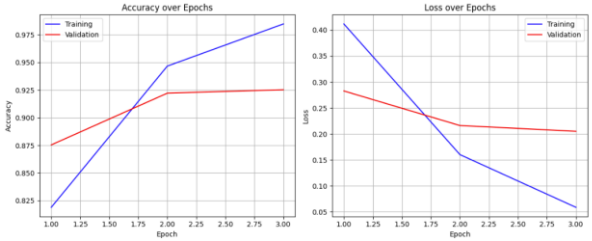
Berdasarkan hasil tabel skenario diatas, grafik skenario 1 menunjukkan grafik akurasi dan *loss* pada data pelatihan dan validasi selama 3 *epoch* dengan *learning rate* 3e-5. Akurasi meningkat secara konsisten pada kedua data, dengan akurasi pelatihan mencapai 96% dan validasi 93%. Sementara itu, nilai *loss* menurun signifikan, menunjukkan proses pelatihan berjalan efektif. Perbedaan antara data pelatihan dan validasi relatif kecil, sehingga model tidak menunjukkan *overfitting*. Berikut grafik akurasi *training* dan *validation* serta *training loss* dan *validation loss* pada skenario 1.



Gambar 4. 20 Grafik Akurasi Training dan Validation serta Grafik Training Loss dan Validation Loss Paslon 2 Skenario 1

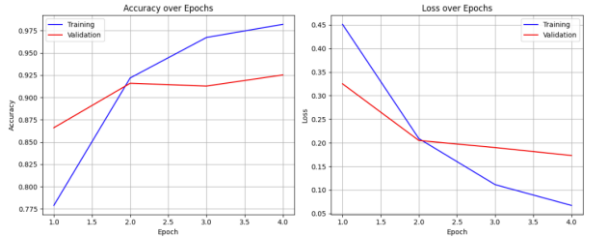
Pada skenario 2, grafik menunjukkan akurasi dan *loss* selama 3 *epoch* dengan *learning rate* 5e-5. Akurasi pelatihan meningkat hingga 98% dan validasi mencapai 93%. *Loss* pelatihan turun signifikan menjadi 0.06, sedangkan *loss* validasi menurun ke 0.21. Model menunjukkan adanya indikasi *overfitting* ringan . Berikut grafik akurasi *training* dan

validation serta *training loss* dan *validation loss* pada skenario 2.



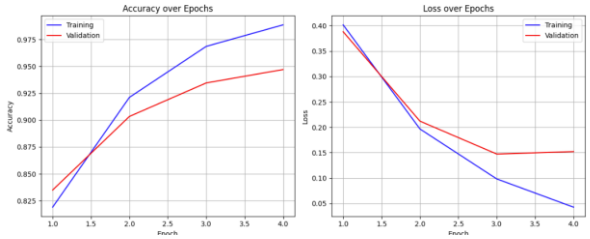
Gambar 4. 21 Grafik Akurasi Training dan Validation serta Grafik Training Loss dan Validation Loss Paslon 2 Skenario 2

Pada skenario 3, grafik menunjukkan akurasi dan *loss* selama 4 *epoch* dengan *learning rate* 3e-5. Akurasi pelatihan meningkat hingga 98%, sedangkan akurasi validasi mencapai 93%. *Loss* pelatihan terus menurun hingga 0.06, dan *loss* validasi juga menurun stabil hingga 0.17. Model menunjukkan performa yang baik tanpa indikasi *overfitting*. Berikut grafik akurasi *training* dan *validation* serta *training loss* dan *validation loss* pada skenario 3.



Gambar 4. 22 Grafik Akurasi Training dan Validation serta Grafik Training Loss dan Validation Loss Paslon 2 Skenario 3

Pada skenario 4, grafik menunjukkan akurasi dan loss selama 4 epoch dengan learning rate 5e-5. Akurasi pelatihan meningkat hingga 98%, dan akurasi validasi mencapai 95%. Nilai loss pada data pelatihan menurun signifikan hingga 0.04, sementara loss validasi menurun hingga 0.15. Model menunjukkan adanya overfitting, terutama jika pelatihan dilanjutkan lebih lama. Berikut grafik akurasi *training* dan *validation* serta *training loss* dan *validation loss* pada skenario 4.

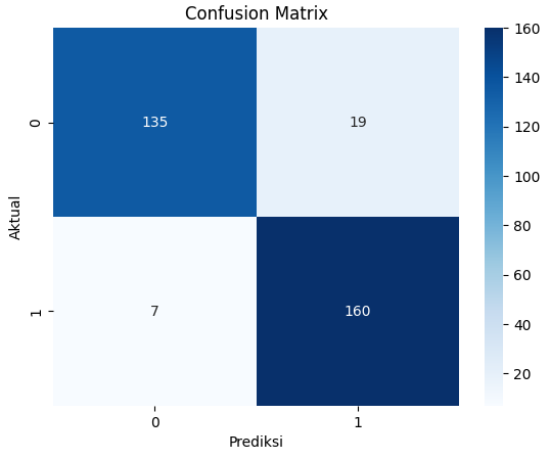


Gambar 4. 23 Grafik Akurasi Training dan Validation serta Grafik Training Loss dan Validation Loss Paslon 2 Skenario 4

Kesimpulan skenario yang diuji pada data paslon 2, skenario 1 dan 3 menunjukkan performa paling stabil tanpa indikasi *overfitting*. Skenario 2 dan 4 berpotensi *overfitting*, khususnya jika pelatihan dilanjutkan lebih lama. Pemilihan kombinasi *epoch* dan *learning rate* yang tepat sangat berpengaruh terhadap kemampuan generalisasi model. Skenario terbaik pada paslon 2 yaitu skenario 1.

H. Evaluasi Matrix

Setelah dilakukan pelatihan dan validasi terpilih opsi terbaik menggunakan skenario 1 karena model menunjukkan performa yang baik dengan menggunakan *epoch* 3, *learning rate* 3e-5, dan *batch size* 32. Berikut merupakan diagram *confusion matrix* dan performa dari model.



Gambar 4. 24 Diagram Confusion Matrix

Dari confusion matrix tersebut jika digambarkan dalam bentuk tabel sebagai berikut.

