# **LAPORAN PROYEK AKHIR**

**PRAKTIKUM DATA SCIENCE**

**ANALISIS SENTIMEN TWEETS TENTANG COVID19 DI INDONESIA**



|  |  |
| --- | --- |
| EKA PRYHANANTO | 123190020 |
| MUHAMMAD AYI M. | 123190037 |

**PROGRAM STUDI INFORMATIKA**

**JURUSAN TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK INDUSTRI**

**UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN” YOGYAKARTA**

**2021**

# **PENDAHULUAN**

Virus Corona (COVID 19) adalah penyakit menular yang disebabkan oleh virus. Sebagian besar orang yang tertular COVID 19 akan mengalami gejala ringan hingga sedang, dan akan pulih tanpa penanganan khusus. Akan tetapi sebagian orang akan mengalami sakit parah dan memerlukan bantuan medis dengan penanganan khusus.

Virus ini dapat menyebar dari mulut atau hidung orang yang terinfeksi melalui partikel kecil ketika orang tersebut batuk, bersin, berbicara atau bernafas. Seseorang dapat tertular dengan menghirup udara yang mengandung virus jika berada di dekat orang yang terinfeksi COVID 19.

Kasus COVID 19 di dunia tercatat terdapat 255 juta total kasus dan 5.12 juta meninggal dunia, sedangkan di Indonesia sendiri 4.25 juta total kasus dan 144 ribu meninggal dunia. Seluruh kehidupan di bumi mengalami perubahan yang sangat mencolok, yang awalnya aktivitas kerja, sekolah dan lainnya dilakukan secara tatap muka menjadi dilakukan secara daring. Kebijakan pemerintah disetiap negara pun bermunculan, contohnya di indonesia saat melonjaknya kasus COVID 19 muncul kebijakan PPKM untuk meredakan lonjakan kasus COVID 19.

Dengan adanya musibah tersebut telah memberi pengaruh kepada masyarakat dengan berkomentar opininya di media sosial, salah satunya adalah Twitter. Banyak komentar positif dan negatif melalui tweet yang menanggapi tentang Virus Corona (COVID 19). Komentar dari tweet tersebut dapat dijadikan bahan untuk melakukan sebuah analisis untuk menangkap komentar-komentar dari masyarakat. Akan tetapi komentar-komentar tersebut pasti terdapat banyak simbol yang membuat kesulitan saat memilih komentar tersebut positif atau negatif. Untuk mengatasi persolana tersebut dapat dilakukan perubahan format yang tidak terstruktur menjadi terstruktur.

Untuk mengklasifikasi berbagai pernyataan tersebut dapat dilakukan dengan analisis sentimen. Salah satu algoritma yang bisa digunakan untuk analisis sentimen adalah K-Nearest Neighbors (KNN). Kelebihan menggunakan KNN adalah pelatihan sangat cepat, sederhana dan mudah dipelajari, tahan terhadap data pelatihan yang memiliki noisy, dan efektif jika data pelatihan besar.

# **METODE**

Data yang digunakan di dapat dari Kaggle.com tentang tweet masyarakat tentang COVID19. Metode yang digunakan berupa metode K-Nearest Neighbour (KNN), KNN merupakan salah satu metode untuk mengambil keputusan menggunakan pembelajaran terawasi dimana data yang dimasukan diklasifikasi dengan data terdekat.

K-Nearest Neighbour (KNN) adalah algoritma yang sering digunakan dalam pengklasifikasian. Tujuan dari algoritma KNN adalah mengklasifikasikan objek ke dalam salah satu kelas yang telah ada dalam dataset original yang sebelumnya telah ditentukan. Metode KNN mengelompokan data ke dalam satu kelas yang telah ditetapkan berdasarkan jarak terdekat atau kemiripan dengan data latih (data training) yang telah ditentukan.

Dalam proyek ini tahapan yang dilakukan adalah

1. Mengambil beberapa kolom yang terpenting dari beberapa data secara random dari dataset original maupun dataset training yang telah ditentukan.
2. Setelah selesai melakukan proses pengambilan data, semua data kemudian dilakukan pembersihan, yaitu menghilangkan simbol, alamat web, emoticon, dan yang lainnya agar data yang sebelumnya tidak terstruktur menjadi data terstruktur.
3. Setelah dilakukan pembersihan data maka akan dilakukan pengklasifikasian data untuk mengetahui apakah tweet tersebut merupakan hal positif atau negatif.

