BAB I
PENDAHULUAN

Bab ini membahas latar belakang, rumusan masalah, tujuan dan manfaat penelitian, ruang lingkup, serta sistematika penulisan yang digunakan dalam penyusunan skripsi mengenai klasifikasi berita berbahasa Indonesia pada media digital menggunakan metode IndoBERT dan *Support Vector Machine*.

1.1 Latar Belakang

Perkembangan teknologi internet telah mengubah lanskap media di Indonesia secara fundamental. Saputra & Al Siddiq (2020) mencatat bahwa lebih dari 132 juta pengguna internet telah mengubah cara masyarakat Indonesia mengakses informasi, beralih dari media cetak tradisional ke *platform* digital. Media digital telah menjadi sumber informasi utama, terutama bagi generasi muda yang lebih memilih akses *online* daripada surat kabar atau televisi (Bangun, 2018; Bae & Zamrudi, 2018).

Tarihoran dkk. (2022) menemukan bahwa sekitar 50% masyarakat Indonesia kini memperoleh berita melalui Facebook, YouTube, dan Instagram. Popularitas ini didorong oleh kemudahan akses dan sifat interaktif *platform* digital. Rodilosso (2024) menjelaskan bahwa algoritma media sosial yang menyesuaikan konten dengan minat pengguna membuat informasi lebih menarik dan relevan, namun seringkali mengutamakan aspek emosional dibanding keakuratan fakta (Khan dkk., 2021).

Pergeseran ke media sosial menimbulkan risiko signifikan berupa penyebaran berita hoaks. Syam & Nurrahmi (2020) menemukan bahwa banyak pengguna, termasuk mahasiswa, kesulitan membedakan berita valid dari hoaks. Algoritma yang mengutamakan keterlibatan pengguna dibanding keakuratan informasi memperburuk disinformasi, berdampak negatif pada opini publik dan stabilitas sosial (Adesope & Musa, 2023).

Faizah (2020) mengidentifikasi bahwa media sosial menjadi lahan subur persebaran berita palsu karena memungkinkan penyebaran informasi sangat cepat tanpa verifikasi ketat. Aktor politik kerap memanfaatkan hoaks untuk membangun narasi tertentu, terutama saat kampanye pemilu (Hartanto & Purnomo, 2023), yang efektif menggerakkan opini publik dan memicu polarisasi sosial Afifi dkk. (2020).

Penyebaran hoaks tidak hanya mempengaruhi opini publik, tetapi juga merusak kepercayaan terhadap institusi resmi. Meningkatnya berita palsu telah melemahkan kepercayaan publik terhadap pemerintah dan media arus utama (Smith, 2020). Sebagian besar hoaks di Indonesia berkaitan dengan isu politik, kesehatan, dan agama yang berpotensi menimbulkan ketegangan sosial, diperparah oleh rendahnya literasi digital masyarakat (Afifi dkk., 2020; Syam & Nurrahmi, 2020).

WhatsApp, Facebook, dan Instagram menjadi saluran utama penyebaran hoaks di Indonesia (Abdillah dkk., 2024). Sekitar 40% masyarakat Indonesia memiliki kebiasaan membagikan berita hanya berdasarkan judul tanpa membaca konten lengkap (Murfianti (2017). Perilaku tersebut tentu berkontribusi langsung terhadap penyebaran berita hoaks karena informasi yang dibagikan belum terverifikasi kebenarannya.

Fenomena ini berpotensi menimbulkan dampak luas yang lebih luas dan masif, sehingga sangat diperlukan metode yang efektif untuk mengklasifikasikan berita di media digital. Salah satu metode yang sering digunakan dalam konteks *Natural Language Processing* (NLP) adalah *Support Vector Machine* (SVM). Dalam penelitian terkait, Pratama & Tjahyanto (2022) menemukan bahwa dalam kasus analisis sentimen dengan kedua metode menggunakan *feature extractor Bag of Words* (BoW), SVM berhasil mencapai akurasi sebesar 80,6%, lebih unggul dibandingkan dengan *Naive Bayes*. Selanjutnya, Syahputra & Wibowo (2023) membandingkan SVM dengan *Random Forest* dalam kasus klasifikasi teks dengan menggunakan pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF), di mana SVM memperoleh akurasi sebesar 97%, jauh melebihi *Random Forest* yang hanya mencapai 92%.

