BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Pada era informasi dapat tersebar cepat, media seperti koran daring adalah alat yang sangat penting dalam menyampaikan informasi dan berita kepada masyarakat [1]. Media massa, khususnya portal berita daring, memiliki tanggung jawab besar dalam menyampaikan informasi yang akurat dan menggunakan bahasa yang baik agar dapat dipahami dengan jelas oleh masyarakat. Untuk menjaga konsistensi dan kemudahan pemahaman bahasa dalam berita, Persatuan Wartawan Indonesia (PWI) menetapkan pedoman bahasa jurnalistik. Salah satu pedomannya adalah penggunaan Pedoman Ejaan Bahasa Indonesia Yang Disempurnakan (EYD) secara konsisten [2]. Walaupun begitu, masih ada media berita online yang masih kurang teliti sehingga terdapat kesalahan yang berakibat teks berita tersebut tidak sesuai dengan EYD [3, 4].

Untuk mengatasi penulisan teks berita yang salah dan tidak sesuai dengan kaidah yang berlaku, telah dikembangkan aplikasi U-Tapis yaitu sebuah sarana bagi calon jurnalis dan jurnalis untuk melakukan pengecekan kesalahan pada penulisan teks berita [5, 6]. U-Tapis menggunakan teknik *Natural Language Processing* (NLP) dan *machine learning* dalam pembuatannya [5]. U-Tapis telah memiliki beberapa modul untuk mendeteksi berbagai kesalahan penulisan seperti kesalahan penggunaan kata konjungsi, kesalahan peluluhan kata, kesalahan tanda baca, kesalahan penulisan kata majemuk, kesalahan penggunaan kata ’di’, dan kesalahan sintaksis bahasa Indonesia. Walaupun demikian, masih banyak kesalahan penulisan yang belum mampu dideteksi oleh U-Tapis sehingga diperlukan pengembangan lebih lanjut, salah satunya adalah kesalahan penggunaan angka dan bilangan yang tidak sesuai dengan EYD.

Kesalahan penggunaan angka dan bilangan yang menyalahi EYD adalah salah satu kesalahan umum yang masih ditemukan pada teks berita [7]. Kompleksitas aturan ini menjadi tantangan utama. Pedoman EYD mensyaratkan penulisan yang sangat kontekstual: sebuah angka seperti ’10’ bisa dianggap salah dalam frasa ”selama 10 hari” (seharusnya ditulis ”sepuluh hari”), namun menjadi benar dalam konteks ”Pasal 10” atau ”pukul 10.00”. Aturan ini meluas ke berbagai kasus lain, seperti penulisan bilangan tingkat (”juara 2” dengan ”juara ke-2”), penulisan

bilangan besar (”1.000.000” dengan ”1 juta”), hingga penulisan nama geografi (”Kelapa Dua” dengan ”Kelapadua”). Kerumitan dan banyaknya pengecualian ini membuat pemeriksaan manual menjadi tidak efisien dan rentan terhadap kesalahan, oleh karena itu diperlukan bantuan berbasis *machine learning* untuk melakukan pengecekan pada penggunaan angka dan bilangan pada teks berita.

Perkembangan teknologi kecerdasan buatan, khususnya *machine learning*, menawarkan peluang untuk menyelesaikan permasalahan ini. Kecerdasan buatan telah mendapatkan banyak perhatian selama satu dekade ke belakang, salah satunya adalah *machine learning* yang juga menjadi topik populer [8]. *machine learning* dapat diartikan sebagai algoritma yang mampu mempelajari sebuah data sampel yang disebut sebagai training set untuk membuat sebuah model prediksi dan melakukan prediksi atau melakukan identifikasi pola kompleks yang sulit diprediksi oleh manusia [9]. *Natural Language Processing* (NLP) adalah salah satu bidang *machine learning* yang berfokus membuat komputer dapat melakukan tugas yang berhubungan dengan bahasa manusia [10]. Berdasarkan penelitian yang sudah ada, munculnya teknik *machine learning* berupa NLP membuat pergeseran dalam pemilihan alat untuk pengecekan tata bahasa, yang awalnya dari teknik *rule-based* ke pendekatan *machine learning* [11].

Berbagai model *machine learning* dapat digunakan untuk NLP, salah satu yang banyak digunakan untuk permasalahan sekuensial dalam NLP adalah *Conditional Random Fields* (CRF). Beberapa penelitian sebelumnya telah menunjukkan keberhasilan CRF dalam berbagai tugas NLP. Salah satu penelitian mengembangkan pendekatan CRF untuk bahasa Urdu, sebuah bahasa yang termasuk *low-resource language* dengan tantangan unik seperti ketiadaan kapitalisasi, urutan kata bebas, dan morfologi kompleks [12]. Hasil penelitian tersebut menunjukkan bahwa penggunaan CRF dengan rancangan *feature engineering* baru, mampu meningkatkan kinerja sistem NER dibandingkan *baseline*, dengan peningkatan nilai F1 sebesar 1,5%–3% [12]. Penelitian lain menunjukkan efektivitas CRF untuk NER dalam bahasa Marathi, yang memiliki sifat morfologi kaya dan struktur kalimat bebas. Dengan memanfaatkan korpus FIRE-2010 berisi lebih dari 27.000 kalimat beranotasi, mereka melatih sistem Mner-CRF yang mencapai akurasi rata-rata F1 sebesar 75,51%. Hasil ini menunjukkan bahwa CRF cukup handal untuk menangani bahasa alami yang kompleks, terutama ketika dipadukan dengan fitur linguistik dan anotasi data yang memadai [13]. CRF efektif digunakan dalam NLP karena mampu memodelkan hubungan kontekstual antar token dalam sebuah kalimat serta mempertimbangkan

informasi sekuensial secara global, tidak hanya bergantung pada fitur lokal.