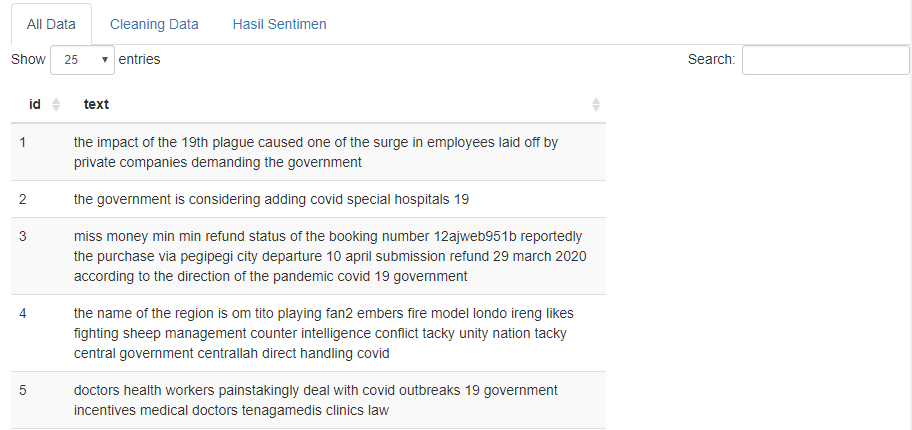
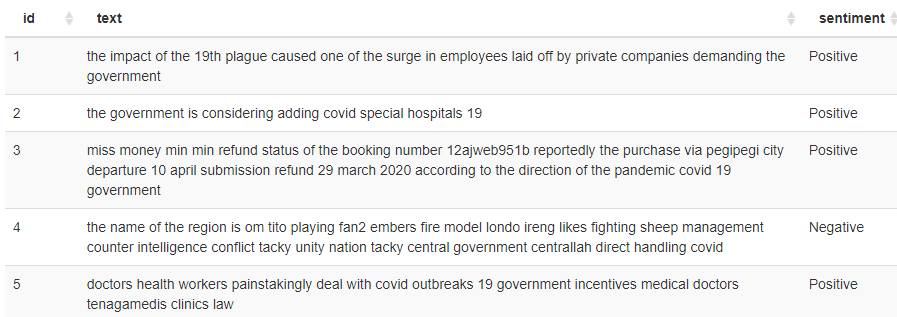
Langkah-langkah kerja KNN :

1. Tentukan parameter K = jumlah dari persekitaran (nearest neighbors)
2. Hitung jarak antara data baru yang ditanyakan dengan seluruh sampel data pelatihan
3. Urutkan seluruh jarak berdasarkan jarak minimum dan tetapkan persekitaran sesuai dengan nilai K.
4. Sesuaikan klasifikasi dari kategori Y dengan persekitaran yang telah ditetapkan
5. Gunakan kelas dengan jumlah terbanyak sebagai dasar menentukan klas dari data baru yang ditanyakan.

