

BAB I

PENDAHULUAN

1.1 Latar Belakang

Berita adalah representasi dari realitas sosial yang merefleksikan berbagai isu yang masih hangat dan sering menjadi bahan perbincangan di tengah masyarakat. Penyampaian berita dapat berlangsung melalui komunikasi lisan dari satu individu ke individu lainnya. Di samping itu, penyebaran informasi juga dapat dilakukan melalui berbagai media konvensional seperti koran, radio, dan televisi. Namun, seiring perkembangan teknologi, kini hadir media baru yaitu internet, yang menjadi sarana utama dalam distribusi informasi di era digital saat ini [1]. Pesatnya perkembangan teknologi informasi saat ini membuat masyarakat tidak dapat terlepas dari kebutuhan untuk mendapatkan informasi melalui internet [2].

Berbagai macam informasi dan berita tersebar secara luas melalui internet yang menyebabkan jumlah pengguna internet semakin meningkat. Menurut Direktur Eksekutif Nielsen Media, Hellen Katherina menyebutkan bahwa “Tingginya frekuensi pengguna internet diantara pembaca media cetak mencapai 86%” Sehingga tidak dapat dipungkiri bahwa masyarakat memiliki kebutuhan untuk mendapatkan informasi berita melalui internet [3]. Berbeda dengan zaman dahulu pada saat berita hanya dapat dilihat melalui surat kabar, kini Masyarakat dapat dengan mudah mengakses portal berita melalui *smartphone* mereka.

Pada survei yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), portal berita menempati peringkat kelima sebagai konten internet yang paling sering diakses oleh Masyarakat [4]. Menurut ketua umum APJII Muhammad Arif menyatakan “pengguna internet Indonesia tahun 2024 mencapai 221.563.479 jiwa, maka tingkat penetrasi internet Indonesia menyentuh angka 79,5%. Dibandingkan dengan periode sebelumnya, maka ada peningkatan 1,4%”. Sehingga dapat dipungkiri bahwasanya peningkatan konsisten grafik tren penetrasi internet Indonesia dalam lima tahun terakhir naik secara signifikan.

Pada umumnya, berita yang ada dalam portal berita dikelompokkan dalam beberapa kategori, seperti berita politik, hiburan, teknologi, olahraga, ekonomi, kesehatan dan lain-lain [5]. Situs-situs berita yang belum dikelola secara

professional tidak melakukan pengorganisasian berita sesuai dengan topik pembahasan berita tersebut dan terdapat ketidakakuratan isi berita dengan judulnya. Salah satu situs yang menjadi sumber informasi utama di Indonesia adalah Kompas.com, yang setiap harinya menghasilkan ribuan artikel berita dengan berbagai macam topik.

Dengan banyaknya jumlah berita yang diunggah setiap harinya, kebutuhan akan pengelompokan berita berdasarkan topik menjadi semakin penting. Namun, pengelompokan topik secara manual memiliki kendala seperti keterbatasan waktu dan efisiensi, terutama karena volume berita yang terus meningkat [6]. Salah satu cara yang dapat dilakukan dengan cepat dan dapat dipahami oleh penerima informasi adalah dengan melakukan klasifikasi teks pada berita berdasarkan topiknya. Oleh karena itu, diperlukan metode otomatis yang mampu mengklasifikasikan berita dengan cepat dan akurat [7].

Klasifikasi teks merupakan salah satu tugas penting dalam bidang Natural Language Processing (NLP), yang bertujuan untuk mengelompokkan dokumen teks ke dalam kategori tertentu berdasarkan isi atau tema yang dibahas [8]. Dalam konteks Text Mining, klasifikasi digunakan untuk menentukan label atau kelas dari suatu data yang belum diketahui kategorinya [5]. Selain klasifikasi, Text Mining juga mencakup proses lain seperti clustering dan information extraction [9]. Berbagai metode telah digunakan oleh para peneliti untuk melakukan klasifikasi teks, dan dalam penelitian ini dipilih metode FastText. FastText merupakan algoritma pembelajaran yang dikembangkan oleh Facebook AI Research (FAIR) [10]. Meskipun FastText telah banyak diaplikasikan dalam klasifikasi teks berbagai bahasa, penggunaannya pada berita berbahasa Indonesia—khususnya yang berasal dari Kompas.com—masih relatif jarang diteliti secara mendalam. Padahal, dengan banyaknya data yang dimiliki Kompas.com, penerapan model FastText berpotensi menghasilkan klasifikasi yang lebih akurat dan relevan.