Penelitian Ketmaneechairat et al. (2020) telah membandingkan CRF dengan model *machine learning* lainnya, termasuk model berbasis *deep learning* seperti *bidirectional Long Short-Term Memory* (Bi-LSTM) [14]. Sebuah studi mengenai *Natural Language Processing* untuk manajemen bencana menunjukkan bahwa CRF yang dioptimalkan (*CRF-optimized*) menghasilkan performa yang lebih baik daripada kombinasi Bi-LSTM-CRF untuk tugas *Named Entity Recognition* (NER) [14]. Hasil ini menunjukkan bahwa model yang lebih sederhana dan dioptimalkan dapat mengungguli model yang lebih kompleks, tergantung pada jenis data dan tugas yang dihadapi. Keunggulan CRF terletak pada kemampuannya untuk mempertimbangkan konteks kata di sekitarnya, sebuah fitur yang sangat krusial untuk kasus penulisan angka dan bilangan dalam bahasa Indonesia. Dengan mempertimbangkan kata-kata di depan dan belakang angka, CRF dapat membedakan konteks seperti ”pukul 10.00” (benar) dan ”sepuluh hari” (salah jika ditulis angka). Ini menjadikan CRF pilihan yang paling sesuai karena menawarkan keseimbangan antara performa yang tinggi, efisiensi komputasi, dan kemampuan untuk diinterpretasikan, yang sangat penting dalam penelitian ini.

Berdasarkan latar belakang tersebut, maka penelitian ini dilakukan untuk merancang dan membangun model *machine learning* untuk mendeteksi kesalahan penggunaan angka dan bilangan yang tidak sesuai dengan EYD, untuk diterapkan pada aplikasi U-Tapis. Dengan adanya penelitian ini, diharapkan dapat membantu para jurnalis dalam melakukan pengecekan dan penyuntingan artikel berita yang telah ditulis agar sesuai dengan EYD.

1.2 Rumusan Masalah

Sesuai dengan latar belakang masalah yang telah dipaparkan, maka dapat dirumuskan masalah sebagai berikut:

1. Bagaimana cara mengimplementasikan algoritma *Conditional Random Fields* pada *website* U-Tapis untuk mendeteksi kesalahan penggunaan angka dan bilangan menurut EYD pada teks berita?
2. Bagaimana akurasi, presisi, *f1-score*, dan *recall* algoritma *Conditional Random Fields* pada *website* U-Tapis dalam mendeteksi kesalahan penggunaan angka dan bilangan menurut EYD?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah tersebut, adapun tujuan yang hendak dicapai dari penelitian ini adalah:

1. Mengimplementasikan algoritma *Conditional Random Fields* pada *website* U-Tapis untuk mendeteksi kesalahan penggunaan angka dan bilangan menurut EYD.
2. Mengukur akurasi, presisi, *f1-score*, dan *recall* algoritma *Conditional Random Fields* pada *website* U-Tapis dalam mendeteksi kesalahan penggunaan angka dan bilangan menurut EYD.

1.4 Urgensi Penelitian

Berdasarkan latar belakang masalah tersebut, maka urgensi dari penelitian ini adalah adanya kesalahan umum dalam penulisan teks berita yaitu penggunaan angka dan bilangan yang tidak sesuai dengan EYD, padahal portal berita memiliki tanggung jawab besar untuk menggunakan bahasa yang baik agar mudah dipahami masyarakat. Selain itu, penggunaan EYD secara konsisten juga adalah pedoman bahasa jurnalistik. Oleh karena itu, perlu dibuat aplikasi yang mampu mendeteksi kesalahan penggunaan angka dan bilangan berbasis *machine learning* untuk membantu jurnalis agar proses penulisan dan penyuntingan menjadi lebih efisien dan efektif.

1.5 Luaran Penelitian

Luaran dari penelitian ini adalah sebagai berikut.

1. API *website* U-Tapis untuk mendeteksi kesalahan penggunaan angka dan bilangan.
2. Artikel ilmiah terakreditasi Sinta.

1.6 Manfaat Penelitian

Adapun manfaat penelitian ini adalah:

1. Menambah wawasan mengenai implementasi algoritma *Conditional Random Fields* dalam mendeteksi kesalahan penggunaan angka dan bilangan menurut EYD.
2. Membantu tim jurnalistik dalam memeriksa penggunaan angka dan bilangan pada teks berita dengan efisien dan efektif.
3. Menambah pengetahuan mengenai performa algoritma *Conditional Random Fields* dalam mendeteksi kesalahan penggunaan angka dan bilangan menurut EYD.