BAB I

PENDAHULUAN

1.1 Latar Belakang

Bahasa Jepang menempati peringkat ke 5 sebagai bahasa asing yang paling banyak dipelajari di Asia Tenggara. Berdasarkan *Survey on Japanese Language Education Abroad* 2021 dari *The Japan Foundation*, jumlah pelajar bahasa Jepang di kawasan ini mencapai sekitar 1,18 juta orang, dengan Indonesia menyumbang hampir 60% dari total tersebut, menjadikannya sebagai negara dengan jumlah pelajar terbanyak di kawasan (Foundation, 2021). Selain itu, menurut BPS Provinsi Bali (2023), sebanyak 116.232 wisatawan mancanegara asal Jepang tercatat berkunjung ke Bali selama tahun 2023. Fakta ini menunjukkan adanya peningkatan interaksi budaya dan kebutuhan komunikasi lintas bahasa, yang mempertegas urgensi penguasaan bahasa Jepang oleh masyarakat lokal, terutama dalam konteks pariwisata dan pelayanan.

Di sisi lain, dinamika bahasa Jepang terus melahirkan kata kerja baru, baik melalui serapan kosakata asing maupun pembentukan kreatif oleh penutur muda. Kata kerja baru tersebut umumnya mengikuti pola konjugasi yang sudah ada, misalnya penambahan akhiran -ru pada kata benda (*guguru* dari "*Google*"), atau pembentukan melalui pola suru seperti *kopī suru* dari "*copy*" (Tsujimura & Davis, 2011;Barešová & Zawiszová, 2015) . Setiap entri baru memerlukan pelabelan ke

dalam kelompok konjugasi yang tepat oleh anotator, baik pakar linguistik maupun tim leksikografis, agar konsisten dengan standar morfologis dan dapat digunakan secara andal dalam pembelajaran maupun pemrosesan bahasa alami.

Namun, proses pelabelan manual oleh anotator menghadapi tantangan konsistensi. Tingkat kesepakatan antar anotator (*inter-annotator agreement*) pada pelabelan *verb sense* dilaporkan hanya sekitar 0,68 (Fry & Bond, 2010), sedangkan pada tugas *word sense disambiguation (WSD)* kesepakatan awal tercatat 0,6878 (Okumura et al., 2010). Nilai ini mengindikasikan adanya variasi penafsiran yang cukup tinggi di antara anotator, yang dapat diperparah oleh variasi ortografi (satu kata dapat ditulis dengan beberapa cara) dan keberadaan heteronim (bentuk tulisan sama tetapi makna berbeda). Kondisi tersebut berpotensi menghasilkan label yang tidak seragam, terutama pada kata kerja baru atau bentuk-bentuk yang jarang digunakan.

Untuk menjaga konsistensi anotasi, berbagai sumber kamus morfologis resmi seperti *UniDic*, *IPADIC*, dan *Sudachi* menyediakan penandaan (tag) kelas kata kerja yang dapat dipetakan ke tiga kelompok konjugasi utama, yaitu K1 (*Godan*), K2 (*Ichidan*), dan K3 (*Fukisoku*). Selain itu, kamus daring multibahasa *JMdict* juga menggunakan kode kategori seperti v5, v1, atau vs yang secara fungsional setara dengan pembagian kelompok tersebut. Pemetaan tag dari berbagai sumber ini (lihat Tabel 1.1) menjadi rujukan penting dalam pelabelan kata kerja, baik untuk anotasi manual maupun untuk pengembangan sistem klasifikasi otomatis.

