BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi telah mengubah cara mendapatkan suatu informasi, terutama adanya kemunculan internet dan buku elektronik dalam era digital [1]. Publikasi digital menjadi salah satu wujud nyata dari perubahan ini [2]. Hal ini memungkinkan penyebaran dan akses informasi dengan lebih cepat, luas, dan efisien. Dengan publikasi digital, informasi dapat disajikan dalam berbagai format, seperti teks, gambar, dan multimedia, yang memungkinkan pengguna untuk meraih pemahaman yang lebih mendalam [2]. Dengan demikian, publikasi digital tidak hanya mencerminkan kemajuan teknologi, tetapi juga mengubah cara pandang penerbitan dan menghadirkan era baru dalam penyebaran dan konsumsi informasi.

Kehadiran teknologi dalam pencarian informasi telah secara mendasar mengubah cara kita mengelola dan menyimpan data [3]. Dengan pertumbuhan secara terus-menerus dalam jumlah informasi yang tersedia, kebutuhan akan filter informasi menjadi semakin penting [4]. Filterisasi informasi menjadi kunci untuk mengatasi masalah ini. Hal ini menyebabkan kebutuhan akan filter informasi menjadi semakin penting ditengah banyaknya sumber informasi yang tersedia secara daring maupun luring. Tantangannya terletak pada kemampuan untuk membedakan antara informasi yang valid dan tidak valid, serta memastikan bahwa informasi yang diterima dapat dipercaya. Proses penyaringan informasi juga memerlukan keterampilan agar dapat mengakses pengetahuan yang akurat dan relevan.

Pada Lembaga Penelitian dan Pengabdian Masyarakat Universitas Multimedia Nusantara (LPPM UMN) memiliki tujuan untuk meningkatkan efisiensi kerja dengan cara melakukan pengelompokkan data artikel penelitian dosen yang berada di UMN. Pengelompokkan ini bertujuan untuk meningkatkan efisiensi kerja karyawan LPPM UMN dikarenakan data artikel yang ada dapat disusun menjadi kategori-kategori yang relevan sehingga mereka dengan cepat dapat melakukan analisis data yang dibutuhkan. Sebelumnya karyawan LPPM UMN melakukan pengelompokkan kategori-kategori tersebut secara manual (Lampiran 1, W02). Pengelompokkan ini dilakukan dengan cara menganalisis judul artikel yang telah dipublikasikan dosen UMN, sehingga hal ini membutuhkan waktu yang cukup lama

dikarenakan banyaknya artikel yang telah dipublikasikan (Lampiran 1, W02 dan W03).

Pada UMN terdapat kurang lebih 200 dosen, dimana dosen-dosen UMN ini telah menerbitkan lebih 3000 artikel yang telah dipublikasikan (Lampiran 1, W06). Penelitian ini dilakukan dengan mengelompokkan artikel-artikel yang telah dipublikasikan oleh dosen UMN menjadi 17 kategori dokumen *United Nations Sustainable Development Goals* (UN SDG) yang telah dikeluarkan oleh Perserikatan Bangsa-Bangsa (PBB) [5]. Tujuan dari PBB menerbitkan 17 kategori UN SDG ini ialah untuk menangani berbagai masalah signifikan di dunia. 17 kategori UN SDG ini mencakup beragam isu seperti, kemiskinan, kelaparan, pendidikan, kesehatan, kesetaraan gender, air bersih, energi terjangkau, pekerjaan layak, dan perlindungan lingkungan [6].

17 kategori UN SDG ini telah menjadi agenda yang diterima dan diakui oleh seluruh negara anggota Perserikatan Bangsa-Bangsa (PBB) serta berbagai pihak pemangku kepentingan, termasuk pemerintah, sektor swasta, dan organisasi masyarakat sipil [5] (Lampiran 1, W04). Sehingga hal ini membuat 17 kategori UN SDG tersebut menjadi parameter utama di UMN yang berguna untuk meningkatkan akreditasi fakultas dan prodi (Lampiran 1, W01 dan W05). Dengan adanya penelitian ini, proses akreditasi terhadap fakultas dan program studi yang ada pada kampus UMN akan menjadi lebih mudah dilakukan sehingga dapat meningkatkan kinerja karyawan LPPM UMN dalam melakukan proses akreditasi.

Algoritma pencarian merupakan solusi dalam membantu menyaring dan menemukan informasi yang sesuai dengan kebutuhan. Pada penelitian yang dikemukakan oleh Jingzhou Liu yang melakukan penelitian terhadap algoritma *Convolutional Neural Networks* (CNN) algoritma ini digunakan untuk melakukan *multilabel text classification* [7]. Penelitian ini dilakukan untuk meningkatkan performa dari CNN sehingga dapat melakukan *multilabel text classification* dengan optimal. Hal ini dilakukan dengan cara melakukan *testing* dan *training* yang dilakukan dengan berbagai macam dataset untuk meningkatkan kualitas dari algoritma CNN sehingga meningkatkan hasil klasifikasinya. Terdapat penelitian yang dikemukakan oleh Samuel Rodriguez Medina, yang membandingkan beberapa algoritma pada penelitiannya, dimana nilai *AUROC* menjadi tolak ukur dalam penelitian ini. Penelitian tersebut membandingkan beberapa algoritma *machine learning* dan *deep learning* untuk mendapatkan nilai *AUROC* yang terbaik [8]. Penelitian yang dikemukakan oleh Santiago Gonzales yang membandingkan keakuratan dalam melakukan *text classification* dimana penelitian

ini membandingkan alogirtma BERT dengan algoritma pada *machine learning* seperti, *multinominal naive bayes, linear support vector classifier*, dan *logistic regression* [9]. Pada penelitian algoritma BERT memperoleh hasil terbaik dalam melakukan keakuratan dibandingkan algoritma yang lain. Hal ini dapat diketahui karena hasil dari algoritma BERT paling mendekati kaggle *score* yang menjadi tolak ukur tingkat akurasi pada penelitian ini.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, berikut rumusan masalah yang terdapat pada penelitian ini.

