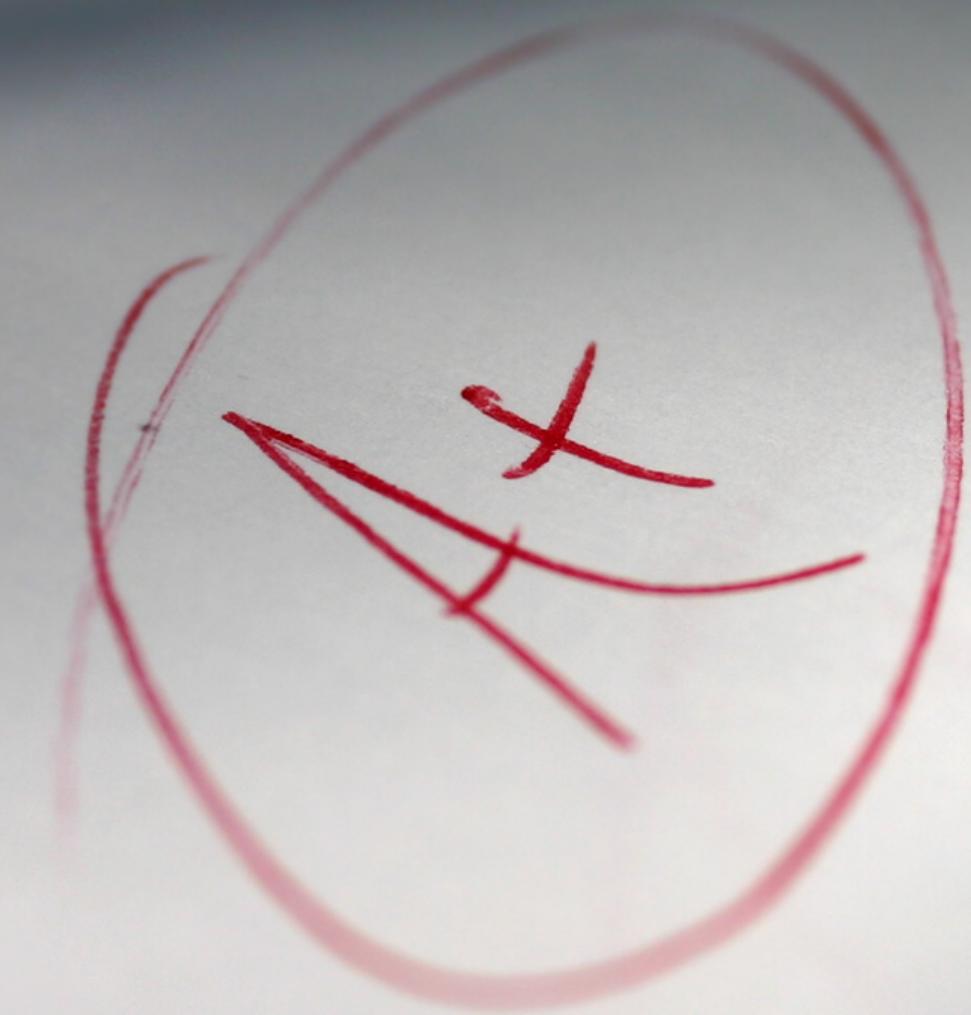


Evaluation Metrics for NLP



OBJECTIVE & OUTLINE

Proprietary document of Indonesia AI 2023



Evaluation Metrics for NLP

Objektif: Memahami konsep dari Evaluation Metrics dalam NLP

Outline:

1. Evaluation Metrics
2. Evaluation Metrics Dasar
3. Evaluation Metrics Lanjut

Apa itu Evaluation Metrics?

EVALUATION METRICS

Proprietary document of Indonesia AI 2023



Evaluation Metrics dalam NLP adalah ukuran atau **metrik** yang digunakan untuk **meng evaluasi kinerja** model dari sistem NLP.

EVALUATION METRICS

Proprietary document of Indonesia AI 2023

Tujuan

mengukur sejauh mana model atau sistem dapat menghasilkan hasil yang akurat dan relevan dalam tugas-tugas NLP seperti klasifikasi teks, penerjemahan mesin, atau analisis sentimen.



