Capstone Project

DETEKSI SPAM DI PLATFORM TWITTER PADA DEMO DPR RI

Presentation by : Frenky Riski Gilang Pratama

LATAR BELAKANG

Twitter saat ini menjadi salah satu media utama bagi masyarakat untuk menyuarakan opini publik. Namun, pada peristiwa Demo DPR RI, marak bermunculan akun anonim yang melakukan aktivitas spam. Keberadaan spam ini tidak hanya mengganggu kualitas informasi, tetapi juga berpotensi memanipulasi opini publik. Oleh karena itu, dibutuhkan sebuah sistem deteksi otomatis berbasis machine learning untuk mengidentifikasi dan memfilter spam secara efektif.

TUJUAN

- Mengidentifikasi dan Mengklasifikasikan Spam: Membangun model untuk secara otomatis membedakan mana tweet yang merupakan spam dan mana yang bukan dalam konteks "Demo DPR RI".
- Menganalisis Pola Spam: Memahami karakteristik dan pola umum dari tweet yang tergolong spam.
- Memanfaatkan Model Bahasa (LLM): Menggunakan IBM Granite LLM untuk melakukan klasifikasi teks dengan kemampuan memberikan penjelasan (explainability) atas keputusannya.
- Menyajikan dashboard interaktif menggunakan Streamlit.

METODE

Proses kerja dalam proyek ini dibagi menjadi beberapa tahap utama:

1. Pengumpulan Data
   - Melakukan crawling data dari Twitter dengan kata kunci terkait "Demo DPR RI".
2. Pemrosesan Awal & Eksplorasi Data
   - Membersihkan data dari duplikasi dan nilai yang hilang.
3. Pemodelan dengan AI
   - Menggunakan IBM Granite LLM untuk mengklasifikasikan setiap tweet sebagai "Spam" atau "Bukan Spam".
4. Analisis & Visualisasi Hasil
   - Menganalisis distribusi hasil, pola kata, dan indikator spam yang paling dominan.
5. Deployment
   - Menyajikan dashboard interaktif menggunakan Streamlit

EKSPLORASI & PERSIAPAN DATA

- Sumber Data: Hasil crawling Twitter dari 1 - 26 Agustus 2025.
- Dataset: demoDPR_gabungan.csv
- Statistik Awal:
  - Jumlah tweet awal: 896
  - Tweet duplikat ditemukan: 28
- Pembersihan Data:
  - Duplikat dihapus berdasarkan konten full_text.
  - Jumlah tweet setelah dibersihkan: 859

MODEL & KRITERIA SPAM

- Model AI: ibm-granite/granite-3.3-8b-instruct melalui Replicate API.
- Tugas Model: Menganalisis tweet dan memberikan label [Spam] atau [Bukan Spam] beserta alasan dan skor kepercayaan.

Kriteria Spam

1. URL/Link mencurigakan (bit.ly, dll.)
2. Promosi berlebihan (GRATIS!, MENANG!)
3. Penggunaan tanda baca atau emoji berlebihan
4. Penggunaan huruf kapital berlebihan
5. Konten duplikat/template yang sama

Kriteria Non Spam

1. Percakapan normal/diskusi
2. Sharing berita/informasi faktual
3. Opini pribadi yang wajar
4. Interaksi sosial normal
5. Konten edukatif
6. Update status personal

HASIL ANALISIS & DISTRIBUSI

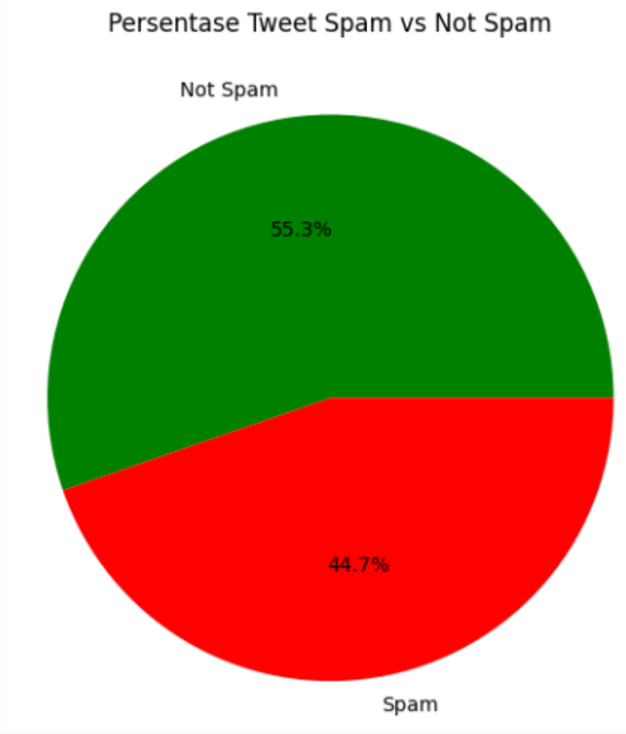
Model berhasil menganalisis 859 tweet dengan hasil sebagai berikut:

