Scoring System TelePhisDebate

1) Triage Classification

Komponen	Nilai / Aturan
SAFE	risk_score = 0 dan (all_whitelisted atau tidak ada URL)
LOW_RISK	0 < risk_score < 30
HIGH_RISK	risk_score >= 30
skip_llm	true hanya pada kondisi SAFE di atas
Batas skor	Risk score di-clamp ke 0..100

Tabel ini menentukan gerbang awal sistem. Jika pesan benar-benar bersih (risk_score = 0) dan URL aman, pesan langsung selesai di triage tanpa biaya token LLM. Jika tidak, pesan diteruskan ke tahap berikutnya.

2) Triage Score Weights (Core)

Flag	Bobot
blacklisted_domain	+50
phishing_keywords	+20
authority_impersonation	+20
suspicious_tld	+15
urgency_keywords	+15
shortened_url	+10
shortened_url_expand_failed	+15
caps_lock_abuse	+10
excessive_punctuation	+5
Bonus shortener -> whitelisted	-10

Bobot core mewakili indikator konten/URL yang paling kuat. Nilai +50 pada blacklisted_domain membuat sistem cepat menaikkan risiko untuk domain berbahaya, sementara bonus −10 mencegah short URL resmi ikut dianggap berisiko tinggi.

3) Triage Score Weights (Behavioral)

Anomali	Bobot Dasar
time_anomaly	+10
length_anomaly	+10
first_time_url	+10
emoji_anomaly	+5

Rumus kontribusi anomali:

```
risk += int(weight * deviation_score)
```

Anomali perilaku tidak dihitung biner, tetapi diskalakan dengan deviation_score. Artinya, semakin jauh pola pesan dari baseline normal user, semakin besar kontribusi skornya.

4) Single-Shot Escalation Rules

Kondisi	Keputusan
`classification = PHISHING`	Selalu eskalasi ke MAD
`classification = SUSPICIOUS`	Selalu eskalasi ke MAD
`classification = SAFE` dan `confidence >= 0.90`	Finalize di single-shot
`classification = SAFE` dan `confidence < 0.70`	Eskalasi ke MAD
`triage_risk >= 50` dan `confidence < 0.80`	Eskalasi ke MAD
Selain kondisi eskalasi	Finalize di single-shot

Single-shot berfungsi sebagai router, bukan hakim final untuk kasus berisiko. PHISHING dan SUSPICIOUS wajib masuk MAD3 agar keputusan akhir lebih robust. Hanya SAFE dengan confidence tinggi yang boleh diputus di tahap ini.

5) MAD3 Scoring

5.1 Bobot Agen

Agent	Weight
`content_analyzer`	1.0
`security_validator`	1.5
`social_context`	1.0

`security_validator` diberi bobot terbesar karena lebih menekankan bukti objektif (indikator keamanan) dibanding opini kontekstual.

5.2 Threshold Keputusan

Kondisi `phishing_prob`	Keputusan
`>= 0.65`	`PHISHING`
`<= 0.35`	`LEGITIMATE` (dinormalisasi jadi `SAFE`)
di antara keduanya	`SUSPICIOUS`

Zona tengah (`0.35 < phishing_prob < 0.65`) sengaja dipetakan ke `SUSPICIOUS` untuk menahan keputusan ekstrem ketika bukti belum cukup kuat.

5.3 Rule Consensus

Kondisi	Consensus
Semua agent satu stance	Unanimous
Minimal 2 dari 3 agent sama stance + rata-rata confidence `>= 0.75`	Strong majority

Consensus dipakai untuk stabilitas antar-agent. Jika tidak ada kesepakatan kuat, aggregator tetap memakai weighted score agar sistem tetap menghasilkan keputusan deterministik.

6) Pipeline Action Mapping

Final Classification	Confidence	Action
SAFE	-	none
SUSPICIOUS	>= 0.60	warn
SUSPICIOUS	< 0.60	flag_review
PHISHING	berapa pun	flag_review

Action mapping memisahkan keputusan klasifikasi dari tindakan operasional bot. Pada implementasi saat ini, PHISHING tidak auto-delete, tetapi diarahkan ke review admin agar jejak audit tetap aman.

7) Runtime Controls (Env)

Variable	Default	Fungsi
MAD_MAX_ROUNDS	2	Batas jumlah ronde debat
MAD_EARLY_TERMINATION	true	Stop dini jika konsensus tercapai
MAD_MAX_TOTAL_TIME_MS	kosong (None)	Batas waktu total debat (opsional)

Tabel ini mengontrol trade-off akurasi vs latency. Semakin besar MAD_MAX_ROUNDS, potensi kualitas debat naik, tetapi waktu proses dan token juga ikut meningkat.

8) Rumus Ringkas

$$R_t = \sum_i w_i \; + \; \sum_j \lfloor b_j d_j \rfloor \; + \; \beta$$

$$p_\phi = \frac{S_\phi}{S_\phi + S_\ell}$$

Keterangan simbol:

- R_t: total skor risiko triage.
- w_i: bobot indikator core ke-i.
- b_j: bobot anomali perilaku ke-j.
- d_j: nilai deviasi anomali ke-j.
- β: bonus shortener ke domain whitelist (nilai negatif).
- p_ϕ: probabilitas phishing hasil agregasi MAD3.
- S_ϕ: skor berbobot untuk stance PHISHING.
- S_ℓ: skor berbobot untuk stance LEGITIMATE.

