

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi dan internet telah menjadi bagian yang tidak terpisahkan dari kehidupan manusia modern. Kemudahan akses terhadap informasi, komunikasi yang instan, serta berbagai layanan berbasis digital telah memberikan dampak signifikan dalam berbagai aspek kehidupan, mulai dari pendidikan, ekonomi, hingga pemerintahan. Namun, di balik kemajuan tersebut, terdapat pula tantangan yang semakin kompleks, khususnya dalam hal privasi dan keamanan data. Semakin tingginya ketergantungan masyarakat terhadap layanan digital menyebabkan peningkatan volume data yang disimpan dan diproses secara daring, sehingga memperbesar potensi risiko kebocoran informasi pribadi maupun institusional [1].

Kebocoran data merupakan salah satu permasalahan serius yang kerap terjadi di era digital saat ini dan berpotensi merugikan banyak pihak, baik individu, institusi, maupun negara. Data pribadi yang tersebar secara tidak sah dapat digunakan oleh pihak yang tidak bertanggung jawab untuk berbagai tujuan yang merugikan, seperti penipuan, pencurian identitas, hingga manipulasi informasi. Kelalaian dalam mengamankan data—terutama data yang bersifat sensitif seperti nomor identitas, informasi keuangan, riwayat kesehatan, atau data biometrik—dapat membuka celah besar bagi penyalahgunaan oleh pihak ketiga. Risiko tersebut tidak hanya menimbulkan kerugian finansial, tetapi juga dapat merusak reputasi individu maupun lembaga terkait. Oleh karena itu, penting bagi seluruh pihak, khususnya instansi pemerintah dan penyedia layanan digital, untuk menerapkan standar keamanan data yang tinggi guna melindungi hak privasi dan keamanan informasi para pengguna. [2].

Perlindungan terhadap informasi elektronik dan transaksi digital telah menjadi fokus penting dalam regulasi nasional seiring dengan meningkatnya ketergantungan masyarakat terhadap teknologi informasi. Salah satu payung hukum yang mengatur hal tersebut adalah Peraturan Pemerintah Republik Indonesia

Nomor 71 Tahun 2019 tentang Penyelenggaraan Sistem dan Transaksi Elektronik. Peraturan ini menegaskan bahwa Pemerintah memiliki tanggung jawab untuk melindungi kepentingan umum dari berbagai bentuk gangguan, terutama yang muncul akibat penyalahgunaan informasi elektronik dan transaksi elektronik. Penyalahgunaan tersebut dapat berupa manipulasi data, pencurian identitas, penyebaran informasi palsu, atau bentuk-bentuk lain yang mengancam keamanan dan ketertiban publik. Oleh karena itu, peraturan ini tidak hanya mengatur tentang tata kelola sistem elektronik dan standar penyelenggara, tetapi juga menekankan pentingnya menjaga integritas dan keamanan data demi menciptakan ruang digital yang sehat dan terpercaya, sesuai dengan ketentuan peraturan perundang-undangan yang berlaku [3].

Perkembangan teknologi informasi dan komunikasi telah membawa kemudahan dalam kehidupan manusia, namun di sisi lain juga menimbulkan risiko baru, terutama terkait dengan privasi dan keamanan data. Salah satu ancaman nyata dalam era digital adalah kebocoran data pribadi yang dapat terjadi akibat kelalaian pengelolaan maupun serangan siber. Kebocoran data bukan hanya mengancam individu sebagai pemilik data, tetapi juga dapat merugikan lembaga penyedia layanan, baik dari sisi reputasi maupun tanggung jawab hukum. Kasus kebocoran data yang terjadi di Indonesia dalam beberapa tahun terakhir semakin meningkat, bahkan pada tahun 2022, publik dihebohkan oleh kebocoran data dari berbagai instansi seperti Indihome, PLN, dan pendaftaran kartu SIM yang dijual di situs gelap (dark web) dengan harga mencapai 50.000 dolar Amerika atau sekitar Rp 743 juta. Hal ini menunjukkan bahwa sistem pengamanan data di Indonesia masih sangat rentan terhadap pelanggaran dan penyalahgunaan oleh pihak tidak bertanggung jawab.

