BAB 1
PENDAHULUAN

1.1 Latar Belakang

Di era digital saat ini, informasi dapat diakses dan disebarkan dengan cepat melalui berbagai platform, seperti media sosial dan situs berita daring. Meskipun kemudahan ini memberikan manfaat besar dalam berbagi informasi, muncul pula tantangan signifikan berupa penyebaran berita palsu atau hoax. hoax didefinisikan sebagai informasi tidak akurat atau menyesatkan yang dibuat dengan tujuan memengaruhi opini publik, memanipulasi pandangan, dan bahkan menciptakan kebingungan dalam masyarakat [1, 2]. Fenomena ini telah memberikan dampak negatif di Indonesia, seperti kebingungan publik, ketegangan politik, hingga potensi perpecahan sosial [3]. Sebagai contoh, selama Pemilu 2019, penyebaran hoax yang bercampur dengan isu politik, agama, dan ras menjadi ancaman serius bagi stabilitas demokrasi di Indonesia [4, 5].

Dampak hoax tidak hanya terbatas pada ranah sosial, tetapi juga merambah ke ranah keamanan siber. Dalam konflik Rusia-Ukraina, hoax digunakan sebagai alat dalam perang siber untuk menyerang infrastruktur vital dan menciptakan instabilitas politik, menunjukkan bahwa penyebaran informasi palsu dapat mengancam keamanan global [6]. Oleh karena itu, pengembangan sistem deteksi berita hoax otomatis menjadi semakin penting untuk mengatasi tantangan ini.

Salah satu solusi yang menjanjikan adalah penerapan teknik klasifikasi teks berbasis machine learning. Pendekatan ini memungkinkan analisis data teks dalam jumlah besar dengan cepat dan akurat. Namun, penelitian sebelumnya dalam bidang ini masih terbatas pada dataset kecil, berkisar antara 1.000 hingga 3.000 sampel [7, 8, 9, 10]. Selain itu, sebagian besar penelitian belum mengeksplorasi secara mendalam pengaruh variasi n-gram terhadap akurasi model deteksi berita hoax.

Penelitian sebelumnya umumnya menggunakan teknik feature extraction seperti TF-IDF atau CountVectorizer dengan konfigurasi default, tanpa eksplorasi mendalam terhadap n-gram atau kombinasi n-gram. Hal ini

menyisakan celah penelitian terkait bagaimana variasi n-gram, seperti unigram, bigram, trigram atau kombinasi ketiganya, dapat memengaruhi performa model dalam mendeteksi berita hoax.

Naive Bayes, salah satu algoritma klasifikasi yang populer, telah terbukti efektif dalam mendeteksi berita hoax karena kesederhanaan dan efisiensinya dalam mengolah data teks [2]. Penelitian oleh Sy. Yuliani [11] menunjukkan bahwa Naive Bayes mampu mencapai akurasi hingga 88%, menandakan potensinya untuk dikembangkan lebih lanjut. Penelitian lain oleh V. O. Yamin [12], menegaskan pentingnya sistem deteksi hoax otomatis dengan memanfaatkan Naive Bayes yang disempurnakan melalui teknik stemming, menghasilkan F1-Score sebesar 85%. Sistem ini terbukti efektif dalam membedakan berita hoax dari berita valid, dengan data yang diambil dari sumber terpercaya.

Penelitian N. E. Febrianty [13] dan penelitian I. F. Ferdiansyah & Wella [14] menunjukkan bahwa algoritma Naive Bayes memiliki performa tinggi dengan akurasi masing-masing sebesar 88% dan 89,48% dalam mendeteksi berita hoaks, penelitian tersebut masih terbatas pada dataset kecil sebesar 1000 data. Selain itu, belum ada eksplorasi mendalam terkait pengaruh teknik ekstraksi fitur seperti n-gram atau kombinasi Count Vectorizer dan TF-IDF Vectorizer terhadap akurasi. Dengan ini bertujuan untuk melengkapi kekurangan tersebut dengan memanfaatkan dataset yang lebih besar dan menguji berbagai variasi teknik ekstraksi fitur, sehingga diharapkan dapat menghasilkan model deteksi berita hoax yang lebih optimal dan relevan, dengan akurasi yang mampu melampaui hasil penelitian sebelumnya.

Dalam penelitian ini, digunakan dua teknik ekstraksi fitur, yaitu Count Vectorizer dan TF-IDF Vectorizer, untuk mengevaluasi kinerja algoritma Naive Bayes. Count Vectorizer merepresentasikan teks berdasarkan frekuensi kemunculan kata, sedangkan TF-IDF Vectorizer menormalkan frekuensi tersebut berdasarkan keberadaan kata dalam dokumen lain, sehingga lebih menonjolkan kata-kata yang relevan. Kombinasi teknik ekstraksi fitur ini dengan pendekatan n-gram (unigram, bigram, dan trigram) diharapkan dapat meningkatkan akurasi deteksi berita hoax, khususnya dalam konteks bahasa Indonesia. Dengan pendekatan ini, penelitian ini bertujuan memberikan kontribusi nyata dalam pengembangan sistem deteksi hoax yang lebih akurat, efisien, dan relevan dengan tantangan di era digital.

1.2 Rumusan Masalah

1. Bagaimana kinerja algoritma Naive Bayes dalam mendeteksi berita hoax menggunakan teknik ekstraksi fitur Count Vectorizer dan TF-IDF Vectorizer?
2. Bagaimana pengaruh variasi parameter n-gram terhadap akurasi model Naive Bayes dalam deteksi berita hoax?

1.3 Tujuan Penelitian

1. Mengevaluasi kinerja algoritma Naive Bayes dalam mendeteksi berita hoax menggunakan dua teknik ekstraksi fitur, yaitu Count Vectorizer dan TF-IDF Vectorizer.
2. Menganalisis pengaruh variasi parameter n-gram (unigram, bigram, dan trigram) terhadap akurasi model Naive Bayes dalam mendeteksi berita hoax.

1.4 Urgensi Penelitian

Penyebaran berita hoax di Indonesia semakin meningkat, terutama melalui media sosial, yang menyebabkan dampak negatif signifikan terhadap stabilitas sosial, politik, dan kepercayaan publik. hoax sering kali memicu konflik dan perpecahan, seperti yang terlihat pada kasus Pemilu 2019, sehingga deteksi otomatis berita hoax menjadi kebutuhan mendesak untuk membantu menangani masalah ini. Namun, penelitian sebelumnya umumnya masih terbatas pada dataset kecil dan belum sepenuhnya mengeksplorasi teknik ekstraksi fitur atau variasi parameter n-gram yang dapat meningkatkan kinerja model deteksi. Oleh karena itu, penelitian ini penting untuk memberikan kontribusi nyata dalam mengembangkan sistem deteksi berita hoax yang lebih akurat dan efisien, khususnya dalam konteks bahasa Indonesia. Selain menawarkan solusi praktis yang dapat diimplementasikan pada berbagai platform digital, penelitian ini juga memberikan kontribusi akademik untuk pengembangan metode pemrosesan bahasa alami (NLP) dalam deteksi hoax.

1.5 Luaran Penelitian

Hasil dari penelitian ini diharapkan dapat menjadi referensi yang bermanfaat bagi penelitian-penelitian selanjutnya dalam bidang deteksi berita hoax. Dengan menyajikan analisis performa berbagai metode ekstraksi fitur, seperti Count Vectorizer, TF-IDF, dengan variasi N-gram, penelitian ini dapat menjadi landasan untuk mengembangkan model yang lebih akurat dan efisien. Selain itu, luaran penelitian ini juga dapat digunakan sebagai acuan dalam pengembangan sistem deteksi hoax yang lebih canggih di masa mendatang, termasuk penerapannya dalam konteks keamanan siber.

1.6 Manfaat Penelitian

Penelitian ini memiliki manfaat yang signifikan baik secara akademik, praktis, maupun sosial. Secara akademik, penelitian ini memberikan kontribusi dalam pengembangan metode machine learning, khususnya algoritma Naive Bayes, dengan mengevaluasi efektivitas teknik ekstraksi fitur seperti Count Vectorizer dan TF-IDF Vectorizer dalam konteks bahasa Indonesia, serta memperkaya literatur terkait pemrosesan bahasa alami (NLP) pada deteksi hoax. Secara praktis, hasil penelitian ini dapat digunakan untuk mengembangkan sistem deteksi otomatis berita hoax yang lebih akurat dan efisien, yang dapat diterapkan pada berbagai platform digital, seperti media sosial atau portal berita, sehingga membantu instansi pemerintah, organisasi media, dan masyarakat dalam menangkal penyebaran informasi palsu. Secara sosial, sistem ini diharapkan mampu mengurangi dampak negatif penyebaran hoax, seperti kebingungan di masyarakat, ketegangan sosial, dan konflik politik, sekaligus mendukung literasi digital masyarakat Indonesia dengan menyediakan akses terhadap informasi yang lebih valid dan terpercaya.