Tabel 4. 14 Hasil Confusion Matrix

	Negatif	Positif	Total
Negatif	135 (TN)	19 (FP)	154
Positif	7 (FN)	160 (TP)	167
Total	142	179	321

Untuk mengevaluasi performa model dalam penelitian ini, akan dihitung nilai *accuracy*, *precision*, *recall*, dan *f1-score* baik secara manual maupun otomatis. Perhitungan manual dilakukan menggunakan rumus matematis, sedangkan perhitungan otomatis dibantu oleh program *Python.*

Sebagai langkah awal dalam mengevaluasi performa model, perhitungan secara manual akan dilakukan terlebih dahulu sebelum beralih ke metode otomatis. Berikut merupakan penjabarannya:

1. Accuracy

   $accuracy = \frac{TP+TN}{Total\ Data}$

   $accuracy = \frac{160+135}{321} = \frac{295}{321} = 0.9190$ (91,90%)

2. Precision tiap kelas

   - Positif

     precision pos $= \frac{TP}{TP+FP}$

     precision pos $= \frac{160}{160+19} = \frac{160}{179} = 0{,}894$ (89,4%)

   - Negatif

     precision neg $= \frac{TN}{TN+FN}$

     precision neg $= \frac{135}{135+7} = \frac{135}{142} = 0{,}951$ (95,1%)

3. Recall Setiap kelas

   - Positif

     Recall pos $= \frac{TP}{TP+FN}$

Recall pos = $\frac{160}{160+7} = \frac{160}{167}$ = 0,9580 (95,80%)

- Negatif

Recall neg = $\frac{TN}{TN+FP}$

Recall neg = $\frac{135}{135+19} = \frac{135}{154}$ = 0,877 (87,7%)

4. F1- Score

- Positif

F1-Score pos = $2\ x\ \frac{precision\ x\ recall}{precision+recall}$

F1-Score pos = $2\ x\ \frac{0{,}894\ x\ 0{,}958}{0{,}894+0{,}958} = 2\ x\ \frac{0{,}855}{1{,}851}$

= 0,925 (92,5%)

- Negatif

F1-Score neg = $2\ x\ \frac{precision\ x\ recall}{precision+recall}$

F1-Score neg = $2\ x\ \frac{0{,}951\ x\ 0{,}877}{0{,}951+0{,}877} = 2\ x\ \frac{0{,}834}{1{,}828}$

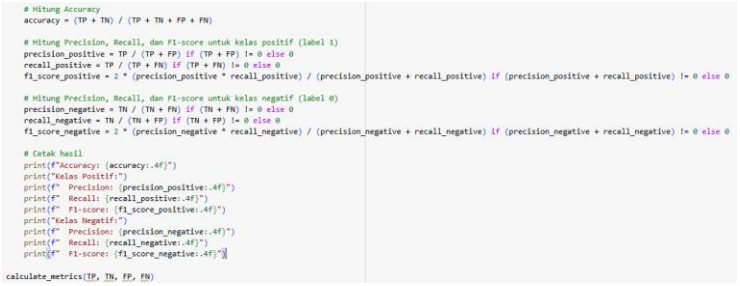
= 0,913 (91,3%)

Berikut merupakan hasil perhitungan manual dari performa model dalam bentuk tabel.

Tabel 4. 15 Hasil Perhitungan Manual Performa Model

Matrix	**Positif**	**Negatif**	**Rata-rata Makro**
Presisi	89,4%	95,1%	92,25%
Recall	95,8%	87,7%	91,75%
F1-Score	92,5%	91,3%	91,9%
Akurasi	91,9%		

Selanjutnya, untuk perhitungan nilai *accuracy*, *precision*, *recall* dan *f1-score* secara otomatis, dapat menggunakan program Bahasa *Python*. Berikut merupakan *source code* yang digunakan untuk menghitung nilai *accuracy*, *precision*, *recall*, dan *f1-score* secara otomatis.



```
# Hitung Accuracy
accuracy = (TP + TN) / (TP + TN + FP + FN)

# Hitung Precision, Recall, dan F1-score untuk kelas positif (label 1)
precision_positive = TP / (TP + FP) if (TP + FP) != 0 else 0
recall_positive = TP / (TP + FN) if (TP + FN) != 0 else 0
f1_score_positive = 2 * (precision_positive * recall_positive) / (precision_positive + recall_positive) if (precision_positive + recall_positive) != 0 else 0

# Hitung Precision, Recall, dan F1-score untuk kelas negatif (label 0)
precision_negative = TN / (TN + FN) if (TN + FN) != 0 else 0
recall_negative = TN / (TN + FP) if (TN + FP) != 0 else 0
f1_score_negative = 2 * (precision_negative * recall_negative) / (precision_negative + recall_negative) if (precision_negative + recall_negative) != 0 else 0

# Cetak hasil
print(f"Accuracy: {accuracy:.4f}")
print("Kelas Positif:")
print(f"  Precision: {precision_positive:.4f}")
print(f"  Recall: {recall_positive:.4f}")
print(f"  F1-score: {f1_score_positive:.4f}")
print("Kelas Negatif:")
print(f"  Precision: {precision_negative:.4f}")
print(f"  Recall: {recall_negative:.4f}")
print(f"  F1-score: {f1_score_negative:.4f}")

calculate_metrics(TP, TN, FP, FN)
```

Gambar 4. 25 Source Code Perhitungan Otomatis Performa Model

Dari gambar tersebut akan menampilkan hasil dan performa model melalui nilai *accuracy*, *precision*, *recall* dan *f1-score* dari setiap kelas sentimen. Berikut merupakan hasil perhitungannya.

```
Accuracy: 0.9190
Kelas Positif:
  Precision: 0.8939
  Recall: 0.9581
  F1-score: 0.9249
Kelas Negatif:
  Precision: 0.9507
  Recall: 0.8766
  F1-score: 0.9122
```

Gambar 4. 26 Hasil Perhitungan Otomatis Performa Model

Berdasarkan hasil perhitungan manual dan otomatis, model klasifikasi sentimen menunjukkan performa yang sangat

baik dengan akurasi tinggi dan keseimbangan yang baik antara *presisi*, *recall*, serta *f1-score*. Berdasarkan hasil evaluasi performa model, diperoleh akurasi sebesar 91,9%. Evaluasi lebih lanjut dilakukan pada masing masing kelas untuk melihat keseimbangan performa klasifikasi. Pada kelas positif, model menghasilkan *precision* sebesar 89,4%, *recall* 95,8%, dan *f1-score* 92,5%. Sementara itu pada kelas negatif, *precision* mencapai 95,1%, *recall* 87,7%, dan *f1-score* 91,2%. Nilai nilai tersebut menunjukkan bahwa model mampu mengklasifikasi data secara cukup seimbang antar kelas.

