BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian mengenai klasifikasi kategori berita telah dilakukan menggunakan metode yang beragam. Penggunaan metode yang beragam dapat menghasilkan akurasi yang berbeda-beda. Oleh karena itu, perlu mencari informasi dan literatur yang relevan untuk mendukung penelitian yang berjudul "Pengaruh Teknik Augmentasi Dalam Klasifikasi Berita Berhasa Inggris Menggunakan Algoritma *BERT*".

Tabel 2.1 Penelitian Terdahulu

No	Penulis	Judul	Metode	Hasil
1	Asep Ripa'i, Firman Santoso, dan Farihin Lazim (2024)	Deteksi Berita Hoax dengan Perbandingan Website Menggunakan Pendekatan *Deep Learning* Algoritma *BERT*	*IndoBERT, SVM,* dan *Random Forest*	Hasil pada penelitian yaitu model *BERT* menghasilkan akurasi sebesar 99% (akurasi = 0.99, *ROC-AUC* = 0.99), *BERT* lebih unggul dibandingkan *SVM* dan *Random Forest* yang menghasilkan akurasi sebesar (96%) dan (94%).
2	Muhammad Basil Musyaffa	Deteksi Spam Berbahasa	*BERT*	Hasil percobaan menggunakan

	Amin, Gibran Hakim, Muhammad Taufik Maulana, Muhammad Fajrul Alwan, Hanna Shafira Anggraheni, Muhammad Jilan Naufal, dan Novanto Yudistira (2024)	Indonesia Berbasis Teks Menggunakan Model *BERT*		*IndoBERT* memiliki nilai akurasi sebesar 98% pada dataset SMS maupun pada dataset Email. Sedangkan, dengan menggunakan model *MultilingualBERT*, diperoleh nilai akurasi sebesar 95% pada dataset SMS maupun pada dataset Email. Data yang sudah dilatih kemudian disimpan dan digunakan untuk mendeteksi pesan SMS maupun Email, apakah pesan tersebut termasuk ke dalam kelas spam atau bukan spam.
3	Luffi Septian, Teguh Aljauza, dan Christina Juliane (2024)	Analisis Sentimen Putusan Mahkamah Konstitusi	*IndoBERT*	Hasil penelitian menggunakan model *IndoBERT* tanpa augmentasi data mendapat akurasi

		Terhadap Batas Usia Capres Dan Cawapres Menggunakan *IndoBERT*		sebesar 0.81 dan F1-score 0.58. Penggunaan *SMOTE* berhasil meningkatkan F1-score. Selanjutnya, setelah diterapkan augmentasi *Random Swap* akurasi dan F1-score mengalami peningkatan menjadi 0.90.
4	Iftitah Athiyyah Rahma dan Lya Hulliyyatus Suadaa (2023)	Penerapan *Text Augmentation* untuk Mengatasi Data yang Tidak Seimbang pada Klasifikasi Teks Berbahasa Indonesia	*SVM* dan *IndoBERT*	Hasil eksperimen menunjukkan bahwa teknik *Back Translation* memberikan peningkatan F1-score sebesar 8%, sedangkan *Synonym Replacement* hanya meningkatkan sebesar 5%.
5	Luthfi Atikah, Diana Purwitasari, dan Nanik Suciati (2022)	Deteksi Kejadian Lalu Lintas pada Teks Twitter dengan	*CNN* dan *LTSM*	Penelitian ini menggunakan *CNN* dan *LSTM* dengan *word embedding word2vec* dan *fastText.*

		Pendekatan Klasifikasi Multi-Label Berbasis *Deep Learning*		Penelitian ini juga menerapkan augmentasi data seperti, *synonym replacement, random exchange,* dan *random deletion.* Hasil penelitian menunjukkan bahwa augmentasi dapat meningkatkan akurasi dari 0,75 tanpa augmentasi menjadi 0,95 menggunakan augmentasi, dengan kombinasi *LSTM* dan *fastText*.

