

BAB I

PENDAHULUAN

1.1 