LAPORAN ANTARA

Judul Proyek: Analisis Dataset RokYou dan Kalkulator Keamanan Password

Diajukan Oleh:
- Farras Dzaky Faisa
- Muhammad Bintang
- Muhammad Bukhori
- Muhammad Virgawan

Rekayasa Perangkat Lunak 2

Universitas Gunadarma

2025

I. Pendahuluan

1.1 Latar Belakang Proyek

Proyek pengembangan perangkat lunak "Password Entropy Calculator" ini dilatarbelakangi oleh tingginya insiden kebocoran data yang disebabkan oleh penggunaan password yang lemah. Berdasarkan studi awal, banyak pengguna yang masih menggunakan kombinasi karakter sederhana. Proyek ini bertujuan membangun aplikasi web yang mampu mengukur kekuatan password menggunakan metode Shannon Entropy dan memberikan edukasi berbasis data dari dataset rockyou.txt.

1.2 Tujuan Pelaporan

Laporan antara ini disusun untuk:

1. Melaporkan progres pengerjaan proyek dari tahap analisis data hingga pemodelan algoritma.
2. Memastikan pengembangan berjalan sesuai dengan metodologi CRISP-DM yang telah direncanakan.
3. Mendokumentasikan hasil analisis data sementara sebelum masuk ke tahap pembuatan antarmuka (user interface).

1.3 Ruang Lingkup Pengerjaan

Pada tahap pelaporan ini, fokus pengerjaan mencakup:

1. Akuisisi dan eksplorasi dataset rockyou.txt.
2. Proses pembersihan data (data cleaning) untuk menghapus data duplikat dan corrupt.
3. Implementasi rumus matematika Entropy ke dalam kode program Python.
4. Evaluasi awal distribusi kekuatan password pada dataset.

II. Realisasi Anggaran

2.1 Timeline dan Capaian Kerja

Tim pengembang telah menyelesaikan fase Data Understanding, Data Preparation, dan sebagian fase Modeling. Berikut adalah rincian realisasi kegiatan:

Kegiatan Utama	Status	Keterangan
Studi Literatur Entropy	Selesai	Memahami rumus $H = L \times log_2 N$.
Pengumpulan Dataset	Selesai	Dataset rockyou.txt (14 juta baris) berhasil diunduh.
Pembersihan Data	Selesai	Reduksi data dari 14,3 juta menjadi 13,1 juta data valid.
Coding Algoritma Dasar	Selesai	Fungsi Python untuk menghitung entropy telah dibuat.
Desain Wireframe UI	Selesai	Rancangan antarmuka Streamlit telah disetujui.
Implementasi UI Web	Selesai	Deployment project menggunakan Streamlit

2.2 Penggunaan Sumber Daya (Tools)

Dalam pelaksanaan tahap ini, perangkat lunak yang digunakan meliputi:

1. Google Colaboratory: Digunakan untuk eksekusi kode Python berat saat memproses 14 juta data password.
2. Library Python: Pandas (manipulasi data), NumPy (perhitungan numerik), dan Matplotlib (visualisasi grafik).
3. VS Code: Digunakan untuk merancang struktur awal aplikasi web.

III. Hasil Pelaksanaan

3.1 Pengumpulan dan Pemahaman Data (Data Understanding)

Langkah pertama adalah membedah dataset rockyou.txt. Data ini merupakan kumpulan password hasil kebocoran situs RockYou pada tahun 2009.

- Total Data Mentah: 14.341.564 baris password.
- Karakteristik: Data berisi teks polos (plaintext) dengan variasi karakter yang sangat beragam, termasuk karakter non-ASCII yang perlu ditangani.

```
def is_valid_password(pw):
    return re.fullmatch(rf"[{ALLOWED_CHARS}]+", pw) and len(pw) <= 12

with open("rockyou.txt", "r", encoding="utf-8", errors="ignore") as f:
    raw_passwords = f.read().splitlines()

```

3.2 Persiapan dan Pembersihan Data (Data Preparation)

Mengingat data mentah mengandung banyak noise, dilakukan proses cleaning menggunakan Python dengan langkah-langkah sebagai berikut:

1. Handling Encoding: Membaca file dengan mode ignore error untuk melewati karakter yang rusak.
2. Duplikasi: Menghapus password yang muncul berulang kali untuk menghindari bias statistik.
3. Filter Panjang: Menghapus password dengan panjang kurang dari 4 karakter atau lebih dari 50 karakter, karena dianggap tidak relevan untuk analisis kekuatan modern.

