BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi yang pesat telah mengubah gaya hidup manusia, terutama dalam cara memperoleh informasi. Portal berita *online* menjadi sumber informasi yang sangat diminati oleh masyarakat, karena banyak penulis yang menyebarkan artikel mereka di berbagai situs web. Jumlah pembaca berita *online* pun terus meningkat setiap harinya. Sebuah survei pada tahun 2017 menunjukkan bahwa jumlah pembaca media *online* telah mencapai 6 juta orang. Portal berita *online* kini tidak hanya berfungsi sebagai sarana untuk menyebarkan informasi faktual, tetapi juga sebagai bisnis yang menjanjikan. Semakin banyak pengunjung yang mengakses suatu situs berita, semakin tinggi *traffic* situs tersebut, dan semakin besar keuntungan yang dapat diperoleh. Salah satu strategi yang sering digunakan untuk meningkatkan jumlah pengunjung (*page views*) dan *traffic* situs berita adalah dengan menggunakan judul artikel yang menjebak (*clickbait*). Tujuan dari penggunaan judul yang menjebak ini adalah untuk membuat pembaca merasa penasaran sehingga mereka akan membuka tautan berita tersebut[1]**.**

Clickbait adalah judul konten yang dirancang untuk menarik perhatian dan mendorong pengunjung mengeklik tautan ke halaman *web* tertentu. *Clickbait* juga dikenal sebagai tautan jebakan, di mana judul kontennya dibuat sedemikian rupa untuk menarik perhatian pembaca. Namun, isi kontennya seringkali biasa saja dan kadang-kadang tidak relevan dengan judulnya. Praktik *clickbait* banyak

ditemukan dalam berita *online* serta konten di media sosial[2]. Judul berita yang mengandung unsur *clickbait* sering menggunakan kata kunci yang sedang tren, namun isi beritanya tidak memberikan informasi yang berarti dan hanya mengandung pesan kontroversial. *Clickbait* memanfaatkan aspek psikologis manusia, yaitu rasa keingintahuan. Rasa ingin tahu ini muncul ketika seseorang ingin mengetahui sesuatu yang baru. Celah ini mudah dimanfaatkan dengan menyajikan pesan-pesan dengan kalimat yang mempengaruhi pembaca dalam *headline clickbait* mengenai informasi baru atau kontroversial, sehingga memancing rasa ingin tahu pembaca dan mendorong mereka untuk mengklik judul tersebut[3].

Website ini diperlukan karena dapat meningkatkan akurasi dan efisiensi dalam mendeteksi judul berita *clickbait* di media *online,* yang selama ini masih menjadi tantangan. Dengan adanya komparasi antara algoritma SVM, LSTM, dan IndoBERT, kita dapat menentukan metode yang paling efektif untuk memisahkan judul *clickbait* dari judul *non-clickbait*. Penggunaan platform ini akan memudahkan masyarakat dalam mendapatkan informasi yang lebih akurat dan terpercaya, mengurangi penyebaran berita yang menyesatkan, serta meningkatkan kualitas konten di media *online*. Selain itu, aplikasi ini juga akan membantu para peneliti dan praktisi media dalam memahami karakteristik berita *clickbait* dan mengembangkan strategi untuk meminimalisir dampaknya. Dengan demikian, aplikasi ini tidak hanya bermanfaat bagi pembaca, tetapi juga mendukung keberlanjutan dan perkembangan industri media yang lebih sehat dan informatif di Indonesia.

Pemilihan algoritma SVM, LSTM, dan IndoBERT dilakukan berdasarkan keunggulan dan kesesuaian masing-masing dalam tugas deteksi clickbait. Menurut penelitian oleh Cortes dan Vapnik yang berjudul *Support-Vector Networks* mengakatakan bahwa, SVM merupakan metode yang efektif dalam klasifikasi data karena kemampuannya untuk menemukan hyperplane yang optimal dalam memisahkan kelas yang berbeda. SVM telah banyak digunakan dalam berbagai aplikasi, termasuk klasifikasi teks, dan terbukti mampu memberikan performa yang baik, terutama dalam kasus dengan data berdimensi tinggi dan jumlah data yang terbatas[4]. LSTM, seperti yang dijelaskan oleh Hochreiter dan Schmidhuber dalam penelitiannya yang berjudul *Long Short-term Memory* bahwa LSTM memiliki kemampuan untuk mengingat informasi dalam jangka panjang, menjadikannya sangat cocok untuk tugas-tugas yang melibatkan data sekuensial seperti teks, di mana urutan kata sangat penting[5]. Sedangkan IndoBERT adalah model berbasis Transformer yang dilatih khusus untuk bahasa Indonesia. Penelitian oleh Koto dan lainnya, menunjukkan bahwa IndoBERT mampu menghasilkan performa yang superior dalam tugas-tugas NLP untuk bahasa Indonesia dibandingkan dengan model-model generik lainnya. Ini karena IndoBERT telah dilatih dengan korpus besar berbahasa Indonesia, sehingga lebih memahami struktur dan karakteristik bahasa tersebut[6].