1. Bagaimana cara mengimplementasikan algoritma BERT dalam melakukan *text classification* pada topik penelitian dosen di UMN?
2. Bagaimana hasil akurasi dari penerapan algoritma BERT untuk melakukan *text classification*?

1.3 Batasan Permasalahan

Berikut merupakan batasan dari penelitian yang sedang dilakukan.

1. Penelitian ini menggunakan data dari artikel penelitian dosen yang telah dipublikasikan dari tahun 2018 - 2023 oleh dosen UMN secara keseluruhan.
2. Model BERT yang digunakan merupakan model yang telah dilatih dengan data berbahasa Inggris sehingga mendapatkan hasil yang lebih akurat dalam mengenali dan memahami bahasa Inggris. Hal ini disebabkan karena BERT dibangun dengan dataset yang melibatkan sumber daya berbahasa Inggris, sehingga membuatnya lebih efektif dalam mengenali dan memahami bahasa tersebut daripada bahasa lain.

1.4 Tujuan Penelitian

Tujuan dari penelitian yang dilakukan ini adalah sebagai berikut.

1. Mengimplementasikan algoritma BERT dalam melakukan *text classification* pada data artikel yang telah dipublikasikan oleh dosen UMN.
2. Mengukur hasil akurasi dari penerapan alogritma BERT dalam mengelompokkan artikel dosen menjadi 17 kategori UN SDG.

1.5 Manfaat Penelitian

Terdapat manfaat yang diperoleh dari penelitian ini yaitu.

1. Bagi peneliti, penelitian yang dilakukan ini memiliki manfaat untuk menambah wawasan dan melakukan penerapan algoritma BERT dalam melakukan *text classification*.
2. Bagi pengguna, penelitian ini berguna untuk membantu karyawan LPPM UMN melakukan proses akreditasi terhadap program studi dan fakultas yang berada pada UMN.

1.6 Sistematika Penulisan

Berisikan uraian singkat mengenai struktur isi penulisan laporan penelitian, dimulai dari Pendahuluan hingga Simpulan dan Saran.Sistematika penulisan laporan adalah sebagai berikut:

Bab 1 PENDAHULUAN
Dalam bab ini, latar belakang masalah, rumusan masalah, batasan masalah, serta tujuan dan manfaat penelitian akan dijelaskan pada bab ini. Serta cara melakukan penulisan sistematika pada laporan skripsi ini.

Bab 2 LANDASAN TEORI
Dalam bab ini, memuat teori dan studi yang digunakan selama penelitian ini berlangsung dalam membuat model algoritma BERT dengan teori yang mendukung penelitian ini.

Bab 3 METODOLOGI PENELITIAN
Dalam bab ini, terdapat metodologi penelitian yang digunakan dalam pembuatan model yaitu, terdapat tahapan perancangan yang berisikan *flowchart*

Bab 4 HASIL DAN DISKUSI
Dalam bab ini, hasil dari implementasi pada penelitian yang dilakukan akan dijelaskan. Penjelasan akan berupa *output* dari penelitian yang telah dilakukan dan analisis pada bab ini.

Bab 5 SIMPULAN DAN SARAN

Dalam bab ini, terdapat hasil dari penelitian yang telah dilaksanakan berupa kesimpulan dan saran terhadap penelitian untuk kedepannya.

No	Judul Penelitian	Algoritma	Hasil
1	*Comparing BERT against traditional machine learning text classification*	*Bidirectional Encoder Representations from Transformers* (BERT) *& Term Frequency - Inverse Document Frequency* (TD-IDF)	Algoritma BERT mendapatkan hasil yang lebih baik daripada algoritma TD-IDF pada melakukan analisis sentimen yang mendapatkan nilai akurasi 90.90% sedangkan algoritma TD-IDF hanya mendapatkan nilai 84.80%
2	*Comparing BERT against traditional machine learning text classification*	*Bidirectional Encoder Representations from Transformers* (BERT) *& Voting Classifier*	Algoritma BERT berhasil mengungguli algoritma voting classifier dalam melakukan text classification yang mendapatkan nilai akurasi sebesar 93.87% sedangkan voting classifer mendapatkan nilai 90.07%
3	*Multi-Label Text Classification with Transfer Learning for Policy Documents*	*Bidirectional Encoder Representations from Transformers* (BERT) *& logistic regression & multinomial naive bayes*	Algoritma BERT mendapatkan hasil nilai AUROC terbaik yaitu sebesar 0.92 yang menjadi tolak ukur dalam penelitian ini mengungguli algoritma logistic regression yang mendapatkan nilai 0.89 dan algortima multinomial naive bayes yang mendapatkan nilai 0.82

Tabel 1.1. Perbandingan Algoritma