BAB 2
LANDASAN TEORI

2.1 Aturan Penggunaan Angka dan Bilangan

Penggunaan angka dan bilangan yang tepat sangat penting untuk menjaga kejelasan dan konsistensi. Menurut pedoman umum, terdapat beberapa aturan utama yang perlu diperhatikan, yaitu [15]:

1. Bilangan yang dapat ditulis dengan satu kata sebaiknya dieja dengan huruf, misalnya ”dua” atau ”seribu”, kecuali jika digunakan dalam perincian yang berurutan. Aturan ini membantu menjaga alur narasi yang lancar dalam teks. Namun, untuk menyatakan ukuran (seperti berat, panjang, atau waktu) dan nilai (seperti uang atau persentase), penggunaan angka lebih dianjurkan untuk presisi. Misalnya, 5 kg atau Rp5.000,00.

2. Ada perlakuan khusus untuk angka di awal kalimat. Angka yang terdiri dari lebih dari satu kata tidak boleh berada di awal kalimat. Untuk mengatasinya, susunan kalimat dapat diubah atau ditambahkan kata bantu seperti ”sebanyak” atau ”sejumlah”. Contohnya, kalimat ”Sebanyak 250 orang peserta...” lebih tepat daripada ”250 orang peserta...”. Selain itu, untuk bilangan besar, kombinasi angka dan huruf diperbolehkan untuk mempermudah pembacaan, misalnya ”500 ribu”.

3. Angka memiliki fungsi spesifik dalam penulisan ilmiah dan teknis. Selain untuk data kuantitatif, angka digunakan untuk penomoran bagian karangan seperti bab dan pasal, serta sebagai bagian dari alamat, seperti nomor rumah dan jalan. Penulisan bilangan tingkat dapat menggunakan gabungan awalan ”ke-” dan angka, seperti ”abad ke-21”, atau dengan huruf, ”abad kedua puluh satu”.

2.2 *Natural Language Processing*

Natural Language Processing (NLP) adalah cabang dari kecerdasan buatan dan linguistik komputasi yang berfokus pada interaksi antara komputer dan bahasa manusia. Tujuan utamanya adalah untuk memprogram komputer agar dapat memproses, menganalisis, dan memahami data bahasa dalam jumlah besar. NLP

juga digunakan sebagai alat untuk membantu komputer dalam memahami bahasa alami manusia [14].

Sebagai salah satu teknologi terpenting di era informasi, NLP merupakan bagian krusial dari kecerdasan buatan (AI). Bidang ini bertujuan agar komputer dapat melakukan tugas-tugas yang melibatkan bahasa manusia [10]. Tantangan dalam NLP sering kali mencakup tiga area utama: pengenalan ucapan (*speech recognition*), pemahaman bahasa alami (*natural language understanding*), dan pembuatan bahasa alami (*natural language generation*). Karena sifat bahasa manusia yang kompleks, NLP dianggap sebagai masalah yang sulit dalam ilmu komputer. Aturan yang mengatur penyampaian informasi melalui bahasa alami tidak mudah dipahami oleh komputer, dan setiap bahasa memiliki aturan sintaksis yang berbeda [14].

2.3 *IOB (Inside, Outside, Beginning) Labelling*

Named Entity Recognition (NER) adalah sebuah teknik dalam bidang *Natural Language Processing* (NLP) yang bertujuan untuk mengidentifikasi dan mengklasifikasikan entitas bernama dalam sebuah teks. Salah satu format anotasi yang umum digunakan untuk melabeli entitas dalam data pelatihan NER adalah format IOB (*Inside, Outside, Beginning*). Format ini sangat berguna untuk memberikan anotasi pada entitas bernama di dalam teks [16]. Format IOB membagi dan melabeli setiap token atau kata dalam teks ke dalam tiga kategori utama [17, 16]:

1. B - *Beginning*: Menandakan bahwa sebuah kata merupakan awal dari sebuah entitas baru. Tag ini diberikan pada token pertama dari sebuah entitas.
2. I - *Inside*: Menandakan bahwa sebuah kata adalah bagian dari entitas yang sudah dimulai oleh token sebelumnya. Tag ini diberikan pada token-token selanjutnya dalam sebuah entitas yang terdiri dari beberapa kata.
3. O - *Outside*: Menandakan bahwa sebuah kata tidak termasuk ke dalam entitas mana pun dan tidak diberikan anotasi khusus.