9) Asumsi Perhitungan

9.1 Asumsi Operasional

Aspek	Asumsi
Penjumlahan triage	Semua flag aktif dijumlahkan (aditif).
Rentang deviasi	`deviation_score` berada pada `0.0..1.0`.
Kontribusi anomali	`int(weight * deviation_score)`.
Batas skor triage	Skor akhir dibatasi ke `0..100`.
Normalisasi MAD	`p_{\phi}` dihitung dari skor PHISHING vs LEGITIMATE; `SUSPICIOUS` netral.
Confidence MAD	`max(p_{\phi}, 1 - p_{\phi})`.
Validitas contoh	Contoh di dokumen ini ilustratif, bukan log produksi langsung.

9.2 Matriks Keputusan Cepat

Tahap	Kondisi Kunci	Output
Triage	`R_t = 0` dan URL aman	`SAFE`, `skip_llm = true`
Triage	`0 < R_t < 30`	`LOW_RISK`
Triage	`R_t >= 30`	`HIGH_RISK`
Single-Shot	`PHISHING` atau `SUSPICIOUS`	Wajib ke MAD3
Single-Shot	`SAFE` dan `confidence >= 0.90`	Finalize `SAFE`
MAD3	`p_{\phi} >= 0.65`	`PHISHING`
MAD3	`p_{\phi} <= 0.35`	`LEGITIMATE -> SAFE`
MAD3	`0.35 < p_{\phi} < 0.65`	`SUSPICIOUS`
Action	`SUSPICIOUS` dan `conf >= 0.60`	`warn`
Action	`PHISHING` (berapa pun confidence)	`flag_review`

10) Contoh Perhitungan Kasus

10.1 Contoh Kasus PHISHING

Asumsi evidence terdeteksi: - `blacklisted_domain` = +50
- `phishing_keywords` = +20
- `urgency_keywords` = +15
- `authority_impersonation` = +20
- `caps_lock_abuse` = +10

Perhitungan triage:

$$R_t = 50 + 20 + 15 + 20 + 10 = 115 \Rightarrow \text{clamp ke } 100$$

Hasil: - Triage = `HIGH_RISK` (karena `R_t >= 30`)
- Lanjut ke Single-Shot
- Single-Shot mengeluarkan `PHISHING` -> wajib eskalasi ke MAD3

Asumsi output MAD3: - Content Analyzer: `PHISHING`, confidence `0.80`
- Security Validator: `PHISHING`, confidence `0.90`
- Social Context: `SUSPICIOUS`, confidence `0.70`

Skor MAD3:

$$S_\phi = (1.0 \cdot 0.80) + (1.5 \cdot 0.90) = 2.15$$

$$S_\ell = 0$$

$$p_\phi = \frac{2.15}{2.15 + 0} = 1.00$$

Keputusan akhir: - `PHISHING` (karena `p_{\phi} \ge 0.65`)
- Action = `flag_review`

10.2 Contoh Kasus SUSPICIOUS

Asumsi evidence terdeteksi: - `shortened_url` = +10
- `excessive_punctuation` = +5
- `emoji_anomaly` dengan `d = 0.60` -> `int(5 * 0.60) = +3`

Perhitungan triage:

$$R_t = 10 + 5 + 3 = 18$$

Hasil: - Triage = `LOW_RISK` (karena `0 < R_t < 30`)
- Lanjut ke Single-Shot
- Single-Shot memberi `SUSPICIOUS` (wajib eskalasi ke MAD3)

Asumsi output MAD3: - Content Analyzer: `SUSPICIOUS`, confidence `0.70`
- Security Validator: `PHISHING`, confidence `0.62`
- Social Context: `LEGITIMATE`, confidence `0.60`

Skor MAD3:

$$S_\phi = 1.5 \cdot 0.62 = 0.93$$

$$S_\ell = 1.0 \cdot 0.60 = 0.60$$

$$p_\phi = \frac{0.93}{0.93 + 0.60} \approx 0.608$$

Keputusan akhir: - `SUSPICIOUS` (karena `0.35 < p_{\phi} < 0.65`)
- Confidence MAD3 = `max(0.608, 0.392) = 0.608`
- Action = `warn` (karena confidence >= `0.60`)

10.3 Contoh Kasus SAFE

Asumsi evidence: - Tidak ada URL berisiko - Tidak ada red flag/anomali aktif

Perhitungan triage:

$$R_t = 0$$

Hasil: - Triage = `SAFE`
- `skip_llm = true`
- Tidak masuk Single-Shot dan MAD3
- Action akhir = `none`

10.4 Contoh SAFE dengan Shortener Whitelist

Asumsi evidence: - URL shortener terdeteksi: `shortened_url = +10` - Hasil expand menuju domain whitelist: bonus `−10`

Perhitungan triage:

$$R_t = 10 + (-10) = 0$$

Hasil: - Tetap `SAFE` (selama kondisi `all_whitelisted` terpenuhi)
- `skip_llm = true`
- Action akhir = `none`