Perhitungan matrik secara otomatis menggunakan python telah sesuai dengan hasil perhitungan manual, sehingga dapat disimpulkan bahwa proses evaluasi dilakukan secara valid dan akurat.

I. Pengujian akhir Model

Pengujian akhir dilakukan untuk mengevaluasi kemampuan generalisasi model dalam mengklasifikasikan data di luar data latih dan validasi. Pada tahap ini, model diuji menggunakan dua jenis data:

1. Dataset berlabel (data yang telah melalui pelabelan manual)
2. Dataset tidak berlabel (data berupa kalimat acak tanpa label)

Pengujian terhadap dataset berlabel bertujuan untuk mengukur akurasi prediksi model mampu mengklasifikasikan data dengan benar berdasarkan label yang diketahui, sedangkan pengujian pada data yang tidak berlabel digunakan untuk melihat sejauh mana model mampu memprediksi sentimen secara umtum terhadap ata baru yang belum dikenali sebelumnya.

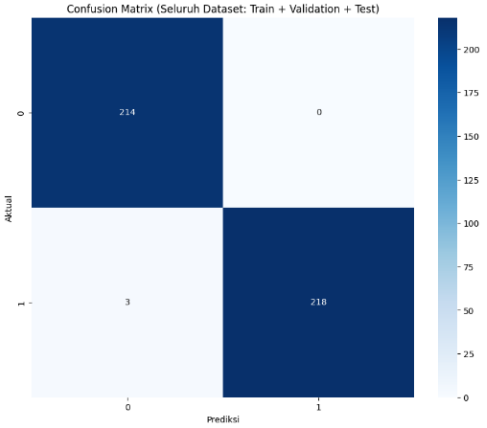
1. Pengujian Dataset Berlabel

   Pada pengujian ini, model akan diuji menggunakan dataset uji yang terdiri dari 426 data, yang terdiri dari 214 data berlabel positif, 221 data berlabel negatif. Berikut merupakan *source code* yang digunakan pada pengujian dengan dataset berlabel.

```
# Training loop
train_losses = []
train_accuracies = []
val_losses = []
val_accuracies = []

print(f"\nMulai training untuk {num_epochs} epochs...")

for epoch in range(num_epochs):
    print(f"\nEpoch {epoch+1}/{num_epochs}")

    # Training
    train_loss, train_acc = train_epoch(model, train_dataloader, optimizer, scheduler, device)
    train_losses.append(train_loss)
    train_accuracies.append(train_acc)

    # Validation
    val_loss, val_acc, val_report = evaluate(model, val_dataloader, device)
    val_losses.append(val_loss)
    val_accuracies.append(val_acc)

    print(f"Train Loss: {train_loss:.4f}, Train Accuracy: {train_acc:.4f}")
    print(f"Val Loss: {val_loss:.4f}, Val Accuracy: {val_acc:.4f}")
    print("Validation Classification Report:")
    print(val_report)
```

Gambar 4. 27 Source Code Pengujian Dataset Berlabel

Dari proses pengujian menggunakan dataset uji berlabel, didapatkan hasil *confusion matrix* sebagai berikut.



Gambar 4. 28 Confusion Matrix Dataset Berlabel

Berdasarkan hasil evaluasi model, confusion matrix yang dihasilkan dengan keseluruhan dataset menunjukkan performa klasifikasi yang sangat baik. Dengan rincian sebagai berikut:

- True Negatif (TN) = 214 sampel
- False Positif (FP) = 0 sampel
- False Negatif (FN) = 3 sampel
- True Positif (TP) = 218 sampel

	precision	recall	f1-score	support
0	0.99	1.00	0.99	214
1	1.00	0.99	0.99	221
accuracy			0.99	435
macro avg	0.99	0.99	0.99	435
weighted avg	0.99	0.99	0.99	435

Gambar 4. 29 Hasil Performa Model Dataset Berlabel

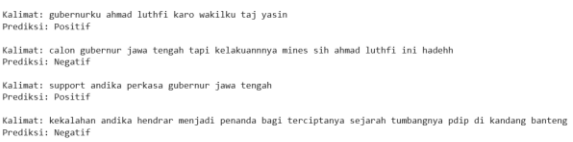
Berdasarkan nilai nilai tersebut, Model menunjukkan nilai yang memuaskan dengan akurasi 99,31%. Selain itu model ini memiliki *recall* sempurna 100% karena tidak terdapat *False Positif*, serta *precision* model juga sempurna 100% karena semua kelas positif benar benar positif dan nilai *f1-score* juga sangat tinggi mencapai 98,64%. Nilai nilai ini mengindikasikan bahwa model memiliki kemampuan klasifikasi yang sangat baik dan generalisasi yang efektif pada dataset yang digunakan.