Pemerintah Indonesia sebenarnya telah mengeluarkan sejumlah regulasi untuk melindungi data dan transaksi elektronik, salah satunya adalah Peraturan Pemerintah Nomor 71 Tahun 2019 tentang Penyelenggaraan Sistem dan Transaksi Elektronik. Dalam peraturan tersebut dijelaskan bahwa pemerintah memiliki tanggung jawab dalam melindungi kepentingan umum dari segala bentuk gangguan akibat penyalahgunaan informasi elektronik. Namun demikian, masih sering terjadi kebocoran data yang melibatkan lembaga-lembaga besar, baik swasta maupun

pemerintah, yang mengindikasikan bahwa pelaksanaan regulasi tersebut belum maksimal. Hal ini kemudian memunculkan berbagai reaksi dan opini dari masyarakat, khususnya di media sosial seperti Twitter, yang menjadi ruang terbuka bagi publik untuk menyuarakan kekhawatiran, kritik, dan harapan mereka terhadap isu kebocoran data. Data ulasan dari Twitter menjadi sumber informasi yang berharga dalam menganalisis persepsi dan emosi pengunjung terhadap suatu destinasi wisata.

Twitter sebagai salah satu platform media sosial paling populer memberikan kemudahan bagi pengguna untuk mengekspresikan pendapat secara real-time. Opini-opini ini, jika dikumpulkan dan dianalisis, dapat memberikan gambaran yang komprehensif tentang persepsi masyarakat terhadap suatu isu tertentu. Dalam konteks kebocoran data PDN (Pusat Data Nasional), reaksi masyarakat di Twitter mencerminkan tingkat kepercayaan publik terhadap keamanan data yang dikelola oleh pemerintah. Analisis terhadap opini-opini ini menjadi penting karena dapat digunakan sebagai masukan dalam merumuskan kebijakan yang lebih baik di masa mendatang.

Untuk memahami pola sentimen masyarakat, digunakan teknik analisis sentimen, yaitu proses mengklasifikasikan opini dalam bentuk teks menjadi kategori positif, negatif, atau netral. Salah satu metode yang umum digunakan dalam klasifikasi sentimen adalah algoritma Naive Bayes. Algoritma ini berbasis pada Teorema Bayes dan bekerja secara efektif pada data teks yang berukuran besar dengan kompleksitas komputasi yang rendah. Naive Bayes banyak digunakan dalam analisis media sosial karena kesederhanaan dan kecepatannya, serta mampu memberikan hasil klasifikasi yang cukup akurat bahkan dengan jumlah data pelatihan yang terbatas.

Penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap isu kebocoran data PDN dengan memanfaatkan data dari Twitter. Metode yang digunakan dalam penelitian ini adalah algoritma Naive Bayes dengan bantuan teknik pembobotan TF-IDF (Term Frequency-Inverse Document Frequency) untuk mengekstraksi fitur penting dari data teks. Ulasan atau tweet yang terkumpul akan dikategorikan berdasarkan sentimen yang terkandung di dalamnya. Hasil dari analisis ini diharapkan dapat memberikan gambaran umum tentang bagaimana

masyarakat memandang kebocoran data PDN serta sejauh mana kepercayaan publik terhadap sistem keamanan data yang dimiliki pemerintah. Temuan dari penelitian ini diharapkan dapat memberikan kontribusi bagi instansi terkait, khususnya dalam hal evaluasi kebijakan dan peningkatan keamanan siber nasional.

Analisis sentimen adalah proses memahami, mengekstrak, dan mengolah data teks secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam kalimat opini. Proses ini memiliki dua perspektif, yaitu pendapat yang negatif atau positif[1]. Teknik klasifikasi yang biasa dipakai dalam melakukan analisis sentimen antara lain *Naive bayes Classifier* (NBC), *K-Nearest Neighbor* (KNN) dan *Support Vector Machines* (SVM).