Indonesia AI

EVALUATION METRICS

Proprietary document of Indonesia AI 2023

Metrics dan kasus:



- Akurasi: Klasifikasi teks
- Presisi: Klasifikasi teks
- Recall: Klasifikasi teks
- F1-Score: Klasifikasi teks
- BLEU: Mesin translasi
- ROUGE: Summarizer
- Perplexity: Language Model

— Any question guys ~

Evaluation Metrics Dasar

EVALUATION METRICS DASAR

Proprietary document of Indonesia AI 2023



- Akurasi
- Presisi
- Recall
- F1-Score

EVALUATION METRICS DASAR

Proprietary document of Indonesia AI 2023

Akurasi

Metrik ini mengukur seberapa akurat model dalam memprediksi kelas yang benar. Akurasi dihitung dengan membagi jumlah prediksi yang benar dengan jumlah total data.

Akurasi = (Jumlah prediksi yang benar) / (Total jumlah contoh)

Contoh = $80/100 = 0,8$ atau 80%



EVALUATION METRICS DASAR

Proprietary document of Indonesia AI 2023

Akurasi

content	label	prediction
kapan maen ke rumah mileakuu	normal	normal
kalo polisi tidak bisa hukum wanita kafir lknat itu s	normal	normal
kepala bin gimana ga bisa deteksi gin	normal	spam
masyaallah keren,normal	normal	spam
andra prabu yg bencong nyinyir itu lu yg bisa y cun	normal	normal
www jualdokumen com jual ijazah sd smp sma ijas	spam	normal
cantik banget gilaaaa,normal	normal	normal
nah telah bahas apa itu wabot sy akan share kan ju	spam	spam
trus gw hrs bilang wow emejing gitu	normal	normal
aneh abis perkosa msh mau duduk sama teman la	normal	normal

$$\text{Akurasi} = 7/10 = 0.7$$

EVALUATION METRICS DASAR

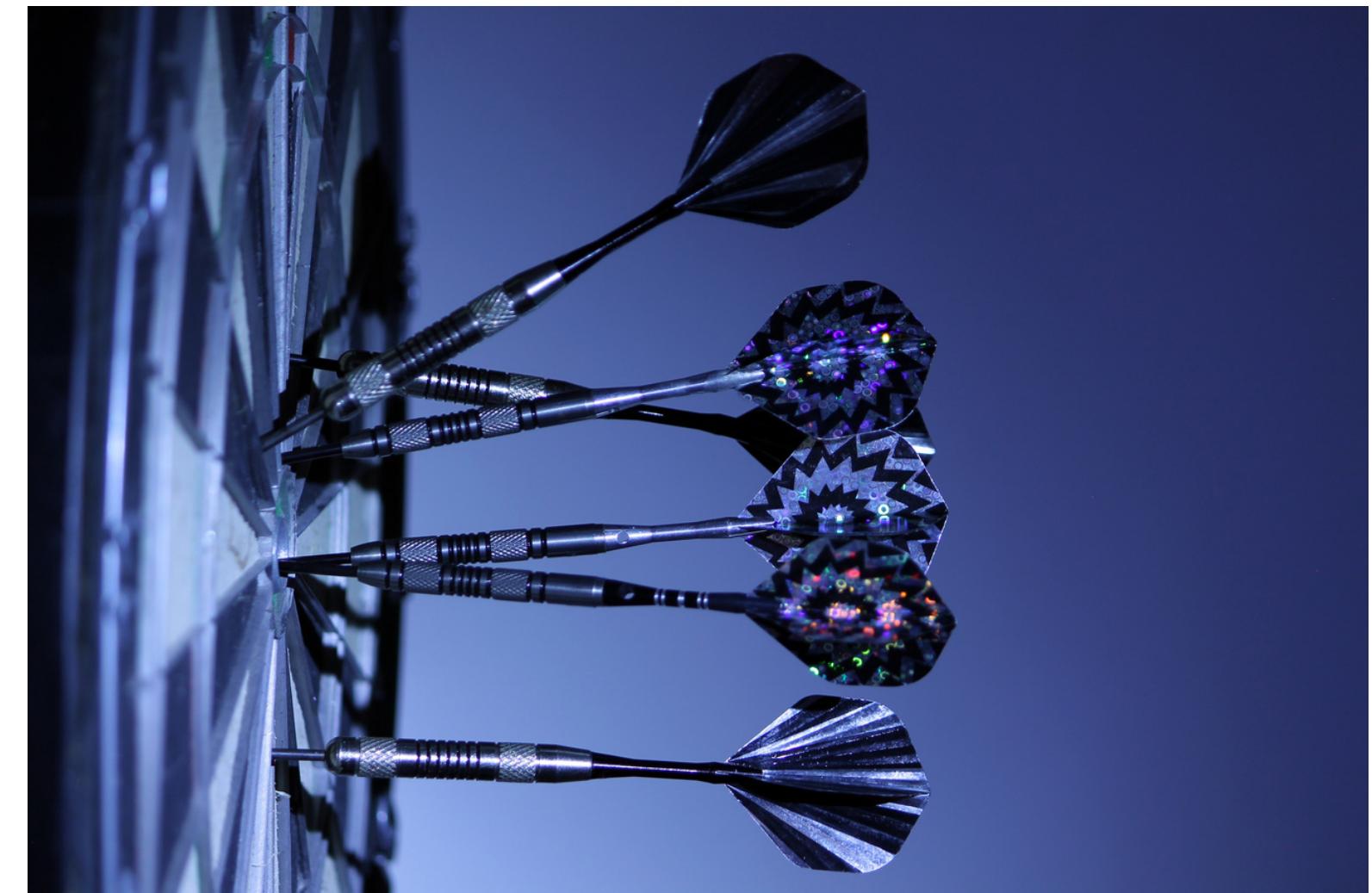
Proprietary document of Indonesia AI 2023