2. Pengujian Kalimat Acak

Pada pengujian kalimat acak, model diuji dengan kalimat acak dan belum memiliki label. Hal ini bertujuan untuk mengevaluasi kemampuan model dalam mengklasifikasikan sentimen terhadap kalimat-kalimat acak yang tidak terdapat dalam data pelatihan maupun pengujian sebelumnya. Kalimat

disusun secara manual berdasrkan isu atau opini yang beredar di masyarakat. Model diminta untuk memprediksi apakah sentimen dari setiap kalimat tersebut bersifat positif atau negatif. Berikut merupakan hasil prediksi dari beberapa kalimat acak:

1. Kalimat:"*gubernurku ahmad luthfi karo wakilku taj yasin*"
   Prediksi Sentimen: Positif
2. Kalimat:"*calon gubernur jawa tengah tapi kelakuannya mines sih ahmad luthfi ini hadehh*"
   Prediksi Sentimen: Negatif
3. Kalimat: "*support andika perkasa gubernur jawa tengah*"
   Prediksi Sentimen: Positif
4. Kalimat: "*kekalahan andika hendrar menjadi penanda bagi terciptanya sejarah tumbangnya pdip di kandang banteng*"
   Prediksi Sentimen: Negatif



Gambar 4. 30 Hasil Prediksi Pengujian Kalimat Acak

Hasil prediksi menunjukkan bahwa model mampu mengidentifikasi sentimen dari kalimat-kalimat tersebut dengan cukup baik, bahkan terhadap variasi bahasa informal dan campuran yang digunakan dalam media sosial. Hal ini membuktikan bahwa model IndoBERTweet yang digunakan dan telah dilatih dapat berfungsi secara efektif dalam konteks nyata, khususnya dalam memahami opini publik di platform digital.

3. Hasil Keseluruhan Pengujian

Secara keseluruhan, hasil pengujian menunjukkan bahwa model memiliki kemampuan klasifikasi yang baik dan generalisasi yang cukup kuat. Model yang tepat berada pada skenario 1 dengan *epoch* 3, *learning rate* 3e-5, dan *batch size* 32. Dengan pasangan calon nomer urut 1 mendapatkan hasil *accuracy* 93%, *precision* 90%, *recall* 84%, *f1-score* 87% sedangkan untuk pasangan calon nomer urut 2 mendapatkan hasil *accuracy* 95%, *precision* 93%, *recall* 94%, dan *f1-score* 93%. Hal ini menunjukkan bahwa skenario 1 mendapatkan hasil yang stabil tanpa adanya indikasi *overfitting*.

BAB V
KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan hasil penelitian dan pembahasan yang telah dipaparkan dapat disimpulkan bahwa:

1. Didapati bahwa mayoritas sentimen masyarakat terhadap pasangan calon gubernur dan wakil gubernur Jawa Tengah 2024 cenderung positif. Dari 1.492 *tweet* yang dikumpulkan pada paslon 1 sebanyak 1.144 *tweet* yang berlabel positif dan 347 *tweet* yang berlabel negatif. Sedangkan dari 1.444 *tweet* yang dikumpulkan pada paslon 2 sebanyak 1.180 *tweet* yang berlabel positif dan 263 *tweet* yang berlabel negatif. Dengan demikian paslon 2 mendapatkan persentase sentimen positif yang lebih tinggi dan sentimen negatif yang cukup rendah dibandingkan paslon 1.
2. Dalam implementasi model IndoBERT untuk analisis sentimen masyarakat terhadap pasangan calon gubernur dan wakil gubernur jawa tengah 2024, penelitian ini menggunakan model *indobertweet-base-uncased* yang telah melalui proses *fine-tuning*. Tahapan utama dalam implementasi mencakup *pre-processing* data yang meliputi pembersihan teks, mengubah huruf besar menjadi huruf kecil (*case folding*), dan normalisasi

ejaan. Untuk menangani ketidakseimbangan kelas dalam dataset, dilakukan *Easy Data Augmentation* (EDA) guna menambah jumlah data negatif. Proses tokenisasi dilakukan dengan *WordPiece Tokenization* yang disediakan oleh IndoBERT. Model dilatih dengan empat skenario *hyperparameter*, dimana skenario terbaik diperoleh pada kombinasi *epoch* 3, *learning rate* 3e-5 dan *batch size* 32, yang memberikan hasil paling optimal.

3. Evaluasi performa model IndoBERT dilakukan melalui pengujian pada dataset uji berlabel yang terdiri dari 426 data serta dilakukan juga pengujian dengan kalimat acak tanpa label. Hasil pengujian menunjukkan bahwa model memiliki akurasi tinggi 99,31%, dengan *precision* dan *recall* mencapai 100%, dan f1-score 98,64%. Dengan nilai – nilai tersebut mengindikasikan bahwa model memiliki kemampuan klasifikasi yang sangat baik dan generalisasi yang efektif. Dalam pengujian dengan kalimat acak, model mampu mengklasifikasikan sentimen dengan akurat. Secara keseluruhan, penelitian ini menunjukkan bahwa IndoBERT dapat menjadi alat yang efektif dalam menganalisis sentimen masyarakat terhadap pasangan calon gubernur dan wakil gubernur Jawa Tengah 2024.

B. Saran

Berdasarkan hasil penelitian ini, terdapat beberapa rekomendasi untuk penelitian dan pengembangan lebih lanjur sebagai berikut:

1. Peningkatan kualitas data sangat diperlukan untuk memastikan representasi yang lebih seimbang antara sentimen posititf dengan negatif. Dataset dalam penelitian ini didominasi oleh sentimen positif, sehingga pengumpulan data tambahan seperti berita online, media sosial lain dan forum diskusi dapat membantu mendapatkan perspektif yang lebih beragam.
2. Model IndoBERT yang telah dikembangkan dalam penelitian ini dapat diuji pada platform media sosial lain, seperti Instagram, Facebook, atau Youtube. Hal ini penting untuk mengetahui bagaimana sentimen masyarakat terhadap pasangan calon gubernur dan wakil gubernur Jawa Tengah 2024 berkembang di berbagai media sosial, yang mungkin memiliki karakteristik pengguna yang berbeda. Dengan demikian, model dapat lebih fleksibel dalam menangkap opini publik dari berbagai sumber.

DAFTAR PUSTAKA

Annizar, B. (2024). *Upaya Pemprov Jawa Tengah Sukseskan Pilkada Serentak 2024*. Tirto.Id.

Azriyan Arham. (2023). *Labeling Sentimen Bahasa Indonesia Secara Otomatis*.

Delfian, A. (2018). *Analisis Sentimen Teks Bahasa Indonesia Pada Media Sosial Menggunakan Algoritma Convolutional Neural Network (Studi Kasus : e-commerce)*.