Penelitian sebelumnya [10] membandingkan kinerja metode word embedding, yaitu *Word2Vec*, *GloVe*, dan *FastText*, dalam klasifikasi teks menggunakan algoritma *Convolutional Neural network (CNN)*. Perbandingan ini dilakukan untuk mengukur efektivitas masing-masing metode dalam menangkap makna semantik dan sintaktik kata pada teks tidak terstruktur. Hasil pada penelitian ini menunjukkan

bahwa metode *FastText* memberikan kinerja terbaik di antara ketiga metode *word embedding*, dengan nilai F-Measure sebesar 0.979 untuk dataset 20 *Newsgroups* dan 0.715 untuk dataset *Reuters Newswire*. Hal ini disebabkan oleh kemampuan *FastText* dalam menangkap informasi subword, sehingga dapat menangani kata-kata yang tidak muncul dalam data penelitian (out-of-vocabulary) [10].

Pada penelitian sebelumnya [5]. Penelitian ini bertujuan untuk membangun sistem klasifikasi berita online secara otomatis menggunakan metode *Text Mining* dan *Algoritma Multinomial Naïve Bayes*. Sistem ini dirancang untuk mengatasi kendala pengelompokan berita secara manual yang memakan waktu dan merepotkan, sehingga dapat mempercepat proses pengkategorian berita dalam tujuh kategori yang telah ditentukan. Dari hasil penelitian ini, sistem klasifikasi berita secara otomatis berhasil dikembangkan menggunakan metode *Text Mining* dan *Algoritma Multinomial Naïve Bayes*, dengan hasil pengujian yang menunjukkan performa tinggi. Model klasifikasi yang dibuat memiliki Tingkat akurasi, *presisi*, *recall*, dan f1-score masing-masing sebesar 96% menggunakan dataset sebanyak 10.500 artikel berita dari dua portal utama, yaitu Kompas dan Tempo [5].

Penelitian yang lain juga [1]. Penelitian ini bertujuan membangun sistem klasifikasi topik berita berbahasa Indonesia secara otomatis menggunakan metode *Multilayer Perceptron (MLP)*. Tujuan akhirnya adalah untuk menilai efektifitas metode *Multilayer Perceptron* berdasarkan nilai akurasi, *precision*, *recall*, dan f1-score dalam mengelompokkan berita sesuai kategorinya. Hasil dari penelitian ini menunjukkan bahwa sistem klasifikasi topik berita berhasil dikembangkan menggunakan metode tersebut dengan struktur model yang terdiri dari tiga layer, yaitu input, hidden, dan output, serta parameter seperti 75 epoch, learning rate 0.001, dan optimizer Stochastic Gradient Descent (SGD). Model klasifikasi mencapai performa terbaik dengan nilai F1-Score sebesar 77,44%, menunjukkan bahwa kombinasi metode TF-IDF tanpa stemming dan multilayer perceptron efektif untuk menangani klasifikasi berita berbahasa Indonesia [1].

Penelitian yang lainnya juga [11]. Tujuan dari penelitian ini adalah untuk mengembangkan metode klasifikasi yang dapat mendeteksi cyberbullying pada platform Twitter dengan mengimplementasikan algoritma Random Forest serta menggunakan penyematan kata *FastText*. Dengan adanya pembatasan karakter

dalam setiap tweet, variasi kata menjadi tantangan dalam klasifikasi, sehingga metode *FastText* diterapkan untuk mengatasi ketidaksesuaian kosakata. Melalui penelitian ini, diharapkan dapat membantu menciptakan ruang media sosial yang lebih aman dengan mendeteksi cyberbullying secara akurat. Hasil penelitian menunjukkan bahwa metode Random Forest dengan penyematan kata *FastText* mencapai akurasi sebesar 88,17% pada skenario terbaik ($K=20$) saat pengujian dengan data latih. Namun, ketika diuji dengan dataset yang berbeda dari hasil crawling, akurasinya menurun menjadi 61,61%. Meskipun demikian, model tetap menunjukkan performa yang baik dalam mengklasifikasikan cyberbullying pada Twitter. Penelitian ini memberikan kontribusi dalam meningkatkan deteksi cyberbullying serta mendorong penggunaan teknologi untuk menjaga lingkungan digital yang lebih positif dan aman [11].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengimplementasikan dan mengevaluasi kinerja model *FastText* dalam klasifikasi topik utama pada berita Bahasa Indonesia. Pada penelitian ini dataset yang digunakan diambil dari halaman Kompas.com yang terdiri dari 1050 data berita menggunakan teknik *scraping*. Kemudian, akan dilakukan *EDA (Exploratory Data Analysis)* terhadap data untuk melihat Gambaran data secara keseluruhan. Selanjutnya, data akan melalui tahap Preprocessing dan splitting hingga akhirnya dilakukan klasifikasi menggunakan metode *FastText*. Diharapkan hasil penelitian ini dapat memberikan kontribusi terhadap peningkatan efisiensi dalam pengolahan berita daring. Selain itu, penelitian ini juga bertujuan untuk memanfaatkan kecepatan dan skalabilitas *FastText* dalam memproses data dengan jumlah yang besar.