Tabel 1. 1 Peta Konversi Tag Kata Kerja

Sumber Tag	Contoh Tag	Kategori	Notasi Penelitian
	*五段-**	*I Group*	K1

Sumber Tag	Contoh Tag	Kategori	Notasi Penelitian
UniDic / IPADIC / Sudachi	*一段-**	*II Group*	K2
	サ変- / 力変-**	*III Group*	K3
JMdict	*v5**	*I Group*	K1
	v1	*II Group*	K2
	vs / vs-i / vs-s / vk	*III Group*	K3

Meskipun sistem konjugasi kata kerja bahasa Jepang secara umum stabil, akhiran -ru yang lazim ditemukan pada kata kerja K2 juga muncul pada K1 dan K3. Sebagai contoh, *taberu (食べる)* termasuk K2, *kaeru (帰る)* termasuk K1, dan *kuru (来る)* termasuk K3. Secara umum, kata kerja dengan huruf sebelum -ru berupa vokal i atau e cenderung termasuk K2 (Taeko Kamiya, 2001). Namun, terdapat pengecualian penting, seperti kaeru (帰る), hashiru (走る), dan shiru (知る) yang meskipun memenuhi pola permukaan tersebut, secara morfologis masuk ke K1. Kesamaan bentuk akhir ini menimbulkan ambiguitas, baik bagi pembelajar bahasa maupun bagi sistem klasifikasi otomatis yang hanya mengandalkan ciri permukaan.

Untuk mengurangi perbedaan penilaian antar anotator, langkah penting yang dapat dilakukan adalah mengidentifikasi kata kerja yang sudah tervalidasi sebelumnya sebagai acuan. Dengan adanya basis data kata kerja berlabel yang telah disepakati, proses anotasi terhadap kata kerja baru dapat dilakukan dengan membandingkan kemiripan pola morfologis dan konjugasi dengan entri yang sudah ada. Pendekatan ini membantu menjaga konsistensi label dan meminimalkan kesalahan klasifikasi yang bersumber dari interpretasi subjektif anotator.

Dalam konteks ini, *machine learning* dapat dimanfaatkan untuk memprediksi kelompok kata kerja baru berdasarkan pola yang dipelajari dari kata kerja yang telah tervalidasi. Model klasifikasi berbasis pembelajaran mesin mampu

mengenali ciri morfologis seperti akhiran, huruf sebelumnya, dan kombinasi karakter yang sering sulit ditangkap oleh aturan linguistik manual. Dengan demikian, sistem berbasis *machine learning* dapat berfungsi sebagai *validator* kedua yang memberikan rekomendasi label secara otomatis, sehingga mempercepat proses anotasi dan mengurangi potensi inkonsistensi.

Penerapan metode ini pada akhirnya diharapkan dapat menghasilkan alat bantu rekomendasi bagi anotator. Sistem ini akan menyarankan label kelompok kata kerja yang paling mungkin berdasarkan analisis morfologis dan hasil pembelajaran dari data historis. Anotator tetap memegang keputusan akhir, namun keberadaan rekomendasi ini dapat meningkatkan efisiensi, menjaga konsistensi pelabelan, serta memperluas cakupan anotasi untuk menangani kemunculan kata kerja baru secara berkelanjutan.

Pendekatan berbasis aturan (rule-based system) memang dapat digunakan untuk mengklasifikasikan kata kerja, tetapi memiliki keterbatasan dalam menghadapi bentuk-bentuk baru, variasi ortografi, dan pengecualian pola. Sistem semacam ini cenderung kaku dan sulit beradaptasi terhadap dinamika bahasa yang terus berkembang. Oleh karena itu, penelitian ini mengadopsi metode machine learning yang lebih fleksibel. Dalam penelitian ini, klasifikasi dilakukan dengan memanfaatkan ciri morfologis permukaan dari bentuk dasar kata kerja, seperti keberadaan akhiran -ru (*ContainRU*), satu huruf sebelum -ru (*OcbRU*), dan dua huruf sebelum -ru (*TcbRU*). Fitur-fitur ini dipilih karena sederhana namun relevan untuk membedakan pola konjugasi, dan detail pemrosesannya dibahas pada Bab 3.