Hasil Akhir Data Preparation:

1. Data Awal: 14.341.564
2. Data Dibuang (Invalid/Duplikat): 1.142.336
3. Data Valid (Siap Analisis): 13.199.228

```python
def is_valid_password(pw):
    return re.fullmatch(rf"[{ALLOWED_CHARS}]+", pw) and len(pw) <= 12
```

3.3 Implementasi Model Algoritma Entropy (Modeling)

Inti dari aplikasi ini adalah fungsi perhitungan kekuatan password. Kami telah berhasil mengonversi rumus matematika Shannon Entropy menjadi fungsi Python. Variabel yang dihitung oleh sistem adalah:

1. L (Length): Jumlah karakter dalam password.
2. N (Charset): Ukuran himpunan karakter (26 huruf kecil, 26 huruf besar, 10 angka, 32 simbol).

Kode program yang telah berhasil diuji coba adalah sebagai berikut:

```python
def calculate_entropy(password):
    charset = 0
    if any(c.islower() for c in password): charset += 26
    if any(c.isupper() for c in password): charset += 26
    if any(c.isdigit() for c in password): charset += 10
    # simbol simbol yang diizinkan
    symbols = "!@#$%^&*()_-=><.,:'+"
    if any(c in symbols for c in password): charset += len(symbols)
    if charset == 0: return 0
    return round(len(password) * math.log2(charset), 2)

```

3.4 Analisis Statistik Awal

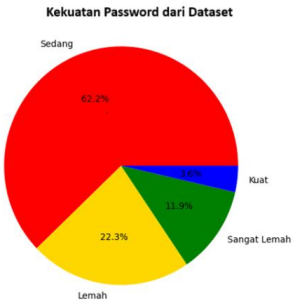
Berdasarkan algoritma di atas, kami telah melakukan running program terhadap 13 juta password. Hasil sementara menunjukkan tren yang mengkhawatirkan:

- Sangat Lemah: 11,9% (Contoh: "123456")
- Lemah: 22,3% (Contoh: "princess")
- Sedang: 62,2%
- Kuat: Hanya 3,6%

Data ini nantinya akan divisualisasikan dalam bentuk Pie Chart di halaman depan aplikasi

Studi Kasus: Analisis Dataset `rockyou.txt`

Sebagai bagian dari penelitian, dilakukan analisis terhadap **13 juta password** dari dataset `rockyou.txt`



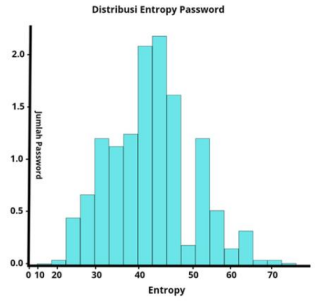


Diagram ini menunjukkan distribusi kekuatan dari 13 juta password yang bocor dari dataset rockyou.txt. Temuan utamanya adalah:

1.Dominasi Kategori Sedang: Sebanyak 62,2% password hanya memenuhi standar keamanan minimal.

2.Zona Bahaya: Jika digabungkan, kategori Lemah (22,3%) dan Sangat Lemah (11,9%) mencakup lebih dari sepertiga total password.

3.Sangat Langka: Password yang benar-benar Kuat hanya berjumlah 3,6%.

Hasil ini menunjukkan betapa pentingnya untuk secara aktif memeriksa dan meningkatkan kekuatan password

Grafik ini menunjukkan di mana kekuatan 13 juta password terdistribusi. Puncak tertinggi berada di skor Entropi sekitar 40-45 bit. Ini berarti mayoritas password di dunia nyata hanya berada di level "Sedang" tidak terlalu lemah, tetapi juga sama sekali tidak kuat. Sangat sedikit password (batang di ujung kanan) yang benar-benar mencapai tingkat keamanan "Kuat".

IV. Kendala dan Solusi

4.1 Kendala Teknis

1. Keterbatasan Memori (RAM): Saat memproses dataset sebesar 13 juta baris secara lokal, komputer mengalami crash (Memory Error).
2. Variasi Karakter: Terdapat password dengan karakter asing (bahasa non-latin) yang menyebabkan error saat perhitungan nilai ASCII.

4.2 Solusi Yang Diterapkan

1. Migrasi ke Cloud: Proses analisis data dipindahkan dari komputer lokal ke lingkungan Google Colaboratory yang menyediakan RAM lebih besar secara cloud.
2. Validasi Regex: Menambahkan filter Regular Expression untuk memastikan hanya password dengan karakter standar (Latin, Angka, Simbol umum) yang diproses.

VI. Penutup

Secara keseluruhan, proyek pengembangan perangkat lunak ini berjalan sesuai jadwal (on-track). Tahap terberat yaitu pemrosesan data raksasa (Big Data) telah berhasil dilalui. Fokus selanjutnya adalah pada aspek estetika dan fungsionalitas antarmuka pengguna (Frontend). Tim pengembang optimis aplikasi dapat diselesaikan tepat waktu pada akhir periode proyek.