Pada Tabel 2.1 menunjuknan beberapa penelitian yang dijadikan sebagai acuan untuk melakukan penelitian ini. Penelitian oleh Asep Ripa'i dkk. (2024) menunjukkan bahwa IndoBERT unggul dibandingkan SVM dan Random Forest dalam mendeteksi berita hoaks, dengan akurasi mencapai 99%, namun belum menggunakan teknik augmentasi [10]. Penelitian oleh Muhammad Basil Musyaffa Amin dkk. (2024) juga menunjukkan performa tinggi IndoBERT untuk deteksi spam dengan akurasi 98%, tetapi tidak membahas penggunaan augmentasi data [11]. Sementara itu, Luffi Septian dkk. (2024) menerapkan teknik *Random Swap* pada analisis sentimen, yang berhasil meningkatkan akurasi dari 0,81 menjadi 0,90, meskipun hanya menggunakan satu jenis augmentasi [7].

Penelitian lain oleh Iftitah Athiyyah Rahma dan Lya Hulliyyatus Suadaa (2023) membandingkan dua teknik augmentasi *Back Translation* dan *Synonym*

Replacement yang terbukti mampu meningkatkan F1-score, tetapi belum menggunakan model *BERT* secara mendalam [8]. Adapun penelitian oleh Luthfi Atikah dkk. (2022) menerapkan beberapa teknik augmentasi dalam model CNN dan LSTM untuk mendeteksi kejadian lalu lintas, dengan peningkatan akurasi dari 0,75 menjadi 0,95 [12].

Berdasarkan acuan penelitian sebelumnya, belum ada penelitian yang menggabungkan tiga teknik augmentasi yaitu *Synonym Replacement*, *Back Translation*, dan *Random Swap* menggunakan model *BERT*. Dengan demikian, penelitian ini dilakukan untuk mengatasi kekurangan tersebut dan menganalisis dampak teknik augmentasi pada klasifikasi berita berbahasa Inggris.

2.2 Berita

Berita merupakan sebuah informasi tentang peristiwa atau kejadian yang sedang terjadi, dan dianggap penting, atau menarik untuk disampaikan kepada publik. Berita juga dapat diartikan sebagai informasi yang sebelumnya tidak diketahui, yang kemudian disampaikan agar orang dapat mengetahuinya [13]. Berita biasanya ditulis oleh seorang jurnalis baru kemudian disampaikan kepada masyarakat melalui berbagai saluran media seperti, surat kabar, televisi, radio, maupun platform digital [14]. Berita juga dapat menjadi sarana untuk menambah wawasan dan pengetahuan masyrakat tentang beberapa topik.

Berita memiliki beberapa fungsi bagi kehidupan masyarakat. Pertama, dalam dunia pendidikan berita dapat memberikan pengetahuan baru dan wawasan yang luas terhadap sesuatu yang belum mereka ketahui seblumnya. Kedua, melalui informasi berupa olahraga, budaya, dan selebriti dapat membuat hiburan bagi masayrakat yang bisa mengurangi stress dan memberikan rasa senang. Ketiga, berita juga dapat berfungsi sebagai pengawasan terhadap kinerja pemerintah selama ini, karena dengan adanya berita kinerja pemerintah dapat dipantau dan menjadi lebih transparan.

2.3 Text Minning

Text Minning merupakan proses pengambilan informasi yang berguna dari

kumpulan data teks yang tidak terstruktur dan dapat di analisis, sehingga memungkinkan untuk menemukan wawasan yang tersembunyi di dalamnya. Penerapan *text mining* umumnya meliputi analisis sentiment, *information retrival, information extraction,* dan *clustering* [15].

Klasifikasi adalah proses pengelompokan data atau informasi ke dalam kategori yang telah ditentukan [16]. Pada berita klasifikasi digunakan untuk mengelompokan berita kedalam beberapa kategori seperti politik, olahraga, hiburan, bisnis, dan sebagainya. Klasifikasi membuat berita menjadi lebih mudah untuk dikses dan dipahami oleh pembaca. Proses ini melibatkan algoritma machine learning yang berfungsi untuk menganalisis berita dan menentukan kategori yang relevan berdasarkan kategori yang terdapat pada berita.