Presisi

Presisi mengukur seberapa akurat model dalam mengidentifikasi kelas positif. Metrik ini menghitung rasio prediksi yang benar positif dibandingkan dengan total prediksi positif.

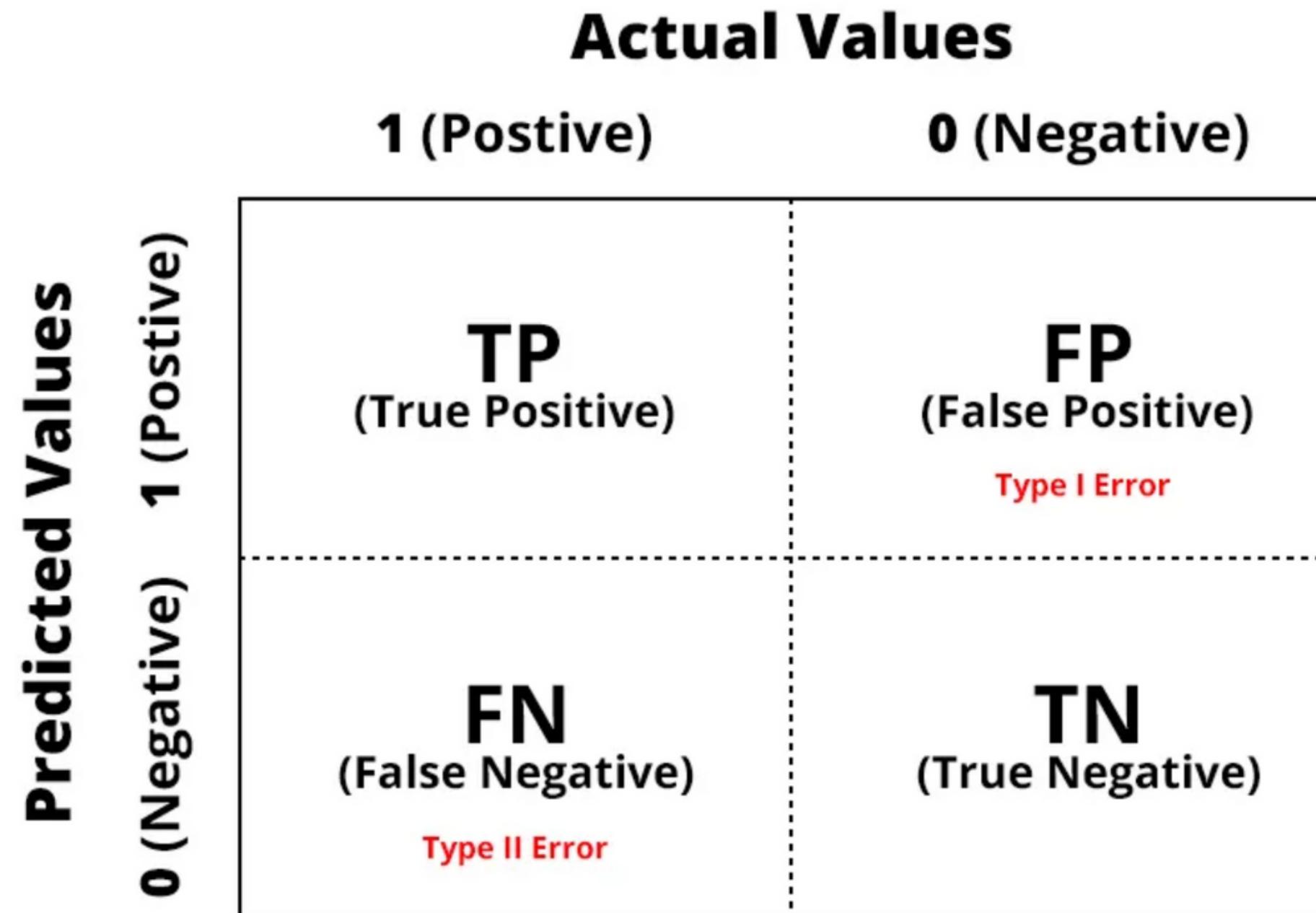
Presisi = True Positives / (True Positives + False Positives)