Eka Sembodo, J., Budi Setiawan, E., & Abdurahman Baizal, Z. (2016). *Data Crawling Otomatis pada Twitter*. 11–16. https://doi.org/10.21108/indosc.2016.111

Endrik, & Nugroho Agung. (2023). Penerapan Algoritma Naive Bayes dan PSO pada. *Remik*.

Fauzan, M. A. (20022). *PENERAPAN SENTIMEN LEKSIKON INDONESIA PADA ANALISIS SENTIMEN MENGENAI OPINI MASYARAKAT DI TWITTER TERHADAP KEBIJAKAN PPKM DARURAT MENGGUNAKAN ALGORITMA MAXIMUM ENTROPY SKRIPSI Oleh*.

Imron, S., Setiawan, E. I., & Santoso, J. (2023). Deteksi Aspek Review E-Commerce Menggunakan IndoBERT Embedding dan CNN. *Journal of Intelligent System and Computation*, *5*(1), 10–16. https://doi.org/10.52985/insyst.v5i1.267

Jayadianti, H., Kaswidjanti, W., Utomo, A. T., Saifullah, S., Dwiyanto, F. A., & Drezewski, R. (2022). Sentiment analysis of Indonesian reviews using

fine-tuning IndoBERT and R-CNN. *ILKOM Jurnal Ilmiah*, *14*(3), 348–354. https://doi.org/10.33096/ilkom.v14i3.1505.348-354

Jayaswal, V. (2020). *Performance Metrics: Confusion matrix, Precision, Recall, and F1 Score*. Medium.

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). *IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP*. http://arxiv.org/abs/2011.00677

Liesting, T., Frasincar, F., & Truşcă, M. M. (2021). *Data Augmentation in a Hybrid Approach for Aspect-Based Sentiment Analysis*. https://doi.org/10.48550/arXiv.2103.15912

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Oktavian, R. (2024). *ANALISIS SENTIMEN AKSI SERUAN BOIKOT PRODUK TERAFILIASI ISRAEL DI TWITTER MENGGUNAKAN ALGORITMA NAÏVE BAYES, K-NEAREST NEIGHBORS, DECISION TREE*.

Pradana, L. S. (2024). *Analisis Sentimen Masyarakat Media Sosial Twitter Terhadap Kinerja Pejabat Gubernur DKI Jakarta Menggunakan Model IndoBERT*.

Pradana Rachman, F., Santoso, H., & History, A. (2021). Perbandingan Model Deep Learning untuk Klasifikasi Sentiment Analysis dengan Teknik Natural Languange Processing. *Jurnal Teknologi Dan Manajemen Informatika*, *7*(2), 103–112. http://http://jurnal.unmer.ac.id/index.php/jtmi

Praptiwi, Y. (2018). *ANALISIS SENTIMEN ONLINE REVIEW PENGGUNA E-COMMERCE MENGGUNAKAN METODE SUPPORT VECTOR MACHINE DAN MAXIMUM ENTROPY*.

Retno, H. (2013). Pilkada Langsung dan Stabilitas Pemerintahan Di Jawa Tengah. *Journal Of Social and Political Studies*.

Rochmawati, N., Hidayati, H. B., Yamasari, Y., Peni, H., Tjahyaningtijas, A., Yustanti, W., & Prihanto, A. (2021). Analisa Learning rate dan Batch size Pada Klasifikasi Covid Menggunakan Deep learning dengan Optimizer Adam. *JIEET (Journal Information Engineering and Educational Technology)*, *05*(02).

Satriawan, E. B., Hadi Wijoyo, S., & Ratnawati, D. E. (2024). Analisis Sentimen Terhadap Pendapat Masyarakat Mengenai Pilkada 2024 Menggunakan Metode Support Vector Machine (SVM). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *1*(1), 2548–2964. http://j-ptiik.ub.ac.id

Sayarizki, P., & Nurrahmi, H. (2024). Implementation of IndoBERT for Sentiment Analysis of Indonesian Presidential Candidates. *Journal on Computing*, *9*(2), 61–72. https://doi.org/10.34818/indojc.2024.9.2.934

Setyo Nugroho, K., Yullian Sukmadewa, A., Wuswilahaken, H. D., Abdurrachman Bachtiar, F., & Yudistira, N. (2021). BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews. *SIET '21: 6th International Conference on Sustainable Information Engineering and*

Technology 2021.
https://research.google/teams/brain.

Suyanto, Nur Ramadhani, K., & Mandala, S. (2019). *Deep Learning Modernisasi Machine Learning untuk Big Data*.

Tata Lukmana, D., Subanti, S., Susanti, Y., & Studi Statistika, P. (2019). *Analisis Sentimen Terhadap Calon Presiden 2019 Dengan Support Vector Machine di Twitter*.

Wei, J., & Zou, K. (2019). *EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks*. http://arxiv.org/abs/1901.11196

Wildan Amru Hidayat, & Nastiti, V. R. S. (2024). PERBANDINGAN KINERJA PRE-TRAINED INDOBERT-BASE DAN INDOBERT-LITE PADA KLASIFIKASI SENTIMEN ULASAN TIKTOK TOKOPEDIA SELLER CENTER DENGAN MODEL INDOBERT. *JSiI (Jurnal Sistem Informasi)*, *11*(2), 13–20. https://doi.org/10.30656/jsii.v11i2.9168

Wilie, B., Vincentio, K., Indra Winata, G., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., Purwarianti, A., & Bandung, I. T. (2020). *IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding*. https://github.com/annisanurulazhar/absa-playground

Zempi, C. N., Kuswanti, A., & Maryam, S. (2023). ANALISIS PERAN MEDIA SOSIAL DALAM PEMBENTUKAN PENGETAHUAN POLITIK MASYARAKAT. *EKSPRESI DAN PERSEPSI : JURNAL ILMU KOMUNIKASI*, *6*(1), 116–123. https://doi.org/10.33822/jep.v6i1.5286