1.2 Rumusan Masalah

Perumusan masalah dalam penelitian ini adalah sebagai berikut:

1.2.1 Permasalahan

Berita dalam Bahasa Indonesia dari berbagai portal berita daring tersebar dengan berbagai topik utama seperti teknologi, politik, ekonomi, hiburan, dan lainnya. Hal ini menimbulkan tantangan dalam mengelompokkan berita berdasarkan topiknya karena jumlahnya yang terus meningkat. Data berita dari Kompas.com digunakan dalam penelitian ini hanya sebagai sampel atau uji coba

untuk menguji efektivitas metode *FastText*. Namun, pemodelan yang dibuat bertujuan untuk dapat diterapkan pada portal berita lainnya yang memiliki struktur teks serupa.

1.2.2 Metode Usulan

Penelitian ini mengusulkan penggunaan model *FastText*, sebuah Teknik embedding teks yang berbasis subword, untuk melakukan klasifikasi berita utama secara langsung berdasarkan data dari Kompas.com. Teknik ini mampu mengenali struktur internal kata, termasuk kata-kata dengan ejaan yang tidak baku, sehingga diharapkan dapat meningkatkan akurasi dalam klasifikasi berita.

1.2.3 Pertanyaan Penelitian

Berdasarkan uraian permasalahan, maka pertanyaan penelitian sebagai berikut:

1. Bagaimana performa model *FastText* dalam mengklasifikasikan topik utama berita pada Kompas.com?
2. Bagaimana tingkat kesulitan model *FastText* dalam mengklasifikasikan berita yang memiliki kemiripan antar topik pada Kompas.com?

1.3 Tujuan & Manfaat

1.3.1 Tujuan Penelitian

1. Menganalisis dan mengukur performa model *FastText* dalam mengklasifikasikan topik utama berita Bahasa Indonesia berdasarkan evaluasi akurasi, precision, recall, dan f1-score.
2. Mengidentifikasi dan mengevaluasi tingkat kesulitan model *FastText* dalam mengklasifikasikan berita yang memiliki kemiripan antar topik pada Kompas.com.

1.3.2 Manfaat:

1. **Manfaat bagi pembaca:**
 - a. Menambah wawasan pembaca mengenai penggunaan *FastText* untuk klasifikasi teks.
 - b. Dapat dijadikan referensi bagi peneliti yang ingin menggunakan *FastText* untuk klasifikasi teks berbahasa Indonesia.
2. **Manfaat bagi media:**

- a. Membantu media seperti Kompas.com mengoptimalkan pengelompokan berita berdasarkan kategori tertentu.
- b. Meningkatkan efisiensi distribusi berita kepada pembaca dengan klasifikasi yang lebih akurat dan cepat.

1.4 Batasan Penelitian

1. Data yang digunakan adalah berita berbahasa Indonesia yang diambil dari portal Kompas.com.
2. Dataset terdiri dari 4189 data text berita pada Kompas.com.
3. Teknik yang digunakan adalah *FastText*.

1.5 Sistematika Penulisan

Sistematika penulisan terdiri dari tiga bab, yaitu:

BAB I PENDAHULUAN

Mencakup latar belakang, perumusan masalah, tujuan dan manfaat, Batasan penelitian, dan sistematika penulisan.

BAB II KAJIAN PUSTAKA

Berisi teori-teori penunjang yang berkaitan dengan penelitian dan penelitian terdahulu.

BAB III METODE PENELITIAN

Berisi metodologi penelitian, alur penelitian mulai dari preprocessing data, hingga evaluasi.

BAB IV HASIL DAN PEMBAHASAN

Dibagian ini dijelaskan mengenai hasil dari penerapan pemodelan hingga ke sistem yang telah dikembangkan, termasuk pengujian dan evaluasi yang telah dilakukan selama penelitian.

BAB V PENUTUP

Bab ini menjelaskan mengenai hasil kesimpulan yang diperoleh dan memberikan saran untuk pengembangan penelitian lebih lanjut.