Dalam penelitian ini, metode Naive Bayes dan metode *K-Nearest Neighbor* digunakan sebagai dasar. Metode *Naive Bayes* dipilih karena dapat menghasilkan tingkat akurasi terbaik dengan jumlah data latih yang sedikit, sedangkan metode *K-Nearest Neighbor* dipilih karena sangat tahan terhadap data suara. Kinerja kedua metode akan dibandingkan untuk menentukan mana yang lebih baik dalam melakukan klasifikasi dokumen. Hasil menunjukkan bahwa metode Naive Bayes lebih baik dengan tingkat akurasi 70%, sedangkan metode *K-Nearest Neighbor* lebih buruk dengan tingkat akurasi 40%[2].

Oleh karena itu, penelitian ini melakukan analisis sentimen terhadap tweets yang berhubungan dengan Universitas Muhammadiyah Malang (UMM) untuk mengetahui opini masyarakat tentang kampus ini. Tweets yang berisi sentimen masyarakat tentang UMM diklasifikasikan sebagai bagian dari analisis. Penelitian ini menggunakan *Naive Bayes* dan *Support Vector Machine (SVM)* dengan pembobotan TF-IDF untuk klasifikasi. Hasil perbandingan kedua teknik menunjukkan bahwa Naive Bayes memiliki akurasi sebesar 73,65% dibandingkan SVM[3].

Pada tahun 2022 lalu, Indonesia dihebohkan dengan bocornya data mulai dari data Indihome, PLN, hingga data pendaftaran kartu identitas ponsel atau kartu SIM yang dijual di situs gelap (*dark web*) dengan harga 50.000 dolar atau setara dengan Rp. 743 juta [4]. Baru-baru ini, PDNS 2 telah disusupi serangan *ransomware* pada tanggal 20 Juni 2024 dengan membobol sistem dan mengunci data-data yang ada di dalamnya. Akibat serangan tersebut, sebagian besar data di

pusat data yang dihuni 282 institusi pemerintahan pusat dan daerah ini terkunci dan tak bisa dipulihkan kecuali pemerintah membayar tebusan untuk membuka kuncinya seharga 8 juta dolar atau sekitar Rp 131,8 miliar.

Insiden kebocoran data kementerian di Indonesia telah menimbulkan keresahan di kalangan masyarakat. Kebocoran ini tidak hanya menyoroti kelemahan sistem keamanan siber yang ada, tetapi juga mempengaruhi tingkat kepercayaan publik terhadap institusi pemerintah. Kebocoran data, khususnya data pemerintah, dapat menimbulkan berbagai implikasi serius baik secara sosial, ekonomi, maupun politik. Oleh karena itu, pemahaman terhadap reaksi masyarakat terkait insiden kebocoran data ini menjadi penting untuk dilakukan. Sentimen masyarakat terhadap insiden ini dapat bervariasi, mulai dari kekecewaan, kemarahan, hingga kekhawatiran terhadap keamanan data pribadi mereka [5]. Mengingat pentingnya isu ini, diperlukan penelitian yang mampu menangkap dan menganalisis sentimen publik secara *komprehensif*.

Twitter merupakan salah satu media sosial yang membantu manusia untuk saling berhubungan dan berbagi informasi satu sama lain. *Twitter* juga dikenal sebagai layanan *micro-blogging*, karena penggunanya diberikan fasilitas untuk menulis, membaca, serta mengirimkan informasi pada forum umum [6]. Dengan fitur-fitur interaktifnya, pengguna dapat dengan mudah menyampaikan pandangan mereka tentang berbagai topik, termasuk isu-isu sosial seperti bantuan sosial. Kemudahan penggunaan *Twitter* sebagai alat untuk berpartisipasi dalam diskusi publik telah membuatnya menjadi sarana yang efektif dalam menyuarakan opini terkait. Untuk melaksanakan tugas analisis sentimen ini, diperlukan metode yang dapat mendukung dalam proses klasifikasi. Salah satu metode yang dapat digunakan untuk melakukan analisis ini adalah *Naive Bayes*. Metode ini dipilih karena berdasarkan penelitian terdahulu yang membandingkan metode *Logistic Regression*, *Naive Bayes*, dan *Support Vector Machine* pada klasifikasi sentimen analisis data *Twitter* yang memberikan hasil terbaik dari metode lain [7]. Hasil dari menggunakan metode *Naive Bayes* pada penelitian tersebut adalah didapatkannya akurasi sebesar 73,7% dengan ekstraksi fitur TF-IDF dan 73,9% dengan menggunakan count vectorizer [7]. Pada penelitian lain yaitu memprediksi *re-tweet* dan *like* pada *Twitter* dengan membandingkan kinerja algoritma Random Forest,