LAMPIRAN LAMPIRAN

Lampiran 1 **Contoh Dokumen Hasil *Crawling* Data Twitter Paslon 1**

No	created_at	full_text
1	Sat Oct 19 11:30:30 +0000 2024	Pelaku seni Jatemg pilih yg ga kaleng². Andika Perkasa-Hendi untuk #jatengperkasa Pilih Rambo bukan Sambo. https://t.co/wrQ8bgA8qA
2	Thu Oct 17 05:11:07 +0000 2024	Pagi tadi saya dan Mas Hendrar Prihadi bertemu dengan Ibu Juliana Richards Wakil Kepala Program Kerjasama Kedutaan Besar Inggris. Pertemuan ini membahas peluang kerjasama strategis di bidang pengentasan kemiskinan kesehatan dan pendidikan. Kedutaan Besar Inggris menyatakan https://t.co/Kl19v03Ddj
3	Fri Oct 18 08:23:22 +0000 2024	Memang benar kata pihak lawan pak Andika ini memang kelasnya calon presiden. Pengalamannya bkn hny nasional tapi juga internasional. Warga Jawa Tengah beruntung

		punya gubernur sekelas presiden. Calon Gubernur dan Wakil Gubernur Jawa Tengah Andika Perkasa dan Hendrar Prihadi https://t.co/Vv7IZmFYtB
4	Sun Oct 20 02:57:36 +0000 2024	Pak Andika Perkasa selalu jadi idola di mana-mana. Pilih Rambo bukan Sambo. https://t.co/m7j5x6tYjB
5	Thu Oct 17 16:26:08 +0000 2024	Pagi tadi Cagub-Cawagub Jateng Andika Perkasa - Hendrar Prihadi bertemu Ibu Juliana Richards Wakil Kepala Program Kerjasama Kedutaan Besar Inggris di Indonesia. Pertemuan ini membahas peluang kerjasama strategis bidang pengentasan kemiskinan kesehatan & pendidikan antara https://t.co/jlTxMqPBqD
.....	....	....
1488	Wed Oct 30 14:21:58 +0000 2024	Tadi nonton debatnya Andika Hendi dan saya terkesan dengan cara beliau menyampaikan visi-

		misi. Ada komitmen nyata untuk membangun Jawa Tengah yang lebih baik! Semoga bisa terealisasi ya! #JatengPerkasa https://t.co/k5QJLMdOYS
1489	Wed Oct 30 18:00:52 +0000 2024	@Jateng_Twit Gak sederhana tapi calon lain gak bisa . Pak Hendi kata gue lah. https://t.co/SQKaT7XM7K
1490	Wed Oct 30 18:00:52 +0000 2024	@Jateng_Twit Gak sederhana tapi calon lain gak bisa . Pak Hendi kata gue lah. https://t.co/SQKaT7XM7K
1491	Wed Oct 30 14:13:34 +0000 2024	@jawafess biyen omahku ng smg utara banjire ngeriii udan sediluk banjir. bar p hendi dadi walikota ng smg utara wes gak ono banjir. contone wes biyen kota lama ki rak kopen bar di revitalisasi karo p hendi saiki dadi uapikk poll. pokoe 01
1492	Wed Oct 30 12:19:31 +0000 2024,13	Luar biasa nih Jateng Pokoknya Andika yg punya ya Jawa Tengah Yakin makin kuat dan perkasa karena ANDIKA - HENDI yg TERBAIK untuk JATENG

		https://t.co/8cYfp9oKkR

Lampiran 2 **Contoh Dokumen Hasil *Crawling* Data Twitter Paslon 2**

No	created_at	full_text
1	Wed Oct 16 04:55:48 +0000 2024	Padahal sudah ngerahin bot Buat nyerang Andika maupun PDIP tapi Ahmad Luthfi - Taj Yasin masih aja ketinggalan jauh dari Andika Perkasa - Hendrar Prihadi. Ini menjadi bukti bahwa masyarakat Jawa Tengah itu butuh Rambo bukan Sambo. https://t.co/ZGdHrST8sJ
2	Sat Oct 19 13:55:34 +0000 2024	Kasus ini jaman si Lutfi (Tukang remes pantat biduan) jadi Kapolda JATENG.. Kasian sekali korban sudah mencari keadilan tapi tidak di gubris oleh ISILOP .. Emang kapret itu lembaga jadi inget motor aing yang ilang pas laporan ke isilop malah di bentak . @WagimanDeep212_ https://t.co/rksaVKxBCs

3	Thu Oct 17 15:22:43 +0000 2024	Calon Gubernur Jawa Tengah Ahmad Luthfi https://t.co/1Y1fJ2dGvi
4	Tue Oct 15 12:39:37 +0000 2024	Sudah waktunya Jateng dipimpin pemimpin yang tegas mengayomi dan ada untuk rakyanya serta dekat dengan ulama dan tokoh agama yang ada di jateng monggo coblos nomor urut 02 Pak Ahmad Luthfi Gus Yasin. #relawanluthfijayapati @SangPengayom24 @gusyasinmaimoen https://t.co/KJSxBCeYbj
5	Fri Oct 18 16:20:29 +0000 2024	Bismillah semua karena Allah yuk dukung Pak Ahmad Luthfi Gus Yasin Ngopeni Ngelakoni nomor urut 02 untuk Jateng lebih baik. #relawanluthfijayapati #NgopeniNgelakoni #santripatibergerak #pati #wongpati_keren https://t.co/2IG3obOQxm
.....	.....	.....

	.....	.....
1440	Mon Nov 04 12:20:09 +0000 2024	Pernyataan surveyor Litbang Kompas Pdhal Taj Yasin itu ex-wagub Jateng yg pengalaman tetapi publik pikir (berharap) Kaesang yg maju. Jadi Pak @Jokowi boleh lah turun bantu Luthfi-Taj Yasin kampanye mantep banget itu.
1441	Mon Nov 04 11:19:39 +0000 2024	Masih banyaknya pemilih bimbang juga karena tidak ada calon gubernur petahana pada Pilkada Jateng kali ini. Memang ada petahana tapi wagub dan kini cawagub (Taj Yasin). Tidak ada sosok kuat di Jateng saat ini kata Toto.
1442	Mon Nov 04 11:20:54 +0000 2024	Di sisa waktu ini kurang lebih 20 hari semua tim dan relawan harus bekerja keras untuk meningkatkan elektabilitas serta popularitas Pak Luthfi dan Taj Yasin kata Yogo melalui telepon pada Senin (4/11/2024). https://t.co/TyqB9YZYGA