- Spam: 384 tweet (44.7%)
- Bukan Spam: 475 tweet (55.3%)

Analisis Skor Kepercayaan:

- Skor Kepercayaan untuk spam adalah >5 dan untuk non spam adalah <5 dari 10
  - Rata-rata skor kepercayaan Spam: 7.70 / 10
  - Rata-rata skor kepercayaan Bukan Spam: 2.99 / 10

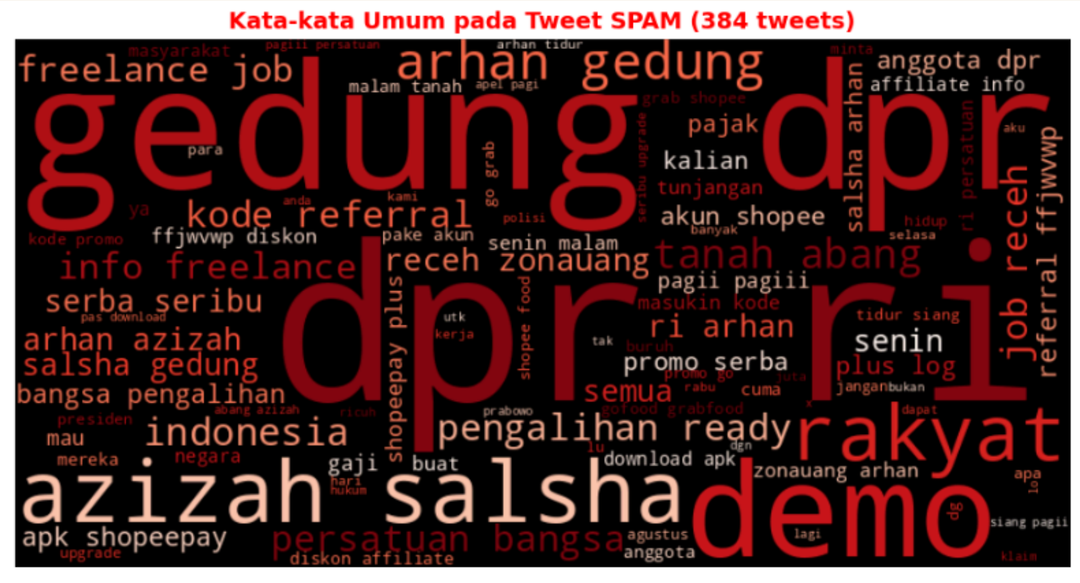
Insight: Model menunjukkan kepercayaan tinggi saat mengklasifikasikan spam, dan kepercayaan rendah untuk non-spam, yang mengindikasikan performa yang baik.



VISUALISASI POLA KATA

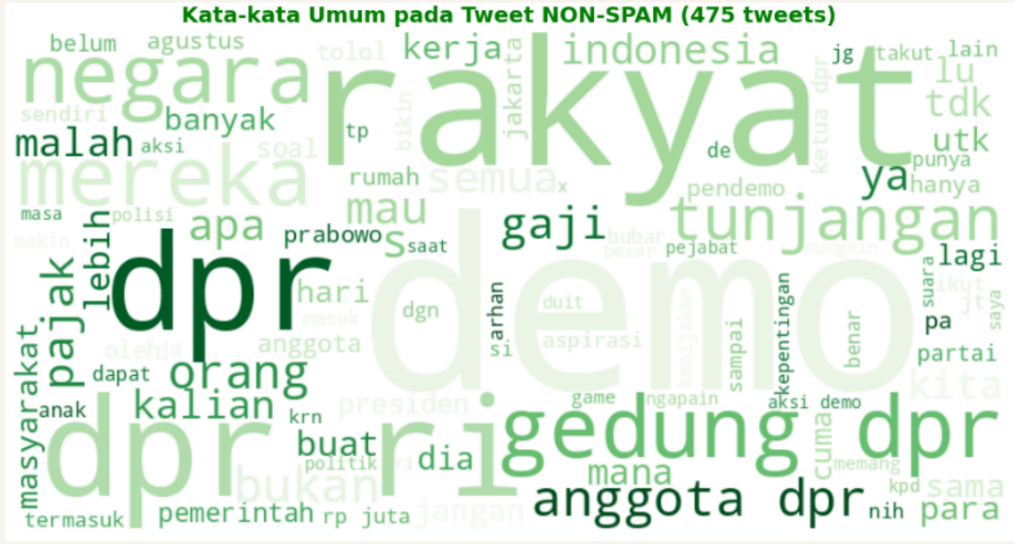
Tweet SPAM

Didominasi oleh kata-kata terkait promosi, nama orang yang tidak relevan, dan hashtag acak.