Sementara itu di dalam konteks klasifikasi berita berbahasa Inggris dengan *feature extractor* BoW, Yuliani dkk. (2019) menunjukkan bahwa SVM memiliki akurasi 92.31%, lebih tinggi dari *Naive Bayes* (88,10%), *Decision Tree* (90,61%), *Logistic Regression* (90,63%), dan *Stochastic Gradient Descent* (88,40%). Hal ini sejalan dengan penelitian oleh Nurhasanah dkk. (2022) di mana SVM memiliki akurasi sebesar 96%, unggul sekitar 7% dibandingkan dengan *Naive Bayes* meskipun kali ini *feature extractor* yang digunakan adalah TF-IDF.

Penelitian lainnya oleh Triyono dkk. (2023) menemukan bahwa SVM juga unggul dari berbagai macam metode *machine learning* tradisional dalam hal klasifikasi berita berbahasa

Indonesia. Memiliki akurasi 83,55%, SVM unggul atas *Logistic Regression*, *Gradient Boosting Classifier*, dan *Random Forest* yang masing-masing memiliki akurasi sebesar 82,61%. SVM juga mengalahkan metode *Decision Tree Classifier* yang hanya memiliki akurasi sebesar 75,33%.

Lebih lanjut, Indra dkk. (2024) dalam penelitiannya mencoba membandingkan performa berbagai metode *machine learning* tradisional yang dikombinasikan dengan pembobotan oleh TF-IDF dalam konteks klasifikasi berita selama Pemilihan Umum (Pemilu) di Indonesia. SVM yang memiliki akurasi sebesar 86,36% unggul terhadap dua metode lainnya, yaitu TF-IDF + *Random Forest* dan TF-IDF + K-NN yang masing-masing memiliki akurasi sebesar 81,82% dan 79,54%.

Selanjutnya, meskipun pendekatan *machine learning* tradisional telah menunjukkan akurasi yang cukup tinggi dalam mengklasifikasi berita berbahasa Indonesia, perkembangan terbaru di bidang *deep learning* juga menawarkan potensi signifikan untuk mengatasi masalah ini. Salah satu metode *deep learning* yang kerap digunakan karena performanya yang sangat baik adalah *Indonesian Bidirectional Encoder Representations from Transformers* (IndoBERT). Dibandingkan dengan berbagai metode *machine learning* tradisional dan bahkan metode *deep learning* lainnya, IndoBERT (*base*) unggul secara performa dengan mencatatkan akurasi sebesar 88%, unggul cukup jauh dari metode lain seperti *One-Hot Encoding* + *Convolutional Neural Network-Long Short-Term Memory* (CNN-LSTM) dengan akurasi 73%, TF-IDF + *Naive Bayes* dengan akurasi 75%, TF-IDF + *Logistic Regression* dengan akurasi 65%, TF-IDF + *Gradient Boosting Classifier* dengan akurasi 55%, TF-IDF + *Decision Tree* dengan akurasi 61%, dan terakhir TF-IDF + *Random Forest* dengan akurasi 60% (Ridho & Yulianti, 2024).

Sinapoy dkk. (2023) dalam penelitiannya juga mengungkapkan bahwa IndoBERT + *Dense* yang memiliki akurasi sebesar 92,07% unggul secara performa dibandingkan dengan Word2Vec + LSTM + *Dense* yang akurasinya hanya sebesar 87,54% dalam hal mengklasifikasi berita berbahasa Indonesia di media sosial Twitter (saat ini X). IndoBERT juga unggul dibandingkan dengan *transformer-based models* sejenis dalam konteks klasifikasi berita (hoaks dan asli) COVID-19 dalam bahasa Indonesia dengan mencatatkan akurasi sebesar 97,67%, lebih tinggi jika dibandingkan dengan *Bidirectional Encoder Representations from Transformers* (BERT) maupun *Multilingual Bidirectional Encoder*

Representations from Transformers (mBERT) yang masing-masing memiliki akurasi sebesar 96,38% dan 97,16% (Suadaa dkk., 2021).