Kegunaan utama dari format IOB adalah kemampuannya untuk merepresentasikan entitas yang terdiri dari beberapa token (*multi-token entities*). Dengan menandai secara jelas mana token yang menjadi awal (B) dan bagian dalam (I) sebuah entitas, serta mana yang berada di luar (O), format ini memudahkan

model NER untuk mempelajari dan memahami batasan-batasan entitas dalam sebuah kalimat dengan benar. Oleh karena itu, menyediakan data pelatihan dalam format ini menjadi langkah penting agar model dapat belajar mengenali dan melabeli entitas bernama secara akurat [16].

2.4 *Conditional Random Fields*

Conditional Random Fields (CRF) adalah sebuah kerangka kerja pemodelan probabilistik yang telah diakui secara luas sebagai teknik yang sangat berguna untuk tugas segmentasi dan pelabelan data sekuensial dalam berbagai aplikasi *Natural Language Processing* (NLP) [12]. CRF merupakan model pembelajaran relasional yang didasarkan pada model grafis tak terarah (*undirected graphical model*) [13]. Secara formal, CRF didefinisikan sebagai *discriminative probabilistic classifier*. Perbedaan mendasar antara model diskriminatif seperti CRF dengan model generatif seperti *Hidden Markov Models* (HMM) dan *Naive Bayes* adalah bahwa model diskriminatif mencoba memodelkan distribusi probabilitas kondisional P(y|x), sedangkan model generatif mencoba memodelkan distribusi probabilitas gabungan P(x,y) [14].

Keunggulan utama CRF adalah kemampuannya untuk memasukkan fitur-fitur yang saling bergantung dan melakukan pembelajaran yang peka terhadap konteks, yang tidak dapat dilakukan oleh model yang lebih sederhana seperti HMM [13]. Hal ini menjadikan CRF sangat efektif untuk tugas-tugas *sequence labeling* seperti *Named Entity Recognition* (NER) dan *Part-of-Speech (POS) Tagging* [14].

Secara konseptual, CRF bekerja dengan menghitung probabilitas kondisional dari sebuah urutan label keluaran (y) berdasarkan urutan masukan (x). Dalam tugas NER, x adalah urutan token atau kata dalam sebuah kalimat, sedangkan y adalah urutan tag atau label entitas yang sesuai [13]. Probabilitas ini dimodelkan sebagai model log-linear dengan vektor parameter (λ_k) yang nilainya dipelajari selama fase pelatihan [12]. Persamaan dari CRF dapat dilihat pada persamaan 2.1

$$P(y|x) = \frac{1}{Z(x)} \times \exp\left(\sum_t \sum_k \lambda_k \cdot f_k(y_{t-1}, y_t, x)\right) \tag{2.1}$$

2.5 *Confusion Matrix*

Confusion Matrix adalah salah satu metrik yang berguna untuk menilai kinerja model *machine learning*. Matriks ini bisa langsung mengukur seberapa baik atau buruknya model berdasarkan sebaran nilai di setiap selnya [18]. Hasil pengukuran dari *confusion matrix* ditunjukan menggunakan 4 istilah yaitu *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN) [19]. Contoh tabel sederhana *confusion matrix* dapat dilihat pada 2.1.

Tabel 2.1. Tabel *Confusion Matrix*

	Actual Positive	*Actual Negative*
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Dengan menggunakan hasil dari *confusion matrix*, dapat dilakukan perhitungan evaluasi yang meliputi *accuracy*, *precision*, *recall*, dan *f1-score* dengan menggunakan rumus 2.2, 2.3, 2.4, 2.5 [20]:

$$Precision = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.2}$$

$$Precision = \frac{TP}{TP+FP} \tag{2.3}$$

$$Recall = \frac{TP}{TP+FN} \tag{2.4}$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.5}$$