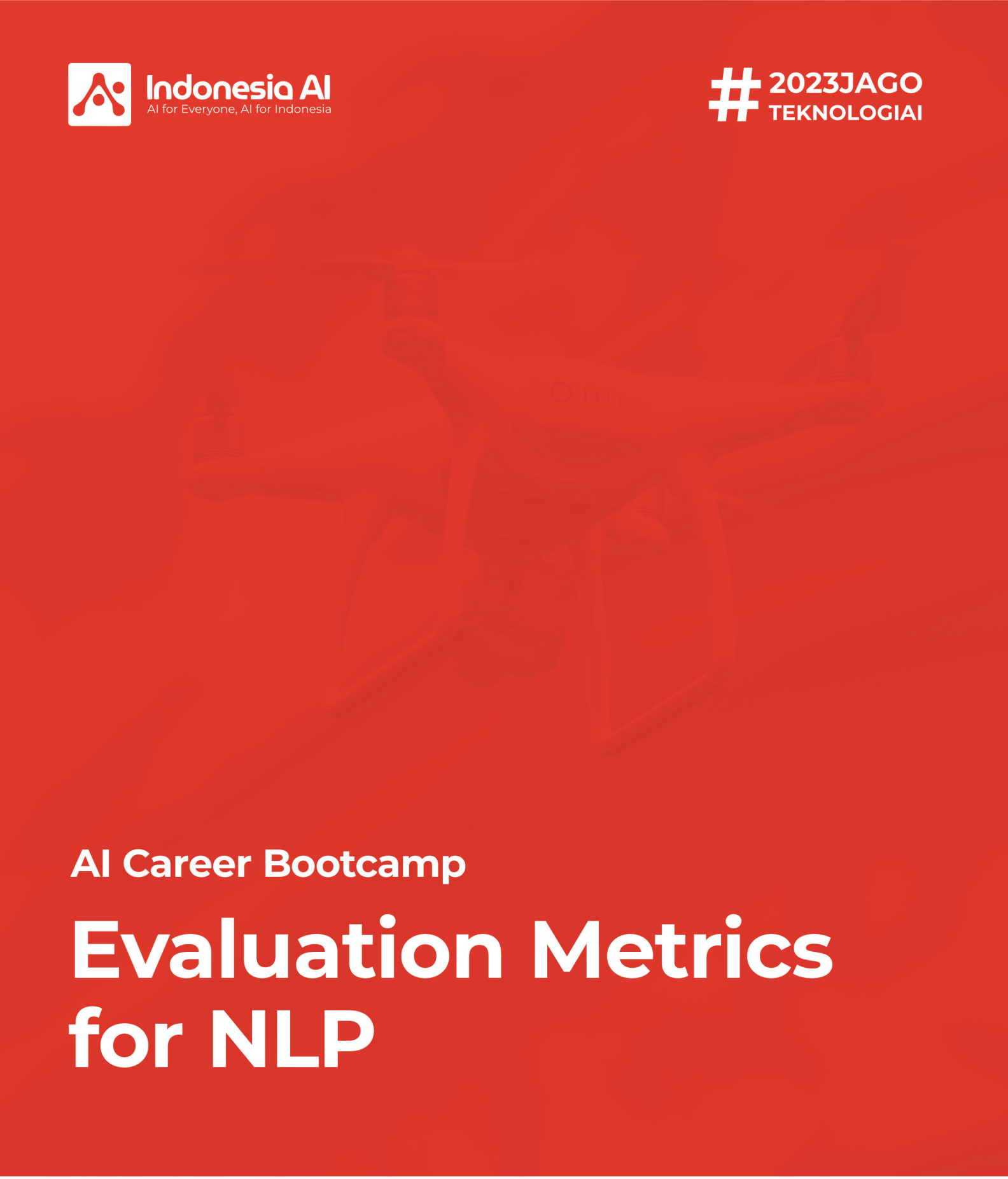
Indonesia AI
AI for Everyone, AI for Indonesia

#2023JAGOTEKNOLOGIAI



AI Career Bootcamp

Evaluation Metrics for NLP

Indonesia AI (PT. Teknologi Artifisial Indonesia)

Pembukaan

Guide Book ini memuat beberapa informasi-informasi utama yang akan disampaikan oleh mentor di program AI Career Bootcamp yang bisa dijadikan pegangan para students untuk mempersiapkan diri sebelum sesi Live Class berlangsung.