Neural Network, Logistic Regression, Support Vector Machine, dan *Naive Bayes* menggunakan ekstraksi fitur TF-IDF dan Doc2Vec menghasilkan akurasi pada prediksi *re-tweet Naive Bayes* dengan kombinasi TF-IDF yaitu 75,9% dan merupakan hasil tertinggi daripada algoritma yang lain [8].

Metode *Naive Bayes* menjadi algoritma yang umum digunakan karena algoritma ini sederhana secara alur kerja namun tetap memiliki performa yang baik sehingga dapat membantu dalam tugas klasifikasi. Hal ini dibuktikan dengan beberapa penelitian yang telah disebutkan yaitu mampu memberikan akurasi yang optimal dan mampu mengungguli metode Machine Learning yang lain. Selain itu, algoritma perhitungan dalam *Naive Bayes* memiliki kecepatan komputasi yang tinggi dalam proses pelatihan dan klasifikasi yang membuat metode ini menarik untuk digunakan dalam tugas klasifikasi dan juga penelitian ini [9].

Penelitian ini dilakukan untuk menjawab permasalahan bagaimana sebaran sentimen masyarakat dalam mengomentari kasus kebocoran data khususnya data PDN yang baru-baru ini terjadi pada media sosial *Twitter* dan klasifikasi analisis sentimen menggunakan *Naive Bayes* dengan ekstraksi fitur TF-IDF. Penelitian ini diharapkan dapat memberikan gambaran yang komprehensif mengenai sentimen publik terhadap insiden kebocoran data yang terjadi.

1.2 Perumusan Masalah

1.2.1 Permasalahan

Di media sosial *Twitter*, banyak pengguna yang menyampaikan opini mereka terkait kasus kebocoran data Pusat Data Nasional (PDN). Namun, tingginya volume cuitan yang tersebar membuat sulit untuk mengetahui secara langsung bagaimana kecenderungan tanggapan masyarakat terhadap isu tersebut. Oleh karena itu, dilakukan analisis sentimen untuk mengidentifikasi apakah opini yang disampaikan bersifat positif atau negatif, dengan memanfaatkan algoritma *Naive Bayes* sebagai metode klasifikasinya.

1.2.2 Metode Usulan

Analisis sentimen dalam penelitian ini dilakukan menggunakan algoritma *Naive Bayes*. Data yang digunakan berupa cuitan berbahasa Indonesia yang

diperoleh dari Twitter terkait topik kebocoran data Pusat Data Nasional (PDN) dengan memanfaatkan teknik *web crawling*. Data tersebut kemudian melalui tahapan pelabelan sentimen, pembersihan data (*preprocessing*), hingga pemodelan klasifikasi. Hasil dari proses ini akan dianalisis menggunakan akurasi, presisi, dan recall sebagai metrik evaluasi untuk menarik kesimpulan mengenai kinerja model dalam mengklasifikasikan opini masyarakat.

1.2.3 Pertanyaan Penelitian

Berapa nilai akurasi, precision, recall, dan f1-score yang dihasilkan dari perbandingan antara metode *K-Fold Cross Validation* dan *Split Validation* terhadap algoritma *Naïve Bayes Classifier* dalam mengklasifikasikan sentimen positif dan negatif pada data opini masyarakat di Twitter mengenai kebocoran data Pusat Data Nasional (PDN)?

1.3 Tujuan dan Manfaat

1.3.2 Tujuan

Untuk Mengetahui nilai *akurasi*, *precision*, *recall* dan *f1-score* yang dihasilkan metode *Naive Bayes Classifier* untuk mengklasifikasikan sentimen positif dan negatif pada data opini masyarakat di Twitter mengenai kasus kebocoran data Pusat Data Nasional (PDN).