Contoh = $40 / (40 + 10) = 0,8$ atau 80%.



EVALUATION METRICS DASAR

Proprietary document of Indonesia AI 2023



EVALUATION METRICS

DASAR

Proprietary document of Indonesia AI 2023

- True Positive (TP)
 - Merupakan data positif yang diprediksi benar. Contohnya, teks spam (class 1) dan dari model yang dibuat memprediksi teks tersebut sebagai spam (class 1).
- True Negative (TN)
 - Merupakan data negatif yang diprediksi benar. Contohnya, teks non-spam (class 2) dan dari model yang dibuat memprediksi teks non-spam sebagai non-spam (class 2).

EVALUATION METRICS DASAR

Proprietary document of Indonesia AI 2023

- False Positive (FP) – Type I Error
 - Merupakan data negatif namun diprediksi sebagai data positif. Contohnya, teks non-spam (class 2) tetapi dari model yang telah memprediksi teks sebagai spam (class 1).
- False Negative (FN) – Type II Error
 - Merupakan data positif namun diprediksi sebagai data negatif. Contohnya, teks spam (class 1) tetapi dari model yang dibuat memprediksi teks sebagai non-spam (class 2).

EVALUATION METRICS

DASAR

Proprietary document of Indonesia AI 2023

Presisi

content	label	prediction	
kapan maen ke rumah mileakuu	normal	normal	TP
kalo polisi tidak bisa hukum wanita kafir lknat itu s	normal	normal	TP
kepala bin gimana ga bisa deteksi gin	normal	spam	FP
masyaallah keren,normal	normal	spam	FP
andra prabu yg bencong nyinyir itu lu yg bisa y cun	normal	normal	TP
www jualdokumen com jual ijazah sd smp sma ijas	spam	normal	FN
cantik banget gilaaaa,normal	normal	normal	TP
nah telah bahas apa itu wabot sy akan share kan ju	spam	spam	TN
trus gw hrs bilang wow emejing gitu	normal	normal	TP
aneh abis perkosa msh mau duduk sama teman la	normal	normal	TP

$$\text{Presisi} = \frac{6}{6+2} = 0.75$$

EVALUATION METRICS DASAR

Proprietary document of Indonesia AI 2023

Recall

Recall mengukur seberapa baik model dalam menemukan semua contoh dari kelas positif yang sebenarnya. Metrik ini menghitung rasio prediksi benar positif dibandingkan dengan total contoh positif yang sebenarnya.

Recall = $\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$

Contoh = $70 / (70 + 30) = 0,7$ atau 70%.



EVALUATION METRICS DASAR

Proprietary document of Indonesia AI 2023

Recall

content	label	prediction	
kapan maen ke rumah mileakuu	normal	normal	TP
kalo polisi tidak bisa hukum wanita kafir lknat itu s	normal	normal	TP
kepala bin gimana ga bisa deteksi gin	normal	spam	FP
masyallah keren,normal	normal	spam	FP
andra prabu yg bencong nyinyir itu lu yg bisa y cun	normal	normal	TP
www jualdokumen com jual ijazah sd smp sma ijas	spam	normal	FN
cantik banget gilaaaa,normal	normal	normal	TP
nah telah bahas apa itu wabot sy akan share kan ju	spam	spam	TN
trus gw hrs bilang wow emejing gitu	normal	normal	TP
aneh abis perkosa msh mau duduk sama teman la	normal	normal	TP

$$\text{Recall} = \frac{6}{6+1} = 0.86$$

EVALUATION METRICS DASAR

Proprietary document of Indonesia AI 2023

F1-Score

F1-score merupakan gabungan antara presisi dan recall, menggabungkan informasi tentang akurasi model secara keseluruhan. Metrik ini memberikan nilai tunggal yang mencerminkan keseimbangan antara presisi dan recall.

