BAB I
PENDAHULUAN

1.1 Latar Belakang

Teknologi informasi menjadi perananan penting dalam kehidupan manusia. Tanpa adanya teknologi, untuk berkomunikasi dan menyampaikan informasi akan sulit bagi manusia [1]. Saat ini teknologi informasi telah menjadi sarana utama dalam penyebaran berita, yang memungkinkan berita dengan cepat disampaikan kepada publik secara luas dan efisien [2].

Berita adalah informasi tentang peristiwa atau kejadian yang sedang terjadi, yang dianggap penting, atau menarik untuk disampaikan kepada publik [3]. Berita biasanya mencakup fakta yang terjadi, dan disusun untuk memungkinkan penguna memahami peristiwa secara jelas dan kronologis. Jenis berita dapat bervariasi, termasuk berita ekonomi, politik, sosial, olahraga, teknologi, dan lain-lain [4].

Seiring meningkatnya jumlah berita yang dihasilkan, beragam kategori berita menjadi semakin luas dan kompleks. Masyarakat kini dapat mengakses berita dalam berbagai topik yang spesifik dan sesuai dengan minat mereka [5]. Namun, dengan volume berita yang sangat besar, pengguna seringkali dihadapkan dengan tantangan dalam menemukan berita yang relevan dengan minat atau kebutuhan mereka.

Klasifikasi berita memainkan peran penting dalam pengelolaan dan penyajian informasi yang efektif, terutama di era digital saat ini di mana volume berita yang tersedia secara online sangat besar. Klasifikasi berita yang baik memungkinkan penyusunan berita yang lebih terstruktur dan memudahkan pengguna dalam mengakses informasi yang diinginkan. Untuk mengatasi masalah ini, teknologi informasi seperti kecerdasan buatan (AI) dan pembelajaran mesin (*machine learning*) mulai digunakan untuk mengotomatisasi klasifikasi berita. Algoritma ini membantu mengelompokkan berita berdasarkan kategori, memungkinkan pengguna untuk menemukan informasi yang mereka butuhkan dengan lebih efisien, serta membantu platform berita menyajikan konten yang lebih

relevan [6]. Namun, masih ada beberapa masalah yang masih dihadapi, seperti ketidakseimbangan data dan keterbatasan variasi data. Teknik augmentasi data menjadi salah satu solusi yang dapat diterapkan untuk meningkatkan performa model dengan menambah variasi pada dataset. Kombinasi antara preprocessing teks, augmentasi data, dan representasi teks diharapkan dapat meningkatkan akurasi klasifikasi berita dalam berbagai kategori.

Pada penelitian sebelumnya [7] yang membahas tentang analisis sentimen masyarakat terhadap Putusan Mahkamah Konstitusi nomor 90/PUU-XXI/2023, penelitian ini menggunakan data dari media sosial X. Model yang digunakan dalam penelitian inni adalah *IndoBERT*. Penelitian ini mengklasifikasikan opini publik menjadi positif, negatif, atau netral. Hasil dari penelitian ini mendapatkan akurasi sebesar 0,81 dan F1-S 0.58 tanpa penerapan teknik augmentasi. Setelah menerapkan teknik augmentasi *random swap*, akurasi dan F1-Score mengalami penigkatan menjadi 0,90. Namun, penelitian ini memiliki keterbatasan pada ketergantungan data yang hanya berasal dari Twitter/X, yang mungkin tidak mewakili pandangan masyarakat secara keseluruhan, serta potensi bias bahasa dalam teks yang digunakan.

Pada penelitian lain [8] yang membahas tentang penerapan augmentasi teks untuk mengatasi data yang tidak seimbang pada klasifikasi teks berbahasa Indonesia. Hasil penelitian ini menunjukkan bahwa penerapan teknik *back translation* terbukti lebih efektif dibandingkan dengan *synonym replacement,* dengan peningkatan F1-Score mencapai 8%, sementara teknik *synonym replacement* hanya mencapai 5%. Penelitian ini juga mencatat potensi masalah *overfitting* pada data latih seiring dengan bertambahnya jumlah data augmentasi. Namun, terdapat keterbatasan pada penelitian ini yaitu, kata-kata tidak baku pada dataset teks informal yang dapat menurunkan kualitas teks.

Penelitian selanjutnya [9] yang membahas tentang klasifikasi multi-label terhadap 18.000 data dari akun Twitter di Surabaya untuk mendeteksi situasi lalu lintas. Metode yang digunakan dalam penelitian ini adalah *CNN* dan *LSTM* dengan *word embedding word2vec* dan *fastText*. Penelitian ini juga menggunakan teknik augmentasi data yaitu, *synonym replacement, random exchange,* dan *random*

deletion. Eksperimen dilakukan menggunakan tiga skenario yaitu, menggunakan 3 label, 4 label, dan 5 label untuk membandingkan data augmentasi dan non-augmentasi. Hasil penelitian menunjukkan bahwa augmentasi dapat meningkatkan akurasi dari 0,75 tanpa augmentasi menjadi 0,95 menggunakan augmentasi, dengan kombinasi *LSTM* dan *fastText*.

Berdasarkan penelitian sebelumnya, maka fokus penelitian untuk pengklasifikasian berita terhadap beberapa kategori yaitu, *sport, business, politics, tech, dan entertainment.* Diharapkan dengan proses klasifikasi ini dapat memberikan solusi terhadap pengguna untuk memilih berita sesuai dengan apa yang diinginkan. Metode yang digunakan dalam klasifikasi berita adalah *Deep Learning* dengan Algoritma *BERT (Bidirectional Encoder Representations from Transformers)*. Pada penelitian ini akan dilakukan preprocessing data yaitu, *cleaning, stemming, tokenizing, stopwords,* dan menggunakan *BERT* untuk representasi teks. Penelitian ini juga akan membandingkan tiga teknik augmentasi data yaitu, *Synonymn Replacement, Back Translation,* dan *Random Swap* untuk mengetahui augmentasi data yang menghasilkan akurasi terbaik.

1.2 Rumusan Masalah

Berdasarkan latar belakang diatas maka peneliti mengangkat rumusan masalah sebagai berikut :

1. Bagaimana kinerja dan peningkatan akurasi *BERT* dalam klasifikasi berita dalam beberapa kategori?
2. Apakah proses *preprocessing* teks seperti, cleaning, stemming, tokenizing, stopwords dan teknik *BERT* dapat meningkatkan akurasi dalam klasifikasi berita?
3. Teknik augmentasi data mana yang menghasilkan akurasi terbaik dalam klasifikasi berita.

1.3 Tujuan Penelitian

1. Mengembangkan model klasifikasi berita berbasis deep learning dengan menggunakan algoritma *BERT* untuk mencapai akurasi tinggi.
2. Melakukan *preprocessing* teks yaitu, *cleaning, stemming, tokenizing, stopwords*, dan menggunakan *BERT* untuk ekstraksi fitur.
3. Membandingkan tiga teknik augmentasi data, yaitu *Synonym Replacement, Back Translation,* dan *Random Swap*, untuk proses klasifikasi berita.

1.4 Batasan Masalah

1. Penelitian ini hanya akan menggunakan algoritma *BERT* sebagai metode klasifikasi utama.
2. Dataset yang digunakan terbatas pada berita dalam bahasa Inggris dengan kategori *sport, business, politics, tech, dan entertainment.*
3. Hanya membandingkan tiga teknik augmentasi data saja, yaitu *Synonymn Replacement, Back Translation,* dan *Random Swap*.
4. Representasi teks hanya menggunakan *BERT*.