BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Bahasa adalah sebuah alat komunikasi yang terdiri dari susunan kata terstruktur dan digunakan manusia untuk saling berkomunikasi secara lisan dan tertulis. Bahasa memiliki hubungan timbal balik dengan konteks sosial. Di satu sisi, bahasa berperan sebagai alat penyebaran makna. Di sisi lain, bahasa berperan sebagai wujud nyata praktik sosial [1]. Oleh karena itu, penggunaan bahasa harus dilakukan sesuai dengan kaidah berbahasa yang baik dan benar.

Kesalahan berbahasa di bidang sintaksis memiliki dampak yang signifikan dalam berbahasa. Berdasarkan penelitian tentang dampak kesalahan berbahasa bidang sintaksis, ditemukan bahwa kesalahan sintaksis dapat menyebabkan hilangnya sebagian hingga seluruh makna dari suatu kalimat [2]. Hal ini dapat berpotensi menyebabkan miskomunikasi di antara pembicara dengan pendengar atau penulis dengan pembaca. Namun, tingkat kemahiran berbahasa dari masyarakat Indonesia masih relatif kurang. Berdasarkan hasil Uji Kemahiran Berbahasa Indonesia (UKBI) Adaptif Merdeka yang telah dilakukan oleh Badan Pengembangan dan Pembinaan Bahasa, diketahui bahwa hanya delapan dari 42 dosen dari perguruan tinggi di Jabodetabek yang mencapai predikat Istimewa dan Sangat Unggul di tahun 2021 dan 89.6% dari 72.195 pelajar SMA memiliki predikat di bawah Unggul (Madya, Semenja, Marginal, Terbatas) di tahun 2022 [3, 4].

Saat ini, terdapat banyak teknologi yang dapat digunakan untuk *natural language processing*, baik model maupun algoritma. Terdapat beberapa teknologi *machine learning* untuk *natural language processing* seperti *Embedding, Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), dan Recurrent Neural Network (RNN)* [5, 6]. Terdapat pula beberapa algoritma untuk *natural language processing* seperti *Regular Expressions, n-grams, WordNet, Hidden Markov Model, Conditional Random Field*, dan *Context Free Grammar* [7, 8].

Suatu kalimat yang baik dan benar terdiri dari kombinasi kata-kata bahasa Indonesia yang sesuai dengan kaidah bahasa Indonesia yang berlaku. Kata-kata tersebut dapat berperan sebagai frasa dan frasa tersebut dapat berperan sebagai subjek, predikat, objek, pelengkap, atau keterangan [9]. Oleh karena itu, diperlukan

algoritma yang dapat memastikan bahwa kombinasi-kombinasi kata suatu kalimat telah sesuai dengan *rule-rule* yang telah ditentukan. Algoritma yang cocok untuk tujuan ini adalah algoritma *Context Free Grammar* [7, 8].

Beberapa kata pada bahasa Indonesia dapat menempati peran yang berbeda (bersifat ambigu). Sebagai contoh, kata *sekolah* dapat berperan sebagai verba dan nomina. Oleh karena itu, diperlukan algoritma *part-of-speech tagging* untuk membuat peran suatu kata tidak ambigu. *Hidden Markov Model* dapat digunakan untuk tujuan ini, tetapi *Hidden Markov Model* tidak dapat memberikan prediksi yang optimal pada kata-kata yang belum pernah dipelajari sebelumnya. *Conditional Random Field* lebih cocok digunakan karena dapat memperhitungkan fitur-fitur yang terdapat pada kata yang belum pernah dipelajari sebelumnya, seperti huruf kapital, sufiks *-nya*, dan sebagainya [8].

Penelitian terkait *Conditional Random Field* telah dilakukan sebelumnya untuk *named entity recognition* dan *opinion extraction* bahasa Indonesia [10, 11]. Penelitian terkait *Context Free Grammar* juga telah dilakukan untuk pengecekan sintaksis kalimat bahasa Indonesia [12]. Berdasarkan penelitian terkait CRF dan CFG tersebut, diketahui bahwa algoritma CRF tidak dapat digunakan untuk analisis semantik dan sintaksis sedangkan algoritma CFG mengalami kesulitan mengenali kata-kata/istilah-istilah baru. Namun, belum ada penelitian yang menggabungkan *Conditional Random Field (CRF)* untuk *part-of-speech tagging* dan *Context Free Grammar (CFG)* untuk pengecekan sintaksis kalimat bahasa Indonesia. Dengan menggabungkan CRF dan CFG, kedua algoritma tersebut dapat mengatasi kekurangan satu sama lain.

Penelitian ini merupakan pengembangan lebih lanjut dari proyek U-Tapis bidang pendeteksi kesalahan sintaksis kalimat [13]. Algoritma U-Tapis ini merupakan gabungan dari algoritma *Conditional Random Field/CRF* (*part-of-speech/POS tagging*) dan *Context Free Grammar/CFG* (*parsing*). Penelitian ini diharapkan dapat menghasilkan algoritma yang dapat meningkatkan kualitas tata bahasa dari suatu artikel berita dengan mendeteksi kesalahan sintaksis kalimat yang ada pada artikel tersebut.

1.2 Rumusan Masalah

Berikut adalah rumusan masalah dari penelitian ini:

1. Bagaimana cara mengembangkan U-Tapis di bidang pendeteksian kesalahan sintaksis kalimat bahasa Indonesia untuk artikel berita?

2. Bagaimana cara melakukan *deployment* ”Algoritma U-Tapis Pendeteksi Kesalahan Sintaksis Kalimat Bahasa Indonesia” dalam bentuk *Application Programming Interface/API*?

