**PAGE REPORT — 1ST CHECKPOINT**

Subtopik 1: **Vaksinasi COVID-19**

Anggota Kelompok E:

1. Arkha Sayoga Mayadi - 1606890252
2. Maulana Wildan Abdurrahman - 1806191641
3. Muhammad Jilham Luthfi - 1806141340
4. Muhammad Ichsanul Amal - 1906353454

**Tahap 2: *Tweet Data Collection***

Pada proses pengumpulan data *tweet*, kami menerapkan dua cara, yakni melalui GitHub dan menggunakan *module* Twint. Proses pengambilan data melalui GitHub menggunakan Twitter API dengan ID *tweet* yang tersedia dari *repository* GitHub. Kemudian, *tweet* yang berhasil diambil akan dicek kembali apakah isinya membicarakan terkait vaksinasi COVID-19 atau bukan.

Pada proses pengumpulan data menggunakan *module* Twint, kami menentukan terlebih dahulu kata kunci pencarian yang akan digunakan. Kami menggunakan kata-kata, seperti vaksin, vaksinasi, dan beberapa produk vaksin yang beredar di Indonesia menurut laman resmi <https://covid19.go.id/tentang-vaksin-covid19>. Namun, kami tidak menggunakan kata janssen dan convidencia karena kami melihat tidak banyak menemukan *tweet* yang mengandung kata-kata tersebut. Hasil dari cara ini adalah kami mendapatkan 12.000 *tweet*. Kemudian, kami melihat beberapa sampel *tweet* dan menemukan ada *tweet* vaksin yang tidak berhubungan dengan covid, seperti hepatitis, cacar, kucing, kitten, dan anabul. Maka dari itu, kami melakukan filter kembali untuk menghilangkan *tweet* tersebut dan hasil akhir yang kami dapatkan adalah 9916 *tweet*.

**Tahap 3: *Text Mining***

Berdasarkan pertanyaan analitik yang dipilih, yakni RQ #1a, kami membuat anotasi data dengan label *stance* yang terdiri dari tiga label, yakni pro, kontra, dan abstain. Sebelum membuat label, kami berdiskusi singkat terkait kriteria dari ketiga label tersebut. Kami menetapkan bahwa *tweet* dengan label pro mengandung intensi berupa dukungan terhadap vaksin dan proses vaksinasi COVID-19. Sementara itu, *tweet* dengan label “KONTRA” mengadung intensi berupa penolakan terhadap vaksin dan proses vaksinasi COVID-19. Terakhir, *tweet* dengan label “ABSTAIN” menandakan tidak ada kecenderungan terkait vaksin, seperti pertanyaan terkait vaksin atau cerita pengalaman setelah vaksinasi.

Selanjutnya, kami melakukan proses *preprocessing* pada data. Kolom yang akan kami gunakan adalah data *tweet*. Pada tahap *preprocessing*, kami menerapkan pembersihan data, seperti *lowercasing*, menghapus emoji, tanda baca, angka, dan proses *trimming*. Selain itu, kami mencoba untuk menghilangkan *stopwords* dan melakukan normalisasi dengan memperbaiki kata *slang* atau kata singkatan. Selanjutnya, kami mentransformasikan data *tweet* menjadi bentuk numerik dengan vektor TF-IDF dengan ukuran 25000 fitur berupa unigram hingga 4-gram. Selanjutnya, kami memilih model *multinomial naive bayes* karena model tersebut memberikan hasil terbaik pada data *tweet* berdasarkan proyek tengah semester. Kami juga melihat bahwa *task* terkait mendapatkan label dari sejumlah kecil data bisa menggunakan pendekatan *semi-supervised learning*. Teknik yang kami gunakan adalah *self-training*, yakni membangun model dari data berlabel untuk memprediksi data tidak berlabel. Kemudian, data dengan prediksi label yang tinggi akan dimasukkan ke dalam *training data* untuk pembangunan model berikutnya.