BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Dalam beberapa tahun terakhir, pertumbuhan *Artificial Intelligence* (AI) telah melakukan revolusi terhadap berbagai aspek kehidupan sehari-hari, khususnya pada bidang *Natural Language Processing* (NLP). Kedatangan *language model* mutakhir seperti *Generative Pre-trained Transformer 3* (GPT-3) telah membawa kemampuan untuk menghasilkan teks seperti buatan manusia pada tingkat yang belum pernah tercapai sebelumnya. Walaupun kemajuan tersebut membawa banyak manfaat, namun juga menimbulkan permasalahan baru, khususnya pada deteksi *AI generated text*. Masalah yang terjadi adalah meningkatnya kesulitan untuk membedakan konten yang dibuat oleh mesin dan manusia.

Dalam bidang pendidikan, misalnya, pengajar menghadapi tantangan besar dalam menilai tugas dan karya tulis siswa. Jika siswa menggunakan *AI* untuk menghasilkan esai atau tugas mereka, sulit bagi pengajar untuk menilai kemampuan dan pemahaman sebenarnya dari siswa tersebut. Selain itu, dalam jurnalisme dan media, penyebaran informasi yang tidak diverifikasi atau palsu yang dihasilkan oleh *AI* dapat merusak kredibilitas sumber berita.

Mendeteksi *AI generated text* membawa beragam tantangan dikarenakan kecanggihan model bahasa kontemporer yang luar biasa. Berdasarkan arsitektur *deep learning*, model-model tersebut memiliki kemampuan untuk menghasilkan perpaduan teks yang akurat secara sintaksis. Kesulitannya terletak pada membedakan pola dan bentuk yang membedakan *AI generated text* dengan teks buatan manusia. *Naïve Bayes*, sebuah algoritma pembelajaran mesin yang seringkali digunakan dalam tugas klasifikasi dokumen, dikenal sebagai algoritma yang menghasilkan model efektif dan memiliki performa baik, khususnya dalam bidang klasifikasi teks dalam sebuah dokumen [1].

Terdapat penelitian terdahulu yang menerapkan algoritma *Naïve Bayes* dalam deteksi *network intrusion*. Pada penelitian tersebut, algoritma *Naïve Bayes* memiliki *detection rate* sebesar 95% [2]. Algoritma *Naïve Bayes* juga pernah digunakan untuk membedakan *human generated* dan *AI generated phising emails*. Dalam penelitian tersebut, algoritma *Naïve Bayes* memiliki akurasi sebesar 94.10%, dengan *standard deviation* sebesar 0.0165 [3]. Deteksi *AI generated text* juga

pernah dilakukan menggunakan *dataset* yang juga akan digunakan pada penelitian terkait deteksi *AI generated text* ini, dengan mengimplementasikan algoritma *Naïve Bayes* dengan *feature extraction Term Frequency-Inverse Document Frequency*, dimana *accuracy* dari model semakin meningkat ketika *hyperparameter* yang digunakan juga meningkat [4].

Multinomial Naïve Bayes adalah sebuah varian dari algoritma *Naïve Bayes* yang merupakan algoritma berbasis frekuensi untuk mengklasifikasikan teks yang direpresentasikan oleh sekumpulan kata yang muncul dalam sebuah dokumen [5]. Kelebihan dari *Multinomial Naïve Bayes* antara lain adalah cocok untuk digunakan pada data yang bersifat kontinu, yaitu data yang dapat diukur seperti tinggi dan berat badan, dan juga data diskrit, yaitu data yang memiliki nilai terbatas, seperti jumlah siswa dalam sebuah kelas [6].

Penelitian ini menyajikan implementasi algoritma *Multinomial Naïve Bayes* untuk mendeteksi *AI generated text* dengan menggunakan *feature extraction Bag of Words*. Metode yang diterapkan meliputi pelatihan algoritma pada *data set* yang berisi teks hasil buatan manusia dan *AI generated text* untuk mengidentifikasi karakteristik unik yang membedakan mereka. Hasil eksperimen awal menunjukkan efektivitas *Multinomial Naïve Bayes* dalam melakukan klasifikasi teks, menyoroti potensinya sebagai algoritma yang layak dalam upaya meningkatkan kemampuan deteksi *AI generated text*.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang sudah disebutkan, berikut adalah rumusan masalah dari penelitian ini.

1. Bagaimana implementasi algoritma *Multinomial Naïve Bayes* untuk deteksi *AI generated text* ?
2. Bagaimana performa algoritma *Multinomial Naïve Bayes* dalam mendeteksi *AI generated text* ?

1.3 Batasan Permasalahan

Berikut merupakan batasan masalah dalam penelitian ini.

1. Data yang digunakan pada penelitian ini merupakan data yang diambil dari *dataset* Kaggle dengan judul LLM - Detect AI Generated Text Dataset,

dengan total sebanyak 29.145 data dengan bahasa Inggris yang diambil pada tahun 2023 [7].

2. Digunakan 60 *input text* pengguna untuk melakukan deteksi *AI generated text*.

1.4 Tujuan Penelitian

Berdasarkan rumusan dan batasan masalah yang telah diuraikan, tujuan dari penelitian ini adalah sebagai berikut.

1. Melakukan deteksi *AI generated text* menggunakan algoritma *Multinomial Naïve Bayes*.
2. Mengukur performa algoritma *Multinomial Naïve Bayes* dalam melakukan deteksi *AI generated text*.

1.5 Manfaat Penelitian

Manfaat yang diperoleh dari penelitian ini adalah sebagai berikut.

1. Mengetahui performa algoritma *Multinomial Naïve Bayes* dalam melakukan deteksi *AI generated text*.
2. Hasil penerapan algoritma dalam penelitian ini dapat dijadikan landasan untuk pembuatan sistem pen deteksi *AI generated text* lainnya.

1.6 Sistematika Penulisan

Berikut adalah sistematika penulisan penelitian yang akan dilakukan.

- Bab 1 PENDAHULUAN
  Pembahasan mencakup latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, dan manfaat penelitian.
- Bab 2 LANDASAN TEORI
  Pembahasan mengenai teori yang berkaitan dengan penelitian yang dilakukan, antara lain *Preprocess data*, *Natural Language Processing*, algoritma *Naïve Bayes*, algoritma *Multinomial Naïve Bayes*, *Bag Of Words*, *Grid Search* dan *Confusion Matrix*.

- Bab 3 METODOLOGI PENELITIAN
  Pembahasan mengenai pendekatan metodologi yang digunakan dalam penelitian beserta alur prosesnya, yang disajikan dalam bentuk diagram alur.
- Bab 4 HASIL DAN DISKUSI
  Analisis terkait pengujian dan evaluasi hasil penelitian yang diperoleh.
- Bab 5 SIMPULAN DAN SARAN
  Membahas mengenai kesimpulan akhir yang didapat melalui penelitian yang telah selesai dilakukan dan juga saran yang dapat diimplementasikan dalam penelitian yang akan datang.