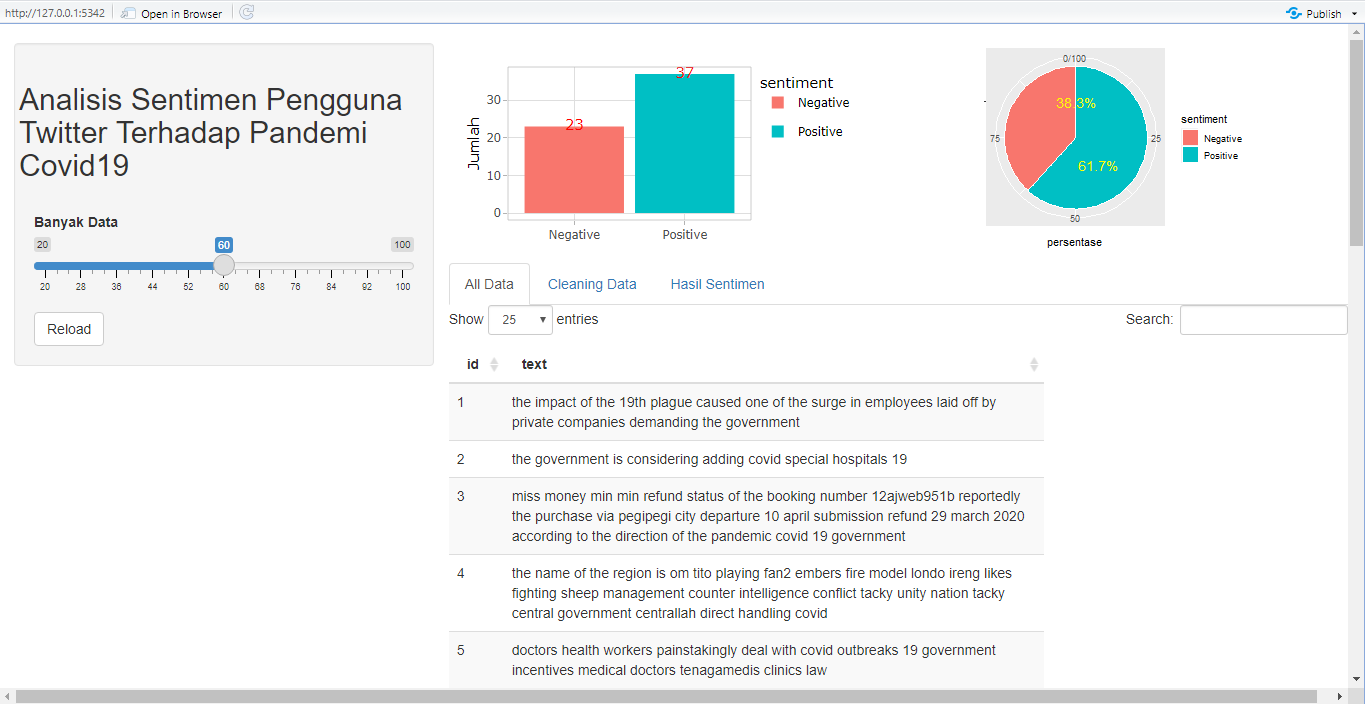
# **HASIL DAN PEMBAHASAN**

Pada proses ini didapatkan 100 data secara random dari dataset original dan 150 data dari dataset training, yang menghasilkan 23 (38.3%) mengandung tweet negatif dan 37 (61.7%) mengandung tweet positif. Data tersebut diproses seperti tabel berikut:

Contoh hasil proses :

* Data :
* Hasil Cleaning data :
* Hasil Sentimen :

Kemudian keseluruhan data disimpan di beberapa file yang dikelompokan berdasarkan hasil dari pemrosessan, yaitu ada data yang akan di prediksi(data\_predict), data prediksi yang sudah dilakukan cleaning(clean\_data\_predict), dan hasil dari prediksi(result\_predict).

Aplikasi halaman dibuat cukup sederhana yang dibuat berupa UI pada bahasa R dengan satu halaman yang menampilkan jumlah data yang akan mau diambil (20-100), tombol reload untuk mengambil ulang data secara random dan hasil plot dari sentimen. Terdapat juga beberapa opsi pilihan, yaitu menampilkan semua data, hasil cleaning data, dan hasil sentimen. 

Pada bagian plot terdapat angka hasil dari klasifikasi sentimen negatif maupun positif. Terdapat dua plot yaitu plot batang dan lingkaran, dimana pada bagian lingkaran menunjukan angka persenan sentimen dari keseluruhan yang sudah di klasifikasi, negatif berwarna oranye sedangkan positif berwarna Hijau.

# **KESIMPULAN**

Dari dataset yang diambil dapat disimpulkan bahwa tweet atau opini dari masyarakat tentang COVID-19 sebagian besar mengandung hal positif. Dari dataset yang sudah di proses dengan KNN didapatkan hasil 38.3% positif dan 61.7 negatif dari 60 data yang diambil random dari 100 data.

Link Github : <https://github.com/ekapryhananto/ProjectDS>

Link Presentasi : <https://youtu.be/OIPVKvJLpTc>