3. Bagaimana cara mengevaluasi performa dari ”Algoritma U-Tapis Pendeteksi Kesalahan Sintaksis Kalimat Bahasa Indonesia”?

1.3 Batasan Permasalahan

Berikut adalah batasan-batasan masalah dari penelitian ini:

1. Keluaran akhir dari penelitian skripsi ini adalah *Application Programming Interface/API* U-Tapis Pendeteksi Kesalahan Sintaksis Kalimat. Tidak ada keluaran aplikasi untuk menampilkan hasil API tersebut.

2. ”Algoritma U-Tapis Pendeteksi Kesalahan Sintaksis Kalimat” tidak akan memperbaiki kesalahan tik dan mengasumsikan bahwa *input* yang diberikan tidak memiliki kesalahan tik.

3. ”Algoritma U-Tapis Pendeteksi Kesalahan Sintaksis Kalimat” tidak akan memperhatikan keserasian makna/semantik dari *input* yang diberikan.

4. ”Algoritma U-Tapis Pendeteksi Kesalahan Sintaksis Kalimat” hanya akan melakukan *part-of-speech tagging* dan tidak akan melakukan *named entity recognition* (nama orang, nama gedung, nama lokasi, dan lain-lain). Semua *named entity* akan digolongkan sebagai nomina.

5. Tahap pengembangan ”Algoritma U-Tapis Pendeteksi Kesalahan Sintaksis Kalimat Bahasa Indonesia” akan menggunakan 60 artikel berita Tribun News dan data-data kalimat buatan penulis. Tahap uji coba akan menggunakan 15 artikel berita Tribun News. Rasio *train-test split* yang digunakan adalah 80% : 20%.

6. Saat digunakan di tahap *deployment*, *input* yang diterima oleh ”Algoritma U-Tapis Pendeteksi Kesalahan Sintaksis Kalimat Bahasa Indonesia” adalah data berita yang berisi maksimal 40 kalimat untuk setiap data berita. Batasan ini dibuat untuk membatasi jumlah data berita yang diterima pada setiap *request* oleh tersebut.

1.4 Tujuan Penelitian

Berikut adalah tujuan penelitian dari penelitian ini:

1. Mengembangkan proyek U-Tapis di bidang pendeteksian kesalahan sintaksis kalimat bahasa Indonesia pada artikel berita dengan menggunakan algoritma *Conditional Random Field (CRF)* dan *Context Free Grammar (CFG)*.
2. Mengembangkan *Application Programming Interface* (API) untuk melakukan *deployment* "Algoritma U-Tapis Pendeteksi Kesalahan Sintaksis Kalimat Bahasa Indonesia" dengan menggunakan *Python Flask Web Framework*.
3. Mengevaluasi performa dari "Algoritma U-Tapis Pendeteksi Kesalahan Sintaksis Kalimat Bahasa Indonesia" dengan menggunakan metrik *accuracy*, *precision*, *recall*, *F1-Score/F-Measure*, dan *run time*.

1.5 Manfaat Penelitian

Hasil dari penelitian ini diharapkan dapat memberikan manfaat-manfaat sebagai berikut:

1. Membantu memastikan kesesuaian artikel berita baru Tribun News dengan kaidah sintaksis bahasa Indonesia yang baik dan benar dengan cara menunjukkan kalimat-kalimat dengan sintaksis yang benar dan salah.
2. Meningkatkan kemampuan peneliti di bidang teori automata dan *natural language processing/NLP*.
3. Mendorong penggunaan teknologi *machine learning* di dalam proses pembuatan artikel berita.
4. Menjadi bahan referensi untuk penelitian serupa di masa depan.

1.6 Sistematika Penulisan

Berikut adalah sistematika penulisan dari laporan penelitian ini:

- Bab 1 PENDAHULUAN

Bab ini membahas tentang latar belakang penelitian ini, rumusan masalah yang ingin diselesaikan, batasan dari masalah tersebut, tujuan penelitian yang ingin dicapai, manfaat dari penelitian ini, dan uraian sistematika penulisan dari laporan skripsi ini.

- Bab 2 LANDASAN TEORI

  Bab ini berisi teori-teori pendukung penelitian skripsi ini, yaitu informasi tentang perusahaan Tribun News, sintaksis kalimat Bahasa Indonesia, kata, frasa, klausa, kalimat, aposisi, suplementasi, *text normalization*, algoritma *Conditional Random Field*, algoritma *Context Free Grammar*, metrik pengukur performa algoritma (*precision, recall, accuracy, F1-score, runtime*), dan *Python Flask Web Framework*.

- Bab 3 METODOLOGI PENELITIAN

  Bab ini berisi uraian metode penelitian yang digunakan di dalam penelitian skripsi ini, yaitu pengumpulan data, *data preprocessing*, pendefinisian *class Conditional Random Field*, perancangan rule-rule *Context Free Grammar*, pelabelan *training data*, *training* algoritma *Conditional Random Field*, evaluasi performa *Conditional Random Field* dan *Context Free Grammar*, *deployment* sebagai *application programming interface/API*, dan *testing* hasil *deployment application programming interface/API*.

- Bab 4 HASIL DAN DISKUSI

  Bab ini berisi uraian implementasi metode penelitian yang telah diuraikan pada bab ketiga serta hasil pengembangan dan implementasi dari ”Algoritma U-Tapis Pendeteksi Kesalahan Sintaksis Kalimat Bahasa Indonesia”.

- Bab 5 KESIMPULAN DAN SARAN

  Bab ini berisi uraian kesimpulan dari penelitian skripsi ini dan saran untuk penelitian serupa di masa depan.