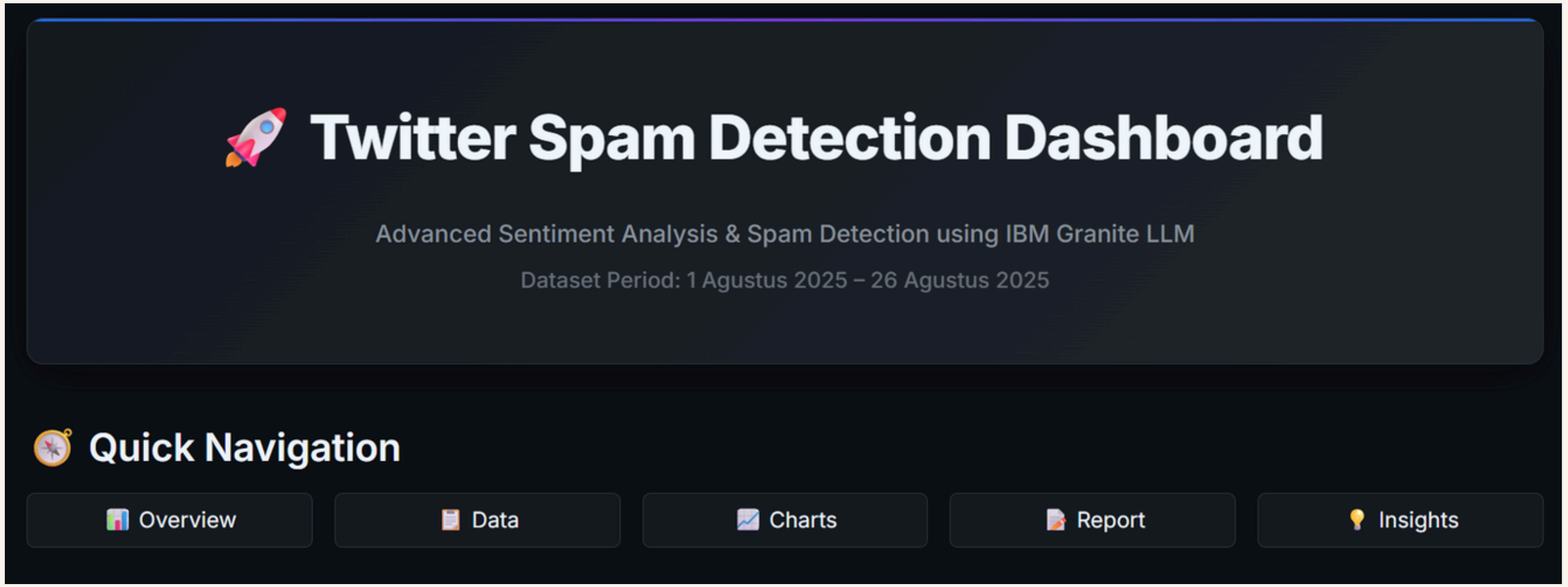
Tweet BUKAN SPAM

Didominasi oleh kata-kata yang relevan dengan topik, seperti "rakyat", "anggota", "negara", dan "tunjangan".



DASHBOARD STREAMLIT

- Aplikasi Web Interaktif: Proyek ini telah di-deploy sebagai aplikasi web menggunakan Streamlit.
- Fungsionalitas: Memungkinkan pengguna memasukkan melihat secara langsung dan mendapatkan hasil klasifikasi spam dari tanggal 1 Agustus 2025 sampai 26 Agustus 2025.
- URL Aplikasi: https://twitter-spam-detection.streamlit.app/



CONCLUSION

Kesimpulan

- Proporsi Spam: Ditemukan 44.7% tweet tergolong spam, menunjukkan adanya gangguan informasi yang signifikan dalam topik ini.
- Karakteristik Spam: Spam pada topik "Demo DPR RI" didominasi oleh promosi produk (e-commerce), clickbait dengan nama figur publik yang tidak relevan, dan penyebaran link pendek.
- Efektivitas AI: Model IBM Granite LLM terbukti efektif tidak hanya dalam mengklasifikasikan spam, tetapi juga memberikan penjelasan yang logis sebagai dasar keputusannya, yang sangat berguna untuk analisis lebih lanjut.

Rekomendasi

- Analisis Sentimen: Melakukan klasifikasi sentimen (positif, negatif, netral) pada tweet yang bukan spam untuk memahami opini publik secara lebih mendalam.
- Analisis Tren Waktu: Menganalisis kapan volume spam meningkat atau menurun selama periode demo untuk mengidentifikasi pemicunya.
- Pengembangan Model Lanjutan: Melatih model dengan kriteria yang lebih spesifik untuk mendeteksi jenis misinformasi atau disinformasi lainnya.

THANK YOU