BAB 1
PENDAHULUAN

1.1 Latar Belakang

Natural language processing (NLP) adalah cabang dari bidang ilmu *artificial intelligence* (AI) yang dapat mempelajari interaksi antar manusia dengan komputer melalui bahasa natural, seperti arti kata, frasa, kalimat, dan sintaksis serta proses semantik [1]. Dalam dekade terakhir manfaat penerapan NLP telah dirasakan oleh manusia seperti *chatbot, translator,* hingga *virtual assistant*. Salah satu penerapan NLP yang digunakan untuk penelitian ini ialah dibidang *similarity*.

Plagiarisme merupakan sebuah tindakan yang mengambil karya, ide, ataupun informasi tanpa sepengetahuan pemilik hak cipta kemudian mengakuinya atas miliknya sendiri. Tindakan tersebut tentu merugikan diri sendiri karena plagiasi mengurangi kemampuan untuk meningkatkan daya kreatifitas. Tugas akhir pemrograman setiap semester rentan terhadap plagiarisme oleh mahasiswa, akibat kurangnya pemahaman dan jarang aktif untuk mempertanyakan bagian yang tidak dimengerti pada saat di kelas maupun lab. komputer.

Dalam penelitian terdahulu Qiubo Huang, Xuezhi Song, dan Guozheng Fang melakukan penelitian dengan menggunakan model *LightGBM* dan mendapatkan nilai F1 score sebesar 83% serta *accuracy* sebesar 94% [2]. Vedran Ljubovic dan Enil Pajic ditahun 2020 mendappatkan nilai akurasi yang sebesar 81% dan *recall* sebesar 38% menggunakan metode *ANN* dan *feature extraction* [3]. Farhan Ullah, Junfeng Wang, Muhammad Farhan, Masood Habib, dan Shehzad Khalid mendapatkan akurasi sebesar 86% dengan menggunakan model *multionomial logistic regression* dan *feature extraction* yaitu *principal component analysis* [4]. Penelitian dari Tomáš Foltýnek, Richard Všianský, Norman Meuschke, Dita Dlabolová, dan Bela Gipp mengunakan kombinasi metode dari *explicit semantic analysis* dan *greedy string tilling* untuk mendeteksi kode program dan mendapatkan akurasi mengenai plagiarisme *monolingual* sebesar 99.2% dan 89% mengenai plagiarisme *cross-lingual* [5]. Selanjutnya penelitian dari Ruchir Puri, David S. Kung, Geert Janssen, Wei Zhang, Giacomo Domeniconi, Vladimir Zolotov, Julian Dolby, Jie Chen, Mihir Choudhury, Lindsey Decker, Veronika

Thost, Luca Buratti, Saurabh Pujar, Shyam Ramji, Ulrich Finkler, Susan Malaika, Frederick Reiss ditahun 2021 menggunakan empat metode yaitu *mlp with bag of tokens*, *siamese network with token sequence*, *spt with handcrafted feature extraction*, dan *gnn with spt* untuk mendeteksi kode program java, python, dan C++ [6].

Berdasarkan latar belakang tersebut, peneliti akan melakukan penelitian dengan menerapkan algoritma *bidirectional encoder representations from transformer* untuk mendeteksi plagiarisme pada kode program bahasa C++.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah diatas, maka rumusan masalah yang akan diselesaikan dalam penelitian ini yaitu :

1. Bagaimana algoritma *bidirectional encoder representations from transformer* bekerja untuk mendeteksi plagiarisme pada kode program bahasa C++?.
2. Berapakah persentase keberhasilan *accuracy* model untuk mendeteksi plagiarisme pada kode program bahasa C++?

1.3 Batasan Masalah

Ditemukan batasan masalah dalam penelitian ini, yaitu:

1. Algoritma yang digunakan pada penelitian ini yaitu bert.
2. Dataset yang digunakan hanya kode program C++.
3. Bahasa yang digunakan adalah python.
4. Penelitian hanya dilakukan pada dataset sebanyak 5000 data per label dan dibagi menjadi data *training* serta *validation* dengan perbandingan 70:30.
5. Penelitian berfokus pada hasil tingkat akurasi dan confusion matrix dalam penerapan algoritma bert untuk mendeteksi plagiarisme pada kode program bahasa C++.

1.4 Maksud dan Tujuan Penelitian

Maksud dan tujuan dari penelitian ini adalah:

1. Mengetahui kinerja dari penerapan algoritma bert untuk mendeteksi plagiarisme pada kode program bahasa C++.

2. Mengetahui seberapa besar persentase akurasi keberhasilan algoritma bert untuk mendeteksi plagiarisme pada kode program bahasa C++.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah :

1. Dapat membuat para pengajar lebih mudah untuk mengatasi plagiarisme kode program bahasa C++.
2. Hasil penelitan dapat dijadikan referensi bagi peneliti berikutnya dan kedepannya bisa mengembangkan plagiarisme kode program bahasa lainnya.

2.2 Sistematika Penulisan

Sistematika penulisan dibutuhkan untuk memberikan susunan skripsi secara ringkas yang terkandung dalam skripsi, sehingga mempermudah dalam pemahaman dan pembahasannya. Adapun sistematika yang digunakan penelitian ini terbagi kedalam beberap bagian, yaitu :

BAB I PENDAHULUAN

Pada bab ini berisikan tentang uraian latar belakang, rumusan masalah, batasan masalah, maksud dan tujuan penelitian, metodologi penelitian serta sistematika penulisan.

BAB II LANDASAN TEORI

Pada bab ini berisikan tentang tinjauan pustaka, serta landasan teori yang memiliki keterikatan terhadap penelitian dalam skripsi ini.

BAB III METODE PENELITIAN

Bab ini membahas tentang metode-metode yang digunakan dalam penelitian ini.

BAB 1V HASIL DAN PEMBAHASAN

Bab ini membahas tentang hasil dari penerapan bert untuk mendeteksi plagiarisme yang sudah dilakukan oleh peneliti.

BAB V PENUTUP

Bab ini merupakan bab terakhir yang memuat kesimpulan-kesimpulan dari penelitian yang telah dilakukan dan berisikan saran-saran yang berguna untuk penelitian selanjutnya.

DAFTAR PUSTAKA