1443	Mon Nov 04 04:14:26 +0000 2024	Bantu analis buat rkyt JTG. Polri ex jtg (independen) & taj Yasin PPP (Ket. ulama bsar Jtg. ex.wkil Ganjar)=) dkg KIM + MUL PDIP Rkyat msh da hrapan baik (makmur) dihandel + bila kel.raja kna hkum VS TNI (PDIPdadakan) & politsi PDIP -) rkyat miskin stabil kan lanjut demi pol.2029
1444	Mon Nov 04 05:00:43 +0000 2024	aku mau kasih sedikit koment tentang calon gubernur di daerah Jawa..dikit aja g banyak2 DKI - RK suswono - Pramono Rano 22nya wakilnya g banget Jateng - Andika Hendra r - Lutfhi taj Yasin kebalikannya 22nya cagubnya g bgt

Lampiran 3 **Contoh Dokumen Data yang diberi Label**

Paslon 1

No	created_at	full_text	sentimen
1	Wed Oct 16 04:55:48 +0000 2024	Padahal sudah ngerahin bot Buat nyerang Andika maupun PDIP tapi Ahmad Luthfi - Taj Yasin masih aja ketinggalan jauh dari Andika Perkasa - Hendrar Prihadi. Ini menjadi bukti bahwa masyarakat Jawa Tengah itu butuh Rambo bukan Sambo. https://t.co/ZGdHrST8sJ	positif
2	Thu Oct 17 05:11:07 +0000 2024	Pagi tadi saya dan Mas Hendrar Prihadi bertemu dengan Ibu Juliana Richards Wakil Kepala Program Kerjasama Kedutaan Besar Inggris. Pertemuan ini membahas peluang kerjasama strategis di bidang pengentasan kemiskinan kesehatan dan pendidikan. Kedutaan Besar Inggris menyatakan	positif

		https://t.co/Kl19v03Ddj	
3	Fri Oct 18 08:23:22 +0000 2024	Memang benar kata pihak lawan pak Andika ini memang kelasnya calon presiden. Pengalamannya bkn hny nasional tapi juga internasional. Warga Jawa Tengah beruntung punya gubernur sekelas presiden. Calon Gubernur dan Wakil Gubernur Jawa Tengah Andika Perkasa dan Hendrar Prihadi https://t.co/Vv7IZmFYtB	positif
4	Sun Oct 20 02:57:36 +0000 2024	Pak Andika Perkasa selalu jadi idola di mana-mana. Pilih Rambo bukan Sambo. https://t.co/m7j5x6tYjB	positif
5	Thu Oct 17 16:26:08 +0000	Pagi tadi Cagub-Cawagub Jateng Andika Perkasa - Hendrar Prihadi bertemu Ibu	positif

	2024	Juliana Richards Wakil Kepala Program Kerjasama Kedutaan Besar Inggris di Indonesia. Pertemuan ini membahas peluang kerjasama strategis bidang pengentasan kemiskinan kesehatan & pendidikan antara https://t.co/jlTxMqPBqD	
.....	....	....	
1488	Wed Oct 30 14:21:58 +0000 2024	Tadi nonton debatnya Andika Hendi dan saya terkesan dengan cara beliau menyampaikan visi-misi. Ada komitmen nyata untuk membangun Jawa Tengah yang lebih baik! Semoga bisa terealisasi ya! #JatengPerkasa https://t.co/k5QJLMdOYS	positif
1489	Wed Oct 30 18:00:52 +0000 2024	@Jateng_Twit Gak sederhana tapi calon lain gak bisa . Pak Hendi kata gue lah. https://t.co/SQKaT7XM7K	positif

1490	Wed Oct 30 16:18:08 +0000 2024,104	Pernyataan Penutup Andika-Hendi di Debat Pilkada Jateng 2024 Pasangan calon (Paslon) gubernur dan wakil gubernur nomor urut 1 @AndikaPerkasa02 - @hendrarprihadi menyampaikan pernyataan penutup dalam debat pertama Pemilihan Gubernur (Pilgub) 2024 di Marina Convention Center https://t.co/XFTMnrtsbC	positif
1491	Wed Oct 30 14:13:34 +0000 2024	@jawafess biyen omahku ng smg utara banjire ngeriii udan sediluk banjir. bar p hendi dadi walikota ng smg utara wes gak ono banjir. contone wes biyen kota lama ki rak kopen bar di revitalisasi karo p hendi saiki dadi uapikk poll. pokoe 01	positif
1492	Wed Oct 30 12:19:31 +0000	Luar biasa nih Jateng Pokoknya Andika yg punya ya Jawa Tengah Yakin makin	positif

	2024,13	kuat dan perkasa karena ANDIKA - HENDI yg TERBAIK untuk JATENG https://t.co/8cYfp9oKkR	

Lampiran 4 **Contoh Dokumen Data yang diberi Label Paslon 2**

No	created_at	full_text	sentimen
1	Wed Oct 16 04:55:48 +0000 2024	Padahal sudah ngerahin bot Buat nyerang Andika maupun PDIP tapi Ahmad Luthfi - Taj Yasin masih aja ketinggalan jauh dari Andika Perkasa - Hendrar Prihadi. Ini menjadi bukti bahwa masyarakat Jawa Tengah itu butuh Rambo bukan Sambo. https://t.co/ZGdHrST8sJ	negatif
2	Sat Oct 19 13:55:34 +0000 2024	Kasus ini jaman si Lutfi (Tukang remes pantat biduan) jadi Kapolda JATENG.. Kasian sekali korban sudah mencari keadilan tapi tidak di gubris oleh ISILOP .. Emang kapret itu lembaga jadi inget motor aing yang ilang pas laporan ke isilop malah di bentak . @WagimanDeep212_ https://t.co/rksaVKxBCs	negatif