Metode yang digunakan membandingkan dua algoritma yang telah banyak digunakan dalam berbagai studi klasifikasi linguistik, yaitu *Support Vector*

Machines (*SVM*) dan *Random Forest* (*RF*). *SVM* dikenal efektif dalam menangani data berdimensi tinggi dan klasifikasi morfologis (Pal & Mather, 2003;Okada & Yamamoto, 2014), sedangkan *RF* unggul dalam menghadapi data tidak seimbang serta menyediakan interpretabilitas model melalui *feature importance* (Breiman, 2001;Fauzi, 2018). Keduanya akan dievaluasi secara menyeluruh berdasarkan akurasi, metrik evaluasi multi-kelas, kualitas prediksi probabilistik, dan efisiensi komputasi, guna mengidentifikasi metode yang paling efektif untuk mendukung proses anotasi dan klasifikasi kata kerja bahasa Jepang.

1.2 Identifikasi Masalah

Berdasarkan uraian pada latar belakang, dapat diidentifikasi beberapa permasalahan utama sebagai berikut:

1. Munculnya kata kerja baru dalam bahasa Jepang, baik dari serapan maupun pembentukan kreatif, yang memerlukan pelabelan kelompok konjugasi yang tepat agar konsisten dengan standar morfologis dan dapat digunakan dalam pembelajaran maupun pemrosesan bahasa alami.
2. Konsistensi label antar anotator masih menjadi kendala, tercermin dari nilai *inter-annotator agreement* yang relatif rendah (0,68) dan kesepakatan awal pada tugas *word sense disambiguation* yang juga rendah (0,6878), sehingga hasil anotasi berpotensi bervariasi.
3. Ambiguitas morfologis akhiran -ru, yang dapat muncul di semua kelompok kata kerja, menyulitkan klasifikasi otomatis apabila hanya mengandalkan ciri permukaan. Hal ini terutama terjadi pada kata kerja dengan huruf sebelum -ru berupa i atau e yang biasanya K2, namun memiliki pengecualian di K1.

4. Ketiadaan sistem acuan berbasis kata kerja tervalidasi membuat proses pelabelan kata kerja baru rentan tidak seragam, terutama pada kasus ambiguitas bentuk dan pola konjugasi.
5. Keterbatasan pendekatan berbasis aturan (rule-based system) yang kaku dan sulit beradaptasi terhadap bentuk kata baru, variasi ortografi, serta pengecualian pola konjugasi.
6. Belum adanya kajian komparatif antara algoritma *Support Vector Machines (SVM)* dan *Random Forest (RF)* untuk klasifikasi kata kerja bahasa Jepang berbasis ciri morfologis sederhana, seperti *ContainRU, OcbRU,* dan *TcbRU*, yang hasilnya dapat dimanfaatkan sebagai alat rekomendasi label bagi anotator guna meningkatkan efisiensi dan konsistensi anotasi.

1.3 Batasan Penelitian

Penelitian ini memiliki ruang lingkup yang telah ditetapkan untuk memastikan fokus dan keterukuran dalam proses pelaksanaan. Batasan batasan ini ditetapkan berdasarkan pertimbangan metodologis, keterbatasan sumber daya, dan relevansi terhadap tujuan penelitian. Adapun batasan penelitian yang diterapkan secara nyata adalah sebagai berikut:

1. Jenis data yang digunakan adalah kata kerja bahasa Jepang dalam bentuk dasar (*dictionary form*) yang diambil dari kamus digital dan sumber referensi tepercaya.
2. Kategori kelas yang dianalisis terbatas pada tiga kelompok konjugasi utama, yaitu K1, K2, dan K3 masing-masing berjumlah seimbang.

3. Fitur morfologis yang digunakan untuk klasifikasi hanya mencakup: keberadaan akhiran -ru (*ContainRU*), satu huruf sebelum -ru (*OcbRU*), dan dua huruf sebelum -ru (*TcbRU*). Fitur lain seperti makna semantik atau konteks kalimat tidak dianalisis.
4. Pendekatan machine learning yang digunakan terbatas pada dua algoritma, yaitu *Support Vector Machines (SVM)* dan *Random Forest (RF)*, dengan pembandingan performa untuk menilai efektivitas masing-masing metode.
5. Hasil penelitian difokuskan sebagai dasar pengembangan alat rekomendasi label bagi anotator, dan tidak mencakup implementasi antarmuka pengguna atau integrasi ke dalam sistem anotasi yang sebenarnya.

