**Petunjuk Teknis Babak Penyisihan**

**Big Data Challenge Satria Data 2024**

Waktu pengumpulan karya sampai dengan 19 Juni 2024

    Selamat datang di kompetisi Big Data Challenge (BDC) Satria Data 2024. BDC Satria Data adalah sebuah ajang tahunan yang mengumpulkan talenta-talenta terbaik dalam bidang *data science*, memberikan kesempatan kepada para peserta untuk menerapkan pengetahuan dan keterampilan mereka dalam menyelesaikan masalah nyata menggunakan *big data*.

**Pendahuluan**

Media sosial telah berkembang menjadi sumber data yang sangat luas dan variatif, memfasilitasi analisis komprehensif mengenai dinamika dan opini publik. Media sosial X, yang dulu dikenal sebagai Twitter, menghasilkan volume data yang masif. Jutaan tweet, retweet, dan komentar diproduksi dalam hitungan detik. Konten yang dihasilkan oleh pengguna merupakan sumber informasi yang berharga untuk memahami pandangan masyarakat, tren, dan respons terhadap isu tertentu.

Selama musim kampanye presiden di Indonesia, intensitas aktivitas di platform media sosial meningkat secara signifikan. Para kandidat dan pendukungnya memanfaatkan media sosial sebagai alat utama untuk komunikasi dan mobilisasi massa. Kampanye ini tidak hanya merupakan persaingan antar individu, tetapi juga sebagai arena pertarungan ide dan ideologi yang diadu di ranah digital. Media sosial X merupakan salah satu platform utama yang menjadi medan vital untuk melaksanakan strategi kampanye, mengukur sentimen dan dukungan, serta melangsungkan debat dan dialog politik.

Ekosistem dinamis yang tercipta dari kondisi ini menghasilkan data yang mampu memberikan pemahaman mendalam tentang dinamika pemilu dan potensi hasilnya. Namun, terdapat tantangan signifikan dalam mengolah data ini karena karakteristiknya yang tidak terstruktur dan volume yang besar. Pemanfaatan pendekatan seperti *complex* *network analysis, machine learning,* hingga *deep learning* menjadi sangat krusial untuk memproses dan memahami data ini dengan efisien dan efektif. Berikut adalah beberapa contoh umum pemanfaatan data tersebut:

* + **Memahami Perilaku Pemilih**: Data media X menyediakan wawasan mendalam mengenai motivasi, kekhawatiran, dan faktor-faktor yang mendorong partisipasi pemilih. Pemahaman ini sangat berharga bagi partai politik dan kandidat dalam merancang kampanye yang lebih resonan dan efektif dalam meningkatkan keterlibatan pemilih di seluruh spektrum demografis.
  + **Optimalisasi Strategi Kampanye**: Data ekstensif dari media X memungkinkan tim kampanye untuk mengenali topik-topik yang paling menarik bagi pemilih. Melalui analisis interaksi dan diskusi, strategi komunikasi dapat disesuaikan sehingga lebih spesifik dan personal, memperkuat efektivitas kampanye dengan menyampaikan pesan yang lebih relevan dan tepat sasaran kepada kelompok demografis yang diincar.
  + **Prediksi Hasil Pemilu**: Dengan mengumpulkan dan menganalisis informasi tentang frekuensi dan konteks diskusi terkait isu-isu kunci serta sentimen terhadap kandidat, analis dapat mengembangkan model prediktif yang lebih akurat sehingga para stakeholder politik untuk memperoleh pemahaman yang lebih dalam tentang potensi hasil pemilu, memfasilitasi penyesuaian strategi yang diperlukan untuk mengoptimalkan hasil.
  + **Deteksi Kampanye Disinformasi**: Analisis pola distribusi dan sumber informasi memungkinkan deteksi efektif kampanye disinformasi. Dengan mengawasi dan menganalisis tren ini, pihak berwenang dan organisasi relevan dapat bertindak cepat untuk memastikan integritas informasi yang disampaikan kepada publik, memastikan bahwa pemilih mendapat akses ke data yang akurat dan tepercaya.
  + **Pemetaan dan Pengukuran Polarisasi**: Data media X dapat digunakan untuk memetakan jaringan pengguna dan mengukur tingkat polarisasi dapat membantu dalam memahami dinamika polarisasi lebih mendalam. Pengetahuan ini penting untuk merumuskan strategi komunikasi, mendorong dialog yang lebih inklusif dan membangun kesepahaman bersama dalam diskusi publik.