3	Thu Oct 17 15:22:43 +0000 2024	Calon Gubernur Jawa Tengah Ahmad Luthfi https://t.co/1Y1fJ2dGvi	positif
4	Tue Oct 15 12:39:37 +0000 2024	Sudah waktunya Jateng dipimpin pemimpin yang tegas mengayomi dan ada untuk rakyanya serta dekat dengan ulama dan tokoh agama yang ada di jateng monggo coblos nomor urut 02 Pak Ahmad Luthfi Gus Yasin. #relawanluthfijayapati @SangPengayom24 @gusyasinmaimoen https://t.co/KJSxBCeYbj	positif
5	Fri Oct 18 16:20:29 +0000 2024	Bismillah semua karena Allah yuk dukung Pak Ahmad Luthfi Gus Yasin Ngopeni Ngelakoni nomor urut 02 untuk Jateng lebih baik. #relawanluthfijayapati #NgopeniNgelakoni #santripatibergerak #pati	positif

		#wongpati_keren https://t.co/2IG3obOQxm	
.....	.....	.....	.....
1440	Mon Nov 04 12:20:09 +0000 2024	Pernyataan surveyor Litbang Kompas Pdhal Taj Yasin itu ex-wagub Jateng yg pengalaman tetapi publik pikir (berharap) Kaesang yg maju. Jadi Pak @Jokowi boleh lah turun bantu Luthfi-Taj Yasin kampanye mantep banget itu.	positif
1441	Mon Nov 04 11:19:39 +0000 2024	Masih banyaknya pemilih bimbang juga karena tidak ada calon gubernur petahana pada Pilkada Jateng kali ini. Memang ada petahana tapi wagub dan kini cawagub (Taj Yasin). Tidak ada sosok kuat di Jateng saat ini kata Toto.	positif
1442	Mon Nov 04 11:20:54 +0000 2024	Di sisa waktu ini kurang lebih 20 hari semua tim dan relawan harus bekerja keras untuk meningkatkan elektabilitas serta popularitas Pak Luthfi	positif

		dan Taj Yasin kata Yogo melalui telepon pada Senin (4/11/2024). https://t.co/TyqB9YZYGA	
1443	Mon Nov 04 04:14:26 +0000 2024	Bantu analis buat rkyt JTG. Polri ex jtg (independen) & taj Yasin PPP (Ket. ulama bsar Jtg. ex.wkil Ganjar)=) dkg KIM + MUL PDIP Rkyat msh da hrapan baik (makmur) dihandel + bila kel.raja kna hkum VS TNI (PDIPdadakan) & politsi PDIP -) rkyat miskin stabil kan lanjut demi pol.2029	positif
1444	Mon Nov 04 05:00:43 +0000 2024	aku mau kasih sedikit koment tentang calon gubernur di daerah Jawa..dikit aja g banyak2 DKI - RK suswono - Pramono Rano 22nya wakilnya g banget Jateng - Andika Hendra r - Lutfhi taj Yasin kebalikannya 22nya cagubnya g bgt	negatif

Lampiran 5 Contoh Data Pendukung "colloquial-indonesian-lexicon.csv"

No	Slang	Normal
1	met	selamat
2	gaharus	enggak harus
3	aminn	amin
4	woww	wow
5	netass	menetas
6	eeeehhhh	eh
7	kata2nyaaa	kata-katanya
8	hallo	halo
9	kaka	kakak
10	ka	kak
11	daah	dah
12	aaaaahhhh	ah
13	yaa	ya
14	smga	semoga
.....	.....	.....
15394	dlm	dalam
15395	wktu	waktu
15396	hr	hari
15397	gatau	enggak tau
15398	fans2	fan-fan
15397	gaharus	enggak harus

Lampiran 6 **Contoh Data Pendukung "*kbba.txt*"**

No	Slang	Normal
1	7an	tujuan
2	@	di
3	ababil	abg labil
4	abis	habis
5	acc	accept
6	ad	ada
7	adlah	adalah
8	adlh	adalah
9	adoh	aduh
10	afaik	as far as i know
11	aha	tertawa
12	ahaha	haha
13	aing	saya
14	aj	saja
15	aja	saja
.....	.....	.....
1315	istaa	mista
1316	benarjujur	benar
1317	benarjujur	jujur
1318	sayan	sayang
1319	mgkin	mungkin

Lampiran 7 **Contoh Data Pendukung "*kamus_slang*"**

No	Slang	Formal
1	woww	wow
2	aminn	amin
3	met	selamat
4	netaas	menetas
5	keberpa	keberapa
6	eeeehhhh	eh
7	kata2nyaaa	kata-katanya
8	hallo	halo
9	kaka	kakak
10	ka	kak
11	daah	dah
12	aaaaahhhh	ah
13	yaa	ya
14	smga	semoga
15	slalu	selalu
.....	.....	.....
15003	gataunya	enggak taunya
15004	gtau	enggak tau
15005	gatau	enggak tau
15006	fans2	fan-fan
15007	gaharus	enggak harus

Lampiran 8 **Contoh Data Pendukung "*slangword.txt*"**

No	Slang	Normal
1	alay	aneh
2	ancur	hancur
3	asal mangap	ceroboh
4	asbun	cerewet
5	asik	asyik
6	bacot	bawel
7	bae	baik
8	baek	baik
9	basbang	basi
10	bcus	becus
11	belagu	songong
12	bete	kesal
13	bgus	bagus
14	bls	balas
15	bner	benar
.....	.....	.....
687	lgi	lagi
688	sampaiken	sampaikan
689	gub	gubernur
690	akherat	akhirat
691	pdi pak	pdip

Lampiran 9 **Daftar Riwayat Hidup**

Riwayat Hidup

A. Identitas Diri

1.	Nama Lengkap	:	Bagus Diaz Pratama
2.	Tempat & Tanggal Lahir	:	Semarang, 12 Desember 2003
3.	Alamat	:	Jl. Beruang Raya V No.13 Rt.5 Rw.2 Kec/ Kel Gayamsari
4.	HP	:	081391541327
5.	Email	:	bagusdiazp@gmail.com

B. Riwayat Pendidikan

1. Sekolah Dasar (SD) Negeri 01 Semarang
2. SMP Kesatrian 1 Semarang
3. Sekolah Menengah Atas (SMA) Negeri 15 Semarang

Semarang, 06 Maret 2025

Pembuat Pernyataan,

Bagus Diaz Pratama

NIM: 2108096059