Berdasarkan uraian sebelumnya, di mana SVM dan IndoBERT telah masing-masing menunjukkan keunggulan dalam mengolah data teks dan mengklasifikasi berita dengan akurasi tinggi, integrasi keduanya menjadi pendekatan yang menarik untuk diteliti lebih lanjut. Sementara IndoBERT unggul dalam memahami konteks dan nuansa bahasa Indonesia, SVM efektif dalam klasifikasi data dengan presisi tinggi. Pendekatan ini menerapkan arsitektur linear di mana IndoBERT berfungsi sebagai *feature extractor* untuk menghasilkan representasi vektor dari teks berbahasa Indonesia, kemudian SVM berperan sebagai *classifier* untuk melakukan klasifikasi berdasarkan fitur yang telah diekstraksi guna menghasilkan sebuah model klasifikasi berita (hoaks dan asli) di media digital dengan akurasi tinggi. Untuk mencapai tujuan tersebut, penulis mengusulkan metode IndoBERT dan SVM dalam klasifikasi berita berbahasa Indonesia pada media digital.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan sebelumnya, rumusan masalah dalam penelitian ini adalah bagaimana mengimplementasikan IndoBERT sebagai *feature extractor*, kemudian memanfaatkan SVM sebagai *classifier* untuk melakukan klasifikasi berdasarkan fitur yang telah diekstraksi tersebut. Proses ini dilakukan secara linear dan berurutan, di mana IndoBERT terlebih dahulu mengekstraksi fitur dari teks *input*, lalu hasil ekstraksi fitur tersebut menjadi *input* bagi SVM untuk proses klasifikasi, sehingga integrasi kedua metode menghasilkan model klasifikasi berita (hoaks dan asli) berbahasa Indonesia di media digital yang optimal, dalam hal ini optimal mengacu pada pencapaian performa terbaik dalam hal *accuracy*, *precision*, *recall*, dan *F1-score* yang tinggi serta konsisten.

1.3 Tujuan dan Manfaat

Tujuan dari penelitian ini adalah mengimplementasikan IndoBERT sebagai *feature extractor* yang terintegrasi dengan SVM sebagai *classifier* dalam arsitektur linear untuk menghasilkan model klasifikasi berita (hoaks dan asli) berbahasa Indonesia di media digital yang optimal dengan performa terbaik dalam hal *accuracy*, *precision*, *recall*, dan *F1-score* yang tinggi serta konsisten.

Manfaat dari penelitian ini adalah tersedianya model klasifikasi berita berbahasa Indonesia di media digital dengan mengintegrasikan IndoBERT sebagai *feature extractor* dan SVM sebagai *classifier* dalam arsitektur linear yang mampu membedakan berita hoaks dan berita asli berbahasa Indonesia di media digital. Model tersebut diharapkan dapat memberikan solusi praktis dalam meningkatkan keakuratan penyaringan informasi yang diterima masyarakat, sehingga mampu mengurangi penyebaran disinformasi dan dampak negatifnya terhadap opini publik. Selain itu, model ini juga diharapkan dapat dikembangkan dan dioptimalkan untuk diterapkan pada skala yang lebih luas di masa mendatang, sehingga mampu mengakomodasi berbagai variasi bahasa dan gaya penulisan yang muncul seiring perkembangan media digital.