1.4 Rumusan Masalah

Berdasarkan latar belakang, identifikasi masalah, dan batasan ruang lingkup yang telah dijelaskan sebelumnya, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana mengatasi permasalahan konsistensi anotasi pada kata kerja bahasa Jepang, khususnya yang berakhiran -ru, yang dapat muncul pada semua kelompok konjugasi?
2. Bagaimana fitur morfologis sederhana, yaitu keberadaan akhiran -ru (*ContainRU*), satu huruf sebelum -ru (*OcbRU*), dan dua huruf sebelum -ru (*TcbRU*), dapat dimanfaatkan untuk membedakan kelompok kata kerja bahasa Jepang?
3. Bagaimana performa algoritma *Support Vector Machines (SVM)* dan *Random Forest (RF)* dalam mengklasifikasikan kata kerja bahasa Jepang ke dalam tiga

kelompok konjugasi, ditinjau dari akurasi, *precision*, *recall, F1-score, ROC-AUC, Log Loss,* serta efisiensi waktu pelatihan dan prediksi?

4. Algoritma manakah yang lebih tepat digunakan sebagai dasar alat rekomendasi label bagi anotator, guna membantu meningkatkan efisiensi dan konsistensi dalam proses pelabelan kata kerja bahasa Jepang?

1.5 Tujuan Penelitian

Adapun uujuan dari penelitian ini adalah sebagai berikut:

1. Mengatasi permasalahan konsistensi anotasi pada kata kerja bahasa Jepang, melalui pendekatan berbasis *machine learning* yang mampu menangani variasi bentuk dan pola konjugasi.
2. Menganalisis kontribusi fitur morfologis sederhana seperti keberadaan akhiran -ru (*ContainRU*), satu huruf sebelum -ru (*OcbRU*), dan dua huruf sebelum -ru (*TcbRU*) dalam membedakan kelompok kata kerja bahasa Jepang.
3. Mengevaluasi dan membandingkan performa algoritma *Support Vector Machines (SVM)* dan *Random Forest (RF)* dalam mengklasifikasikan kata kerja bahasa Jepang ke dalam tiga kelompok konjugasi berdasarkan metrik akurasi, *precision, recall, F1-score, ROC-AUC, Log Loss*, serta efisiensi waktu pelatihan dan prediksi.
4. Menentukan algoritma yang paling tepat untuk dijadikan dasar alat rekomendasi label bagi anotator, sehingga dapat membantu meningkatkan efisiensi dan konsistensi dalam proses pelabelan kata kerja bahasa Jepang.

1.6 Manfaat Penelitian

Penelitian ini diharapkan memberikan manfaat baik dari segi teoretis maupun praktis. Manfaat tersebut dijabarkan sebagai berikut:

1. Manfaat Teorities.
    a. Memberikan kontribusi pada pengembangan metode klasifikasi morfologis kata kerja bahasa Jepang berbasis machine learning, khususnya dengan memanfaatkan fitur morfologis sederhana seperti *ContainRU*, *OcbRU*, dan *TcbRU*.
    b. Menambah referensi ilmiah mengenai perbandingan performa algoritma *Support Vector Machines (SVM)* dan *Random Forest (RF)* dalam domain linguistik komputasional, serta mengisi kekosongan kajian komparatif keduanya untuk klasifikasi kata kerja bahasa Jepang.
2. Manfaat Praktis
    a. Menjadi dasar bagi pengembangan alat rekomendasi label untuk anotator, sehingga dapat membantu meningkatkan efisiensi dan konsistensi proses pelabelan kata kerja bahasa Jepang, termasuk kata kerja baru yang belum terdokumentasi di kamus konvensional.
    b. Memberikan acuan bagi pengembang sistem *Natural Language Processing (NLP)* atau perangkat lunak analisis linguistik dalam memilih metode klasifikasi yang sesuai untuk pengolahan data bahasa Jepang.
    c. Mendukung pembelajaran bahasa Jepang, khususnya pada aspek konjugasi kata kerja, dengan menyediakan pendekatan teknologi yang adaptif dan terotomatisasi.