**Permasalahan**

Data UGC (*User Generated Content*) bervolume masif membawa peluang yang luas tetapi juga menimbulkan tantangan yang signifikan dalam pengolahan dan analisisnya. Pada kompetisi Big Data Challenge (BDC) Satria Data 2024, para peserta ditantang untuk melakukan klasifikasi isu yang diangkat oleh penulis dalam tweet-tweet mereka terkait Pilpres 2024 berdasarkan kelas-kelas Astagatra yang merupakan komponen ketahanan nasional. Klasifikasi ini meliputi delapan kelas yaitu Ideologi (nilai-nilai dasar, prinsip, dan pandangan dunia yang mengarahkan bangsa dan negara), Politik (sistem pemerintahan, kebijakan, dan proses politik dalam negara), Ekonomi (pengelolaan ekonomi, keuangan, dan sumber daya untuk menghasilkan kemakmuran bagi masyarakat), Sosial Budaya (nilai sosial, norma budaya, dan aspek kehidupan masyarakat yang berpengaruh terhadap keutuhan dan identitas bangsa), Pertahanan dan Keamanan (pertahanan negara dan keamanan dalam negeri dari ancaman atau gangguan), Sumber Daya Alam (pengelolaan dan pemanfaatan sumber daya alam yang berada di wilayah negara untuk menunjang pembangunan dan keamanan nasional), Geografi (lokasi, kondisi fisik, dan lingkungan alam yang mempengaruhi kehidupan dan kebijakan di negara tersebut), serta Demografi (struktur populasi, pertumbuhan penduduk, distribusi, serta dinamika penduduk yang berpengaruh pada kebijakan dan pembangunan nasional).

Untuk memastikan keadilan dan integritas analisis, tim peserta diberikan dataset yang mencakup teks lengkap tweet beserta labelnya. Peserta hanya diperkenankan untuk menggunakan dataset yang telah disediakan oleh panitia dan tidak diperbolehkan menambahkan data dari sumber eksternal.

Peserta diharapkan mengembangkan model *multiclass text classification* yang akurat untuk mengkategorikan tweet ke dalam kedelapan kelas Astagatra. Model yang dikembangkan harus mampu menangani kompleksitas dan keragaman bahasa dalam tweet, serta dapat mengatasi tantangan seperti ambiguitas, sarkasme, dan kesalahan penulisan yang sering ditemukan dalam data UGC. Peserta diharapkan mengembangkan solusi yang orisinal, kreatif, dan inovatif dalam memecahkan masalah yang ditetapkan oleh panitia.

**Dataset**

Data yang digunakan dalam Big Data Challenge (BDC) Satria Data 2024 merupakan kumpulan interaksi dari media sosial X terkait tiga pasangan calon presiden pada pemilihan presiden 2024 pada masa kampanye. Data disimpan dalam format .csv, termasuk text tweet dan labelnya. Dataset dapat diakses pada link berikut: [https://bit.ly/dataset\_bdc\_2024](https://bit.ly/dataset_bdc_2024" \t "_blank). Peserta diminta untuk melakukan prediksi label pada dataset berikut: [https://bit.ly/dataset\_unlabeled\_bdc\_2024](https://bit.ly/dataset_unlabeled_bdc_2024" \t "_blank).

**Pengumpulan Pekerjaan**

Peserta Big Data Challenge (BDC) Satria Data 2024 mengumpulkan hasil analisis dalam bentuk *comma separated value* (.csv). Gunakan [template](https://bit.ly/template_jawaban_bdc_2024) yang telah disediakan oleh panitia untuk mengumpulkan jawaban. Kumpulkan file dengan penamaan “kode\_kelompok.csv”. Pengiriman jawaban dilakukan melalui mekanisme unggah di sistem ini hingga 19 Juni 2024.

Metrik pengukuran yang digunakan oleh Tim Juri BDC pada tahap penyisihan yaitu *balanced accuracy*. *Balanced accuracy* adalah metrik yang digunakan untuk mengevaluasi kinerja model klasifikasi, terutama ketika data memiliki distribusi kelas yang tidak seimbang. Metrik ini memberikan gambaran yang lebih adil tentang kinerja model dibandingkan dengan metrik akurasi biasa, terutama saat jumlah sampel dalam masing-masing kelas berbeda secara signifikan. *Balanced accuracy* dihitung sebagai rata-rata dari *recall* (atau sensitivitas) untuk setiap kelas. *Recall* untuk sebuah kelas diukur dengan membandingkan jumlah benar positif (TP) untuk kelas tersebut dengan jumlah total yang sebenarnya adalah kelas tersebut (TP + FN, di mana FN adalah *false negatives*).