1.4 Ruang Lingkup

Penelitian skripsi ini memerlukan penetapan ruang lingkup agar kegiatan penelitian dapat berjalan secara terfokus dan menghasilkan *output* yang sesuai dengan tujuan yang diharapkan. Adapun ruang lingkup yang diterapkan dalam penelitian ini adalah sebagai berikut:

1. Penelitian ini menggunakan istilah "media digital" karena mencakup ekosistem informasi yang lebih komprehensif dibandingkan hanya "media sosial". Media digital meliputi seluruh *platform* berbasis teknologi digital untuk penyebaran informasi, termasuk *platform* media sosial dan situs digital lain yang biasa menjadi tempat penyebaran berita, sehingga memberikan cakupan yang lebih luas.
2. Berita yang dikumpulkan terdiri dari dua kategori, yaitu berita hoaks dan berita asli. Berita hoaks diperoleh dari situs TurnBackHoax.ID yang berasal dari berbagai media sosial seperti Instagram, X, Tiktok, dan Facebook, sedangkan berita asli berasal dari tiga media berita *online* yaitu CNN, SINDOnews, dan detik.
3. Pengumpulan berita dilakukan secara manual dan diformat menjadi *Comma Separated Values* (CSV).
4. Rentang waktu pengumpulan data adalah dari tanggal 3 Februari 2025 hingga tanggal 15 Februari 2025.
5. Data berhasil terkumpul sejumlah 800 berita dengan rincian 400 berita asli dan 400 berita hoaks.
6. Model IndoBERT yang digunakan sebagai adalah IndoBERT-base dengan aktivasi *sigmoid*, *loss function Binary Cross Entropy* (BCE), dan fungsi optimasi AdamW.

Adapun dari 12 *hidden layer* IndoBERT-base, hanya dilakukan pembaruan bobot pada 6 *layers* terakhir, sedangkan 6 *layers* awal di-*freeze* menggunakan bobot dari *pre-trained model*.

7. Fungsi optimasi AdamW menggunakan *Grid Search* untuk mencari kombinasi *hyperparameter weight decay*, *learning rate*, dan *batch size* terbaik.
8. Dalam proses klasifikasi dengan SVM, kernel yang digunakan adalah *Radial Basis Function* (RBF). Selain itu, untuk mencari *hyperparameter* C dan gamma terbaik, digunakan *Grid Search Cross Validation*.

1.5 Sistematika Penulisan

Struktur penulisan laporan ini dibagi ke dalam lima bab utama yang membahas topik klasifikasi berita berbahasa Indonesia pada media digital menggunakan metode IndoBERT dan *Support Vector Machine*. Pembagian bab ini dimaksudkan agar penulisan menjadi lebih terstruktur dan mudah dipahami. Berikut adalah ringkasan singkat dari masing-masing bab yang ada:

BAB I PENDAHULUAN

Bab ini membahas latar belakang, rumusan masalah, tujuan dan manfaat penelitian, ruang lingkup, serta sistematika penulisan yang digunakan dalam penyusunan skripsi mengenai klasifikasi berita berbahasa Indonesia pada media digital menggunakan metode IndoBERT dan *Support Vector Machine*.

BAB II TINJAUAN PUSTAKA

Bab ini memaparkan landasan teori dari dua metode utama penelitian: *Indonesian Bidirectional Encoder Representations from Transformers* (IndoBERT) dalam pemrosesan bahasa dan *Support Vector Machine* (SVM) sebagai metode klasifikasi, ditambah teori relevan lainnya untuk klasifikasi berita berbahasa Indonesia di media digital.

BAB III METODE PENELITIAN

Bab ini menjelaskan metodologi yang digunakan dalam penelitian ini, meliputi alur penelitian, proses pengumpulan data, pembagian *dataset*, *pre-processing* data, dan tokenisasi dengan *Indonesian Bidirectional Encoder Representations from Transformers* (IndoBERT). Selanjutnya, dijelaskan

proses pelatihan model dengan *fine tuning* IndoBERT-base yang kemudian diintegrasikan dengan *Support Vector Machine* (SVM) melalui *Grid Search Cross Validation* untuk menghasilkan model terbaik. Bab ini ditutup dengan perancangan aplikasi simulasi klasifikasi berita berbahasa Indonesia di media digital.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjelaskan analisis dan hasil penelitian, termasuk evaluasi performa model dan implementasi aplikasi klasifikasi berita berbahasa Indonesia di media digital.

BAB V PENUTUP

Bab ini merangkum hasil penelitian secara komprehensif dan memberikan saran praktis untuk penelitian selanjutnya, berdasarkan seluruh pembahasan yang telah dilakukan.