$$\text{F1-score} = 2 * (\text{Presisi} * \text{Recall}) / (\text{Presisi} + \text{Recall})$$

Contoh = $80/100 = 0,8$ atau 80%



EVALUATION METRICS

DASAR

Proprietary document of Indonesia AI 2023

F1-Score

content	label	prediction	
kapan maen ke rumah mileakuu	normal	normal	TP
kalo polisi tidak bisa hukum wanita kafir lknat itu s	normal	normal	TP
kepala bin gimana ga bisa deteksi gin	normal	spam	FP
masyallah keren,normal	normal	spam	FP
andra prabu yg bencong nyinyir itu lu yg bisa y cun	normal	normal	TP
www jualdokumen com jual ijazah sd smp sma ijas	spam	normal	FN
cantik banget gilaaaa,normal	normal	normal	TP
nah telah bahas apa itu wabot sy akan share kan ju	spam	spam	TN
trus gw hrs bilang wow emejing gitu	normal	normal	TP
aneh abis perkosa msh mau duduk sama teman la	normal	normal	TP

$$\begin{aligned} \text{F1-Score} &= \\ &2 * (0.75 * 0.86) \\ &/ (0.75 + 0.86) \\ &= 2 * 0.4 \\ &= 0.8 \end{aligned}$$

— Any question guys ~

Evaluation Metrics Lanjut

EVALUATION METRICS LANJUT

Proprietary document of Indonesia AI 2023



- BLEU
- ROUGE
- Perplexity

EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

BLEU (Bilingual Evaluation Understudy)

Metrik ini digunakan untuk mengevaluasi kualitas penerjemahan mesin. BLEU mengukur sejauh mana teks terjemahan yang dihasilkan oleh sistem mendekati teks referensi manusia



EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

BLEU (Bilingual Evaluation Understudy)

origin	translation	candidate
kapan maen ke rumah mileakuu	when are you going to mileakuu's	when are you going to mileakuu's
kalo polisi tidak bisa hukum wanit	If the police can't punish women v	If the police can't punish women v
kepala bin gimana ga bisa deteks	head bin how come it can't detect	head bin how come it can't detect
masyallah keren,normal	mashallah cool	mashallah cool
andra prabu yg bencong nyinyir it	Andra Prabu, the one who is a sn	Andra Prabu, the one who is a sn
www jualdokumen com jual ijazah	ww selldocuments.com selling dip	ww selldocuments.com selling dip
cantik banget gilaaaaa,normal	so freaking beautiful	so freaking beautiful
nah telah bahas apa itu wabot sy	So, I've discussed what Wabot is,	So, I've discussed what Wabot is,
trus gw hrs bilang wow emejing	then I have to say wow that's awe	then I have to say wow that's awe
aneh abis perkosa msh mau dudu	It's weird that after being raped, I	It's weird that after being raped, I

EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

BLEU (Bilingual Evaluation Understudy)

Teks translasi: "Saya suka makan nasi goreng di restoran ini."

Teks kandidat: "Aku suka makan nasi goreng di tempat ini."

EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

BLEU (Bilingual Evaluation Understudy)

1. Membagi teks referensi dan teks prediksi menjadi token-token:

Teks translasi: ["Saya", "suka", "makan", "nasi", "goreng", "di", "restoran", "ini"]

Teks kandidat: ["Aku", "suka", "makan", "nasi", "goreng", "di", "tempat", "ini"]

EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

BLEU (Bilingual Evaluation Understudy)

2. Menghitung jumlah n-gram yang cocok antara teks translasi dan teks kandidat:

Jumlah n-gram cocok ($n = 1$): 7 <- BLEU-1

Jumlah n-gram cocok ($n = 2$): 6 <- BLEU-2

Jumlah n-gram cocok ($n = 3$): 5 <- BLEU-3

Jumlah n-gram cocok ($n = 4$): 4 <- BLEU-4

EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

BLEU (Bilingual Evaluation Understudy)

3. Menghitung jumlah n-gram yang muncul dalam teks kandidat:

Jumlah n-gram kandidat ($n = 1$): 8

Jumlah n-gram kandidat ($n = 2$): 7

Jumlah n-gram kandidat ($n = 3$): 6

Jumlah n-gram kandidat ($n = 4$): 5

EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

BLEU (Bilingual Evaluation Understudy)