Pada tahap penyisihan, tim yang lolos ditentukan berdasarkan nilai *balanced accuracy* tertinggi. Jika terdapat tim dengan nilai yang sama, faktor penentu selanjutnya adalah waktu pengumpulan.

**Batasan**

Dalam kompetisi Big Data Challenge (BDC) Satria Data 2024, peserta harus mematuhi batasan-batasan berikut untuk menjaga integritas dan keadilan kompetisi:

* + Peserta hanya diperbolehkan menggunakan data yang telah disediakan oleh panitia. Tidak diperkenankan untuk menggunakan data dari sumber lainnya.
  + Peserta tidak diperkenankan menggunakan *large language model* (LLM) seperti GPT, Claude, Gemini, dan sejenisnya.
  + Peserta dilarang menggunakan isu sensitif, termasuk SARA (Suku, Agama, Ras, dan Antar-golongan), dan dilarang keras menyalahgunakan data untuk tujuan yang tidak etis atau melanggar norma-norma sosial.
  + Hasil dikumpulkan sesuai dengan template dan format yang telah ditetapkan oleh panitia untuk memastikan konsistensi dan memudahkan proses evaluasi.
  + Hasil diserahkan sesuai dengan batas waktu yang telah diinformasikan oleh panitia. Keterlambatan dalam pengumpulan jawaban tidak diterima.

**Aturan Tambahan**

Berikut adalah peraturan tambahan untuk kompetisi Big Data Challenge (BDC) Satria Data 2024:

* + Batas waktu untuk mengunggah hasil dari *challenge* penyisihan yaitu sampai dengan tanggal 19 Juni 2024 pukul 23:59 WIB.
  + Tim peserta mengerjakan *challenge* penyisihan dan dapat melakukan submisi secara online maksimum 3 kali selama periode pengumpulan.
  + Berdasarkan sistem evaluasi yang telah ditetapkan, juri akan memilih 22 tim dengan skor tertinggi (5 tim potensial klaster dan 17 tim dengan metrik terbaik). Pengumuman semifinalis dilaksanakan pada tanggal 22 Juni 2024.
  + Ke-22 tim semifinalis menyelesaikan *challenge* semifinal dan mengunggah hasil secara online sampai dengan 6 Juli 2024 pukul 23:59 WIB.
  + Juri menyeleksi 12 tim dengan metrik atau metodologi terbaik untuk maju ke babak final BDC Satria Data 2024. Pengumuman finalis dilaksanakan pada tanggal 22 Juli 2024.
  + Ke-12 tim finalis akan melakukan *technical meeting* secara luring (waktu dan tempat akan diinformasikan kemudian), wajib mendemonstrasikan kode yang digunakan pada *challenge* penyisihan untuk memastikan keabsahan hasil babak penyisihan, dan menyiapkan bahan presentasi hasil analitika *challenge* semifinal.
  + Babak final dilaksanakan pada 29 Juli – 2 Agustus 2024.

**Diskualifikasi atau Sanksi**

Diskualifikasi atau sanksi akan dikenakan kepada peserta kompetisi yang melanggar ketentuan berikut:

* + Karya yang dikumpulkan tidak mematuhi panduan yang telah ditetapkan, termasuk ketidaksesuaian dengan tema atau topik, penggunaan data di luar yang disediakan, atau penyimpangan dari format dan template yang ditentukan.
  + Karya yang dikumpulkan terbukti mengandung unsur plagiarisme.
  + Peserta melakukan kecurangan, termasuk tetapi tidak terbatas pada upaya melakukan prediksi secara manual atau menggunakan metode yang dilarang dalam kompetisi.
  + Pelanggaran etika kompetisi, termasuk namun tidak terbatas pada penggunaan bahasa atau konten yang tidak pantas, atau tindakan yang merugikan peserta lain atau integritas kompetisi.

Sanksi ini diterapkan untuk menjaga integritas dan keadilan kompetisi, memastikan bahwa semua peserta berkompetisi dalam kondisi yang setara dan menghormati aturan yang telah ditetapkan.