2.4 Augmentasi Data

Augmentasi merupakan sebuah teknik yang digunakan untuk menambah variasi data agar dapat meningkatkan kualitas model [17]. Teknik ini biasanya digunakan untuk mengatasi masalah data yang terbatas atau tidak seimbang [18]. Augmentasi dapat diterapkan pada beberapa jenis data seperti data citra, data teks, data tabular, dan data suara.

Dalam pemrosesan bahasa alami (NLP), augmentasi data teks bertujuan untuk menambah variasi data teks tetapi tidak menghilangkan makna aslinya. Beberapa teknik yang biasanya digunakan dalam data teks antara lain *synonym replacement*, *back translation, random swap,* dan sebagainya. Pada *synonym replacement*, beberapa kata dalam teks akan diganti dengan sinonimnya tetapi tidak menghilangkan makna utamanya. *Back translation* dilakukan dengan menerjemahkan teks ke bahasa lain, lalu diterjemahkan kembali ke bahasa asal untuk mendapatkan struktur kalimat yang berbeda tetapi tetap memiliki makna yang sama [8]. *Random swap* menukar posisi dua kata secara acak tanpa mengubah makna asli kalimat [7].

2.5 BERT

Algoritma BERT (Bidirectional Encoder Representations from Transformers) merupakan teknologi revolusioner dalam bidang *Natural Language*

Processing (NLP). *BERT* menghasilkan representasi pada setiap kata dalam kalimat sebagai *output* [19]. *BERT* menggunakan pendekatan pembelajaran dua arah yaitu bidirectional learning yang mempertimbangkan token sebelum dan sesudah dalam sebuah urutan. Proses embedding pada *BERT* dilakukan melalui tiga skema yaitu *vocab embedding, segment embedding,* dan *position embedding*. Pendekatan ini dianggap mampu memahami makna sebuah kata berdasarkan posisinya dalam kalimat [20].

Untuk meningkatkan kinerja *BERT* dalam tugas-tugas *NLP* setelah dilatih pada dataset besar, diperlukan *fine-tuning.* Pada proses *fine-tuning* ditambahkan lapisan output khusus dan melakukan pembelajaran, yang secara konsisten meningkatkan kemampuan model dalam berbagai tugas *NLP* [21].

2.6 Pengujian Peforma Model

Eror analisis adalah proses penting dalam pengembangan model klasifikasi yang bertujuan untuk memahami jenis-jenis kesalahan yang dibuat oleh model dan mencari tahu penyebabnya. Dengan melakukan eror analisis, dapat mengevaluasi kelemahan model serta memperbaiki performanya. Untuk memperkuat analisis ini, digunakan juga metrik evaluasi sebagai berikut:

a. Akurasi menunjukan presentase dari total prediksi yang benar pada seluruh data yang diuji.

$$Akurasi = \frac{True\ Positive\ +\ True\ Negatives}{Total\ Sample}$$

b. Precision mengukur seberapa banyak dari semua prediksi positif yang benar-benar positif.

$$Precision = \frac{True\ Positive}{True\ Positive\ +False\ Positive}$$

c. Recall mengukur seberapa banyak kasus positif yang benar-benar terdeteksi oleh model.

$$Recall\ = \frac{True\ Positive}{True\ Positive\ +False\ Negative}$$

d. F1-Score adalah matrix evaluasi yang menggabungkan nilai dari precision dan recall.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Keterangan :

True Positive (TP) : Jumlah data kelas positif yang diprediksi benar sebagai positif.

True Negative (TN) : Jumlah data kelas negatif yang diprediksi benar sebagai negatif.

False Positive (FP) : Jumlah data kelas negatif yang salah diprediksi sebagai positif.

False Negative (FN) : Jumlah data kelas positif yang salah diprediksi sebagai negatif.

Total Sample : Jumlah total data dalam dataset (TP+TN+FP+FN)