4. Menghitung precision:

Precision (n = 1): $7 / 8 = 0.875$

Precision (n = 2): $6 / 7 \approx 0.857$

Precision (n = 3): $5 / 6 \approx 0.833$

Precision (n = 4): $4 / 5 = 0.8$

EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023



BLEU (Bilingual Evaluation Understudy)

5. Menghitung nilai BP (brevity penalty)

Panjang teks translasi = 8

Panjang teks kandidat = 8 (sama)

BP = 1 (tidak ada penalti karena panjang terjemahan sama dengan referensi terdekat)

EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

BLEU (Bilingual Evaluation Understudy)

Cara untuk menghitung Brevity Penalty (BP) pada metrik BLEU:

1. Hitung panjang teks referensi terpanjang (c) dan panjang teks terjemahan (r).
2. Jika $r \leq c$, $BP = 1$. Ini berarti teks terjemahan memiliki panjang yang sama atau lebih pendek dari teks referensi terpanjang.
3. Jika $r > c$, $BP = \exp(1 - c/r)$. Ini memberikan penalti yang lebih tinggi ketika terjemahan lebih panjang dibandingkan dengan teks referensi terpanjang. Penalti ini dihitung sebagai perbandingan antara panjang referensi terpanjang dan panjang terjemahan.
4. Hitung nilai BP terakhir sebagai $BP = \min(1, BP)$.

Secara matematis, rumus lengkapnya adalah: $BP = \min(1, \exp(1 - c/r))$

EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023



BLEU (Bilingual Evaluation Understudy)

6. Menghitung nilai BLEU

$$\text{Skor BLEU} = \text{BP} * \exp(\sum(W_i * \log(\text{Prec}_i)))$$

Kecocokan sempurna menghasilkan skor 1,0, sedangkan ketidakcocokan sempurna menghasilkan skor 0,0.

EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

BLEU (Bilingual Evaluation Understudy)

Kelebihan:

1. Kesesuaian dengan referensi
2. Kemudahan perhitungan
3. Interpretasi relatif



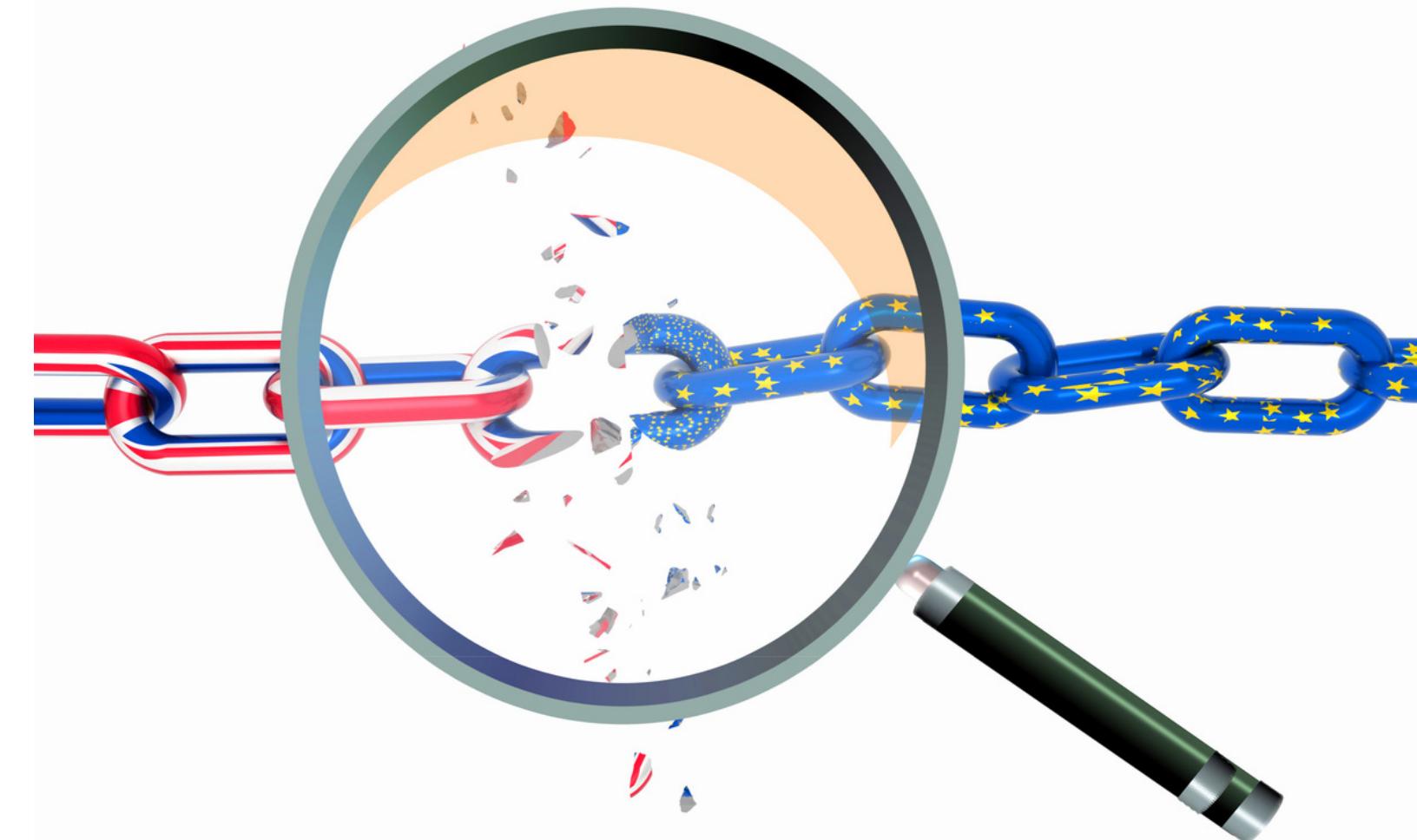
EVALUATION METRICS LANJUT

Proprietary document of Indonesia AI 2023

BLEU (Bilingual Evaluation Understudy)

Kekurangan:

1. Tidak memeriksa secara makna teks



EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Metrik digunakan untuk mengevaluasi kualitas ringkasan teks atau gisting. Metrik ini mengukur sejauh mana ringkasan yang dihasilkan oleh sistem cocok dengan ringkasan referensi manusia.



EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

origin	summarized	candidate
kapan maen ke rumah mileakuu	kapan maen ke rumah mileakuu	kapan maen ke rumah mileakuu
kalo polisi tidak bisa hukum wanit	kalo polisi tidak bisa hukum wanit	kalo polisi tidak bisa hukum wanit
kepala bin gimana ga bisa deteks	kepala bin gimana ga bisa deteks	kepala bin gimana ga bisa deteks
masyaallah keren,normal	masyaallah keren,normal	masyaallah keren,normal
andra prabu yg bencong nyinyir it	andra prabu yg bencong nyinyir it	andra prabu yg bencong nyinyir it
www jualdokumen com jual ijazah	www jualdokumen com jual ijazah	www jualdokumen com jual ijazah
cantik banget gilaaaa,normal	cantik banget gilaaaa,normal	cantik banget gilaaaa,normal
nah telah bahas apa itu wabot sy	nah telah bahas apa itu wabot sy	nah telah bahas apa itu wabot sy
trus gw hrs bilang wow emejing	trus gw hrs bilang wow emejing	trus gw hrs bilang wow emejing
aneh abis perkosa msh mau dudu	aneh abis perkosa msh mau dudu	aneh abis perkosa msh mau dudu

EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Teks referensi: "Saya suka makan"

Teks kandidat: "Suka makan ayam"

EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

1. Pembentukan n-gram

N-gram referensi: ["Saya suka", "suka makan"]

N-gram kandidat: ["Suka makan", "makan ayam"]

EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

2. Hitung intersection (tumpang tindih) antara n-gram sistem dan n-gram referensi

Intersection: 1 (n-gram "suka makan" tumpang tindih)

EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

3. Hitung precision, recall, dan F1-score ROUGE.

- Precision: $1/2 = 0.5$
- Recall: $1/2 = 0.5$
- F1-score: $2 * ((0.5 * 0.5) / (0.5 + 0.5)) = 0.5$

EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Kelebihan:

1. Melibatkan recall
2. Kemudahan perhitungan
3. Lebih sensitif terhadap keseluruhan kualitas ringkasan



EVALUATION METRICS

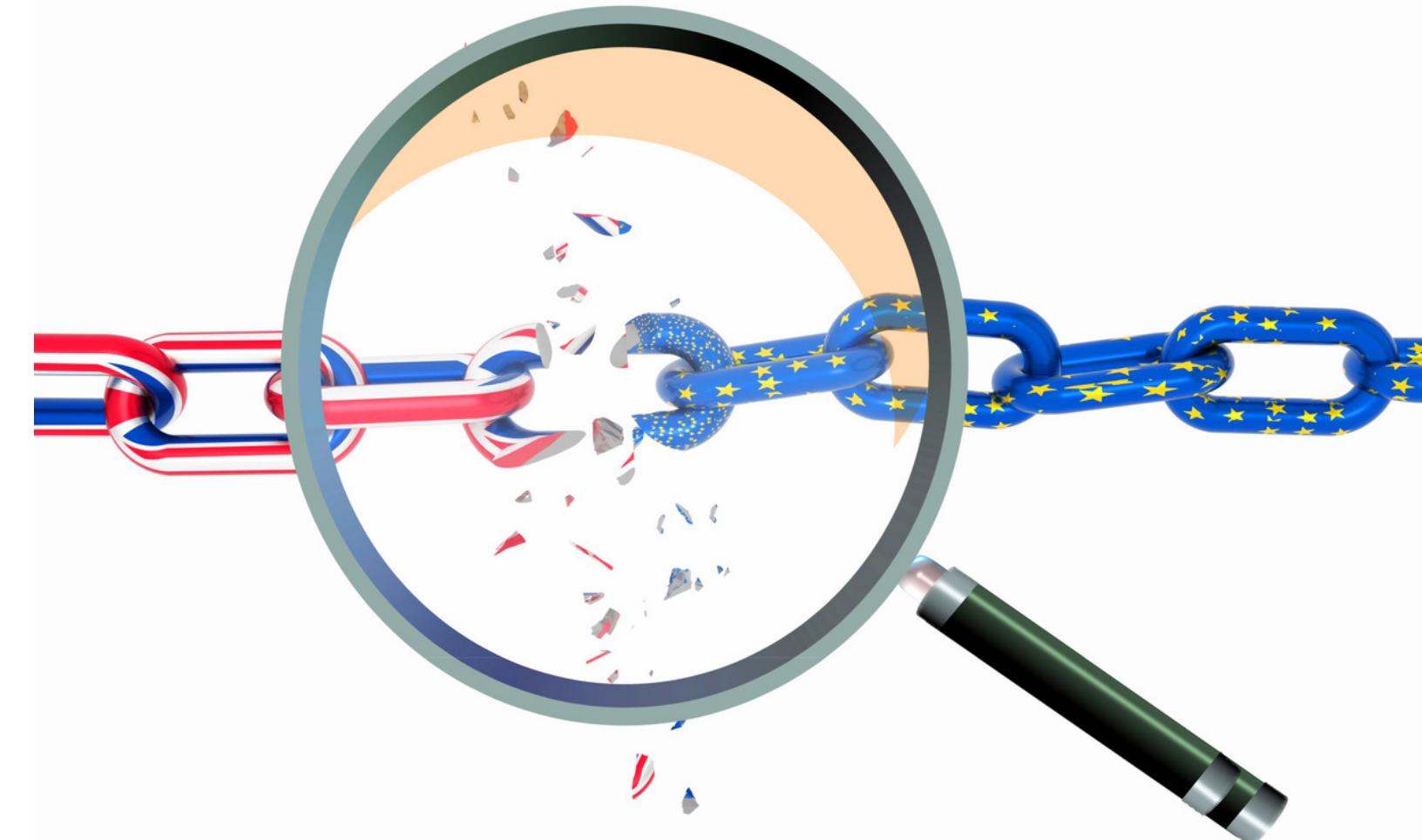
LANJUT

Proprietary document of Indonesia AI 2023

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Kekurangan:

1. Tidak memeriksa secara makna teks



EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

Perplexity

Metrik evaluasi yang umum digunakan dalam pemodelan bahasa (language modeling) dalam NLP. Metrik ini digunakan untuk mengevaluasi seberapa baik sebuah model bahasa dapat memprediksi atau menghasilkan teks yang sesuai dengan distribusi probabilitas dari data pelatihan.



EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

Perplexity

Perplexity dihitung berdasarkan **probabilitas prediksi** model untuk setiap kata dalam urutan teks. Perplexity dihitung sebagai **kebalikan logaritma** dari **probabilitas geometris rata-rata**. **Semakin rendah perplexity, semakin baik model bahasa** dalam memprediksi urutan teks yang diberikan.

$$\text{Perplexity} = \exp(-1 * (\sum(\log(P(x_i|x_1, \dots, x_{i-1})))) / t)$$

EVALUATION METRICS

LANJUT

Proprietary document of Indonesia AI 2023

Perplexity

Preplexity biasanya digunakan hanya untuk **menentukan seberapa baik model telah mempelajari set pelatihan**. Metrik lain seperti BLEU, ROUGE, dll., Digunakan pada set pengujian untuk mengukur kinerja pengujian.

— Any question guys ~

Terima Kasih!