Deskripsi

Evaluasi metrik untuk Pemrosesan Bahasa Alam (NLP) digunakan untuk mengukur kinerja model atau sistem dalam tugas-tugas NLP seperti klasifikasi teks, pemodelan bahasa, penerjemahan mesin, dan lainnya. Beberapa metrik evaluasi yang umum digunakan dalam NLP adalah sebagai berikut:

1. Akurasi (Accuracy): Metrik ini mengukur persentase kebenaran prediksi secara keseluruhan. Akurasi menghitung jumlah prediksi yang benar dibagi dengan jumlah total sampel. Meskipun metrik ini penting, perlu diingat bahwa akurasi tidak memberikan informasi tentang kinerja model pada kelas-kelas yang tidak seimbang atau saat terjadi kesalahan tertentu.

2. Presisi (Precision): Presisi mengukur sejauh mana prediksi positif benar-benar relevan. Presisi dihitung dengan membagi jumlah prediksi positif yang benar dengan jumlah prediksi positif secara keseluruhan. Presisi berguna ketika fokus pada mengurangi jumlah kesalahan positif palsu.

3. Recall (Recall): Recall mengukur sejauh mana model dapat mengidentifikasi semua contoh positif yang ada. Recall dihitung dengan membagi jumlah prediksi positif yang benar dengan jumlah contoh positif secara keseluruhan. Recall penting ketika fokus pada mengurangi jumlah kesalahan negatif palsu.

4. F1-Score: F1-score adalah penggabungan presisi dan recall yang memberikan ukuran yang seimbang tentang kinerja model. F1-score dihitung sebagai harmonic mean dari presisi dan recall. F1-score berguna ketika keseimbangan antara presisi dan recall diinginkan.

Selain metrik di atas, terdapat juga metrik lain seperti Area Under the Curve (AUC) untuk kurva Receiver Operating Characteristic (ROC) yang digunakan dalam klasifikasi biner, Mean Average

Precision (MAP) untuk evaluasi peringkat, dan Bleu Score untuk evaluasi sistem penerjemahan mesin. Pemilihan metrik evaluasi yang tepat tergantung pada tugas NLP yang dilakukan dan tujuan spesifik yang ingin dicapai.

Mengenal Bleu Score

Bleu score (Bilingual Evaluation Understudy score) adalah metrik evaluasi yang umum digunakan dalam sistem penerjemahan mesin untuk mengukur sejauh mana terjemahan sistem mendekati referensi manusia yang benar. Bleu score mengukur kesamaan antara terjemahan sistem dan referensi manusia dengan membandingkan n-gram (urutan kata) yang muncul dalam keduanya.

Metrik Bleu score berkisar antara 0 hingga 1, di mana semakin tinggi skornya menunjukkan kesamaan yang lebih tinggi antara terjemahan sistem dan referensi manusia. Skor Bleu dihitung dengan menghitung rasio n-gram yang cocok antara terjemahan sistem dan referensi, serta memperhitungkan penyesuaian brevity penalty untuk memperhitungkan terjemahan yang terlalu pendek.

Bleu score umumnya digunakan untuk membandingkan hasil terjemahan mesin dari berbagai sistem atau model, dan memberikan pengukuran kualitas yang relatif. Meskipun Bleu score memberikan gambaran kasar tentang kualitas terjemahan sistem, namun juga memiliki keterbatasan, seperti tidak mempertimbangkan konteks global atau aspek semantik. Oleh karena itu, Bleu score sering digunakan bersama dengan metrik evaluasi lainnya untuk memberikan pemahaman yang lebih komprehensif tentang kualitas terjemahan mesin.