Dengan membandingkan ketiga algoritma ini, dapat dikembangkan menjadi sebuah *website* yang dapat mendeteksi berita *clickbait* dengan menggunakan hasil model terbaik. *Website* ini dapat meningkatkan pemahaman tentang penggunaan teknik-teknik tersebut dalam meningkatkan kualitas konten berita,

mempromosikan judul yang informatif, dan meminimalkan penyebaran informasi yang menyesatkan.

1.2 Tujuan dan Manfaat

1.2.1 Tujuan

Tujuan dari skripsi ini adalah melakukan komparasi antara algoritma SVM, LSTM, dan IndoBERT pada sistem deteksi judul berita *clickbait* dan *non-clickbait* di media *online*.

1.2.2 Manfaat

Manfaat yang dapat diperoleh dari penelitian ini antara lain:

1. Penelitian ini memberikan pemahaman yang lebih dalam tentang performa dan karakteristik dari algoritma SVM, LSTM, dan IndoBERT dalam konteks deteksi clickbait. Ini bisa membantu peneliti dan praktisi memahami kelebihan dan kekurangan masing-masing algoritma.
2. Penelitian ini membantu meningkatkan akurasi dalam mengklasifikasi judul berita sebagai *clickbait* atau *non clickbait* menggunakan berbagai pendekatan algoritma yang telah dibandingkan.
3. Penelitian ini dapat membantu masyarakat dalam mengidentifikasi dan menyimpulkan situs-situs yang lebih banyak menggunakan clickbait melalui *website* monitoring, sehingga mereka dapat lebih selektif dalam memilih sumber berita yang dapat terpercaya.

1.3 Tinjauan Pustaka

Peneliti mengumpulkan data dari studi sebelumnya sebagai bahan perbandingan untuk mengevaluasi kelebihan dan kelemahan dari algoritma SVM, LSTM, dan IndoBERT dalam sistem deteksi judul berita *clickbait* dan *non-clickbait* di media *online.* Selain itu, referensi dari jurnal-jurnal terkait digunakan untuk mendalami teori-teori yang relevan dengan penerapan ketiga algoritma tersebut dalam deteksi *clickbait*.

Penelitian Selanjutnya, berfokus pada membandingkam LSTM dengan SVM dan *Naïve Bayes* pada klasifikasi *hoax*. Hasilnya, penelitian ini menunjukkan bahwa model *Long Short-Term Memory* (LSTM) yang digunakan untuk mengklasifikasikan empat kategori *hoax*, yaitu *fabricated content*, *manipulated content, misleading content, dan false context*, tidak memberikan hasil yang optimal. Meskipun akurasi pelatihan mencapai 0.49 pada *epoch* ke-480, akurasi ini menurun dan tidak signifikan saat diuji dengan data validasi, yang hanya mencapai 0.37. Hal ini menunjukkan bahwa penggunaan LSTM dalam klasifikasi *hoax* belum memberikan performa yang memadai. Sebagai perbandingan, model *Support Vector Machine* (SVM) dengan *kernel 'linear'* dan *Multinomial Naïve Bayes Classifier* digunakan. Hasil pengujian menunjukkan bahwa SVM memberikan kinerja yang lebih baik dengan nilai *precision, recall, dan f1-score* sebesar 0.74, dibandingkan dengan *Multinomial Naïve Bayes Classifier* yang masing-masing hanya mencapai nilai sekitar 0.62-0.64. Dengan demikian, SVM lebih unggul dalam mengklasifikasikan data *hoax* dibandingkan dengan LSTM dan *Naïve Bayes Classifier* dalam penelitian ini[7].