1.3.3 Manfaat

Adapun manfaat dari dilakukannya penelitian ini, yaitu :

1. Penelitian ini dapat menjadi rujukan, perbandingan, dan referensi tambahan dalam pengembangan ilmu pengetahuan di bidang analisis sentimen terhadap data teks, khususnya yang bersumber dari media sosial..
2. Untuk mengetahui kinerja metode *Naïve Bayes Classifier* dalam melakukan klasifikasi data teks ke dalam kategori sentimen positif dan negatif.
3. Hasil penelitian ini diharapkan dapat memberikan informasi yang dapat dijadikan bahan evaluasi bagi pemerintah atau instansi terkait, khususnya dalam upaya peningkatan keamanan data dan pengelolaan sistem informasi nasional, serta sebagai bahan pertimbangan dalam menangani opini publik terhadap kebocoran data Pusat Data Nasional (PDN).

1.4 Batasan Masalah

Adapun batasan masalah dalam penelitian ini adalah:

1. Data yang digunakan berupa cuitan berbahasa Indonesia yang diperoleh dari media sosial *Twitter*, dengan topik pembahasan seputar kebocoran data Pusat Data Nasional (PDN)
2. Data dikumpulkan melalui proses *crawling* menggunakan kata kunci tertentu seperti “data bocor”, “PDN bobol”, “PDN bocor”, dan “PDN hack”.
3. Jumlah data awal yang diperoleh dari proses *crawling* adalah sebanyak 1.262 tweet, kemudian ditingkatkan menjadi 1.440 tweet setelah dilakukan proses *oversampling* menggunakan teknik *SMOTE*
4. Data dikelompokkan berdasarkan dua sentimen: positif dan negatif.
5. Proses pelabelan dilakukan secara manual dan divalidasi oleh pakar di bidang bahasa Indonesia.
6. Evaluasi dilakukan dengan menggunakan dua teknik validasi, yaitu *K-Fold Cross Validation* (dengan nilai $K=5$) dan *Split Validation* (dengan rasio 80:20, 70:30, 60:40, dan 50:50) untuk mengukur kinerja model.

1.5 Sistematika Penulisan

Adapun sistematika tugas akhir yang digunakan terdiri dari beberapa bagian, yaitu :

BAB I PENDAHULUAN

Bab ini membahas mengenai latar belakang, perumusan masalah, permasalahan, metode usulan, pertanyaan penelitian, tujuan dan manfaat, tujuan penelitian, dan manfaat penelitian, batasan masalah, sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Pada Bab ini membahas mengenai dasar-dasar teori yang digunakan pada penelitian ini.

BAB III METODE PENELITIAN

Secara garis besar pada Bab ini dijelaskan mengenai perancangan alur analisis sentimen, tahapan penelitian, arsitektur sistem, data penelitian, preprocessing data, Smote,

dijelaskan tentang TF-IDF, tahapan klasifikasi data menggunakan metode *Naïve Bayes Classifier* (NBC), evaluasi, skenario uji coba.

BAB IV HASIL DAN PEMBAHASAN

Pada Bab ini dijelaskan mengenai implementasi sistem yang dikerjakan sesuai dengan rancangan yang dibuat pada Bab sebelumnya. Dimana terdapat skenario pengujian, hasil dari tahap crawling data pada situs platform *Twitter*, hasil dari tahap pelabelan, hasil dari tahap preprocessing data (*cleaning, tokenizing, konversi slangword, stopword, stemming*), hasil dari SMOTE, hasil dari tahap pembobotan dengan TF- IDF, hasil dari tahap *Naive Bayes Classifier*, hasil dari tahap evaluasi model menggunakan *confusion matrix*, terakhir terdapat hasil dari uji skenario.

BAB V PENUTUP

Pada bab ini dijelaskan tentang kesimpulan dari hasil penelitian yang meliputi visualisasi kata, hasil *Accuracy* serta hasil nilai *Precision, Recall* dan *F-1 score* yang didapatkan. Selain itu juga terdapat saran bagi penelitian berikutnya yang akan mengangkat topik yang sama yaitu analisis sentimen.