Penelitian selanjutnya, Peneliti membandingkan antara algoritma LSTM dan *IndoBERT* dalam mengidentifikasi berita *hoax* di *twitter*. Hasil penelitiannya adalah nilai rata-rata yang diperoleh dari eksperimen menggunakan 10-*fold cross-validation*. Model *IndoBERT* menunjukkan kinerja yang baik dengan nilai akurasi rata-rata sebesar 92,07%, sedangkan model LSTM memberikan nilai akurasi rata-rata sebesar 87,54%. Model *IndoBERT* dapat menunjukkan kinerja yang baik dalam tugas deteksi *hoax* dan terbukti lebih unggul dibandingkan dengan model LSTM yang memberikan hasil akurasi rata-rata terbaik dalam penelitian ini[8].

Penelitian selanjutnya berjudul Analisis sentiment opini publik menggunakan SVM pada media sosial *twitter*. Penelitian ini dapat disimpulkan bahwa metode *Support Vector Machine* (SVM) efektif diterapkan dalam analisis sentimen terhadap opini publik mengenai topik *childfree* di media sosial *Twitter*. Melalui perhitungan manual dengan 8 data sampel (5 data *training* dan 3 data *testing*), hasilnya menunjukkan bahwa dua dari tiga data *testing* diklasifikasikan sebagai sentimen *positif*, sementara satu data *testing* diklasifikasikan sebagai sentimen *negatif*. Penelitian ini juga berhasil membangun sistem analisis sentimen berbasis *web* menggunakan bahasa pemrograman *PHP* dan *database MySQL* untuk mengklasifikasikan 461 *tweet* terkait topik *childfree*. Hasilnya, mayoritas *tweet* (265) mengandung sentimen *positif*, diikuti oleh 129 *tweet* dengan sentimen *negatif*, dan 67 *tweet* dengan sentimen netral. Tingkat akurasi sistem analisis SVM ini mencapai 69,69%, dengan *recall* sebesar 45,60%, *precision* 51,56%, dan *F1-Score* 46%. Hasil ini mengindikasikan bahwa opini masyarakat mengenai *childfree* di *Twitter* cenderung positif, menunjukkan bahwa topik ini mulai

diterima di tengah masyarakat. Kinerja analisis SVM menunjukkan performa yang cukup baik dalam melakukan klasifikasi sentimen, dan penelitian ini berkontribusi dalam memperluas pengetahuan tentang penggunaan SVM untuk mengklasifikasikan opini publik serta memantau perkembangan topik *childfree* di media sosial *Twitter* di Indonesia[9].

Penelitian selanjutnya, melakukan Klasifikasi Opini Publik Terhadap Bakal Calon Presiden Indonesia Tahun 2024 di Platform Twitter, menunjukan bahwa hasil klasifikasi dengan menggunakan model LSTM menghasilkan akurasi 76% dan Parameter yang diuji antara lain *batch size, dropout*, dan *learning rate*. Model LSTM diterapkan pada *website* yang menyajikan *dashboard* dengan fitur peta warna yang menunjukkan tingkat sentimen positif untuk setiap calon di setiap provinsi, jumlah klasifikasi sentimen, serta opsi filter berdasarkan provinsi dan waktu[10].

Penelitian selanjutnya melakukan perbandingan antara SVM dan IndoBERT pada untuk Analisis Sentimen Pembangunan Sirkuit Balap di Indonesia. Berdasarkan hasil dan diskusi dari eksperimen yang dilakukan dengan 2624 data terkait pembangunan sirkuit balap di Indonesia, ditemukan bahwa rasio antara sentimen positif dan negatif adalah 61,7%: 38,3%. Dataset tersebut digunakan untuk membangun model *IndoBERTweet* dan *Support Vector Machine* (SVM) dengan *K-Fold Cross Validation*. Model terbaik dari segi akurasi dihasilkan pada iterasi keenam, dengan nilai akurasi *IndoBERTweet* sebesar 94% dan SVM sebesar 93%. Pada setiap iterasi k, model *IndoBERTweet* menunjukkan kinerja yang lebih baik dibandingkan SVM dalam hal akurasi, *presision, recall*,

dan f1-score. Perbedaan nilai evaluasi secara keseluruhan pada seluruh dataset menunjukkan bahwa *IndoBERTweet* memiliki nilai yang lebih tinggi pada semua atribut dibandingkan SVM, dengan selisih akurasi sebesar 4%, *presision* 1%, *recall* 4%, dan *f1-score* 3%. Dengan demikian, dapat disimpulkan bahwa *IndoBERTweet* adalah model yang lebih baik daripada SVM, terutama dalam penelitian analisis sentimen. Untuk penelitian lebih lanjut, penggunaan lebih banyak data dengan lebih banyak label dapat memberikan bukti tambahan untuk kesimpulan ini, terutama dalam perbandingan model. Selain itu, penggunaan parameter lain pada setiap model dapat dilakukan untuk menemukan nilai yang lebih sesuai guna meningkatkan kinerja model[11].

Tabel 1.1 Gap Penelitian

No	Topik	Teknologi	hasil	Pembeda
1.	Perbandingan LSTM dengan SVM dan *Naïve Bayes* pada klasifikasi *hoax*	LSTM, SVM, *Naïve Bayes*	Dengan model pelatihan dua fase mencapai akurasi tertinggi sebesar 0,8247, atau meningkat sebesar 0,064 (6%) dibandingkan dengan akurasi	Studi kasus penelitian ini berbeda dengan penelitian sebelumnya dalam beberapa hal. Pada pelatihan LSTM, penelitian sebelumnya menggunakan batch size 64 dengan 500 epoch, sementara penelitian ini

			klasifikasi SVM menggunakan model *bag-of-words* sebesar 0,7607	menggunakan batch size 32 dengan 10 epoch. Untuk SVM, penelitian sebelumnya menggunakan kernel linear, sedangkan penelitian ini menggunakan kernel RBF. Selain itu, penelitian sebelumnya mengajukan model SVM dengan akurasi tertinggi, sementara penelitian ini mengajukan model IndoBERT.
2.	Perbandingan antara algoritma LSTM dan *IndoBERT* dalam mengidentifikasi berita hoax di twitter	LSTM dan *IndoBERT*	Model *IndoBERT* menunjukkan kinerja yang baik dengan nilai akurasi rata-rata sebesar	Studi kasus penelitian ini berbeda dengan penelitian sebelumnya. Pada LSTM, penelitian sebelumnya menggunakan Word2Vec untuk

			92,07%, sedangkan model LSTM memberikan nilai akurasi rata-rata sebesar 87,54%.	representasi kata sedangkan penelitian saat ini menggunakan CountVectorizer. Model yang diajukan adalah model IndoBERT karena dipenelitian sebelumnya/sejenis hasilnya baik
3.	Analisis sentiment opini publik menggunakan SVM di twitter	SVM	Dengan menggunakan data 461 tweet, tingkat akurasi sistem analisis SVM ini mencapai 69,69%, dengan recall sebesar 45,60%, precision 51,56%, dan F1-Score 46%.	Studi kasus penelitian ini berbeda dengan penelitian sebelumnya dan menggunakan perhitungan manual untuk mencari parameter terbaik. Pada penelitian saat ini melakukan komparasi sehingga model terbaik dengan akurasi tertinggi yang diambil yaitu IndoBERT

			Hasil ini mengindikasikan bahwa opini masyarakat mengenai childfree di Twitter cenderung positif, menunjukkan bahwa topik ini mulai diterima di tengah masyarakat.	
4.	Klasifikasi opini publik bakal calon presiden 2024 menggunakan model LSTM	LSTM	hasil klasifikasi dengan menggunakan model LSTM menghasilkan akurasi 76% dan Parameter yang diuji antara lain batch size,	Studi kasus penelitian ini berbeda dengan penelitian sebelumnya. penelitian sebelumnya menggunakan Word2Vec untuk representasi kata pada LSTM sedangkan penelitian saat ini

			dropout, dan learning rate.	menggunakan CountVectorizer. Pada penelitian saat ini melakukan komparasi sehingga model terbaik dengan akurasi tertinggi yang diambil yaitu IndoBERT
5.	Perbandingan antara SVM dan IndoBERT pada untuk Analisis Sentimen Pembangunan Sirkuit Balap di Indonesia	SVM dan IndoBERT	Model terbaik dari segi akurasi dihasilkan pada iterasi keenam, dengan nilai akurasi IndoBERTweet sebesar 94% dan SVM sebesar 93%. Pada setiap iterasi k, model IndoBERTweet menunjukkan kinerja yang	Studi kasus penelitian ini berbeda dengan penelitian sebelumnya. Pada pelatihan IndoBERT, penelitian sebelumnya menggunakan batch size 12 dengan 3 epoch untuk mencapai akurasi tertinggi, sementara penelitian ini menggunakan batch size 32 dengan 10 epoch. Untuk SVM, penelitian sebelumnya

			lebih baik dibandingkan SVM dalam hal akurasi, presisi, recall, dan f1-score.	menggunakan kernel linear, sedangkan penelitian ini menggunakan kernel RBF. Sehingga IndoBERT diajukan menjadi model terbaik

1.4 Data Penelitian

1. Data Model

Data untuk model diperoleh dari repositori *GitHub* yang diunggah oleh Andika William yang berjudul *CLICK-ID* pada tahun 2020. Berdasarkan jurnalnya, data tersebut berisi judul berita yang dikumpulkan dari 12 penerbit berita lokal di Indonesia, dengan total 46.517 judul. Dari jumlah tersebut, 15.000 judul telah dilabeli sebagai *clickbait* atau *non-clickbait* oleh anotator. *File* data tersedia dalam format .csv dan .xlsx, dengan kolom tambahan untuk label dan skor label. Data yang telah dianotasi terbagi dalam *file* utama, *all_agree* (label yang disetujui semua penilai), dan *does_not_agree* (label dengan ketidaksetujuan). Hasil anotasi menunjukkan ada 8.710 judul *non-clickbait* dan 6.290 judul *clickbait*, dengan 5.297 label *non-clickbait* dan 3.316 label *clickbait* pada *file all_agree*. Skor *Fleiss'* K 0.42 menunjukkan tingkat kesepakatan yang sedang di antara penilai. Distribusi *clickbait* bervariasi antar penerbit, dengan detikNews memiliki tingkat *clickbait* terendah dan wowkeren tertinggi. Kategori judul dikelompokkan ke dalam 9

kelompok kategori, menunjukkan perbedaan signifikan dalam distribusi *clickbait* di berbagai kategori seperti Bisnis & Ekonomi dan Berita lainnya[12]. Selain itu, data tambahan diperoleh dari proyek yang berjudul *Detecting Clickbait Headline* di *Kaggle*. Dataset tersebut berformat CSV dan berisi judul-judul berita yang sudah terlabeli secara manual sebagai *clickbait* atau *non-clickbait.* Berikut sampel data yang dihasilkan:

Tabel 1.2 Sampel Dataset

No	Title	Label
1.	Jadwal Persib Bandung: Awas, Hindari Petaka Kartu Kuning!	Clickbait
2.	Ada Motor Nyangkut di Atas Bambu di Sleman, Kok Bisa?	Clickbait
3.	Viral! Driver Ojol di Bekasi Antar Pesanan Makanan Pakai Sepeda	Clickbait
4.	Adu Kekuatan Angkatan Laut Cina Vs AS, Mana Lebih Unggul?	Clickbait
5.	Air Kelapa Kaya Nutrisi untuk Ibu Hamil, Rasakan 9 Manfaatnya	Clickbait
6.	Masuk Radar Pilwalkot Medan, Menantu Jokowi Bertemu DPW NasDem Sumut	Non-Clickbait
7.	Terkait Mayat Bayi Mengenaskan di Tangerang, Seorang Pria Ditangkap Polisi	Non-Clickbait
8.	Peringati Tahun Baru Islam, Banyuwangi Kembali Gelar Festival Muharam	Non-Clickbait
9.	Lokasi Istana Presiden RI di Papua 15 Menit Lewat Jembatan Holtekamp	Non-Clickbait
10.	Pengacara Ajukan Penangguhan Penahanan Kivlan Zen	Non-Clickbait

2. Data Implementasi

Data yang digunakan dalam implementasi *website* ini diperoleh melalui proses *scraping* dari dua sumber utama, yaitu Detik.com dan Insertlive. Detik.com adalah sebuah *website* berita yang menyediakan berbagai kategori berita terbaru, termasuk berita nasional, internasional, politik, bisnis, dan lainnya. Sementara itu, Insertlive merupakan portal berita yang menyajikan informasi dan berita terkini dari berbagai kategori, termasuk *lifestyle*, teknologi, hiburan, dan lainnya. Data yang diambil meliputi judul berita, link berita, dan tanggal publikasi berita.

1.5 Alat Peneliltian

Penelitian ini memanfaatkan berbagai peralatan utama dan pendukung dalam perancangan sistem. Peralatan yang digunakan untuk merancang dan membangun sistem meliputi:

1. Perangkat Keras :

a. Laptop HP Processor 13th Gen Intel(R) Core(TM) i3-1315U 1.20 GHz
b. SSD 512GB
c. Ram 8 GB

2. Perangkat Lunak

Berikut perangkat Lunak yang digunakan dalam pembuatan penelitian ini

Tabel 1.3 Perangkat Lunak

No	Perangkat Lunak	Fungsi
1	Windows 10	Sistem Operasi
2	SQLite	Database
3	Visual Studio Code	*Text Editor*
4	Chrome	Melihat hasil dari kode
5	Flask	Tools Web
6.	Python	Bahasa Pemrograman
7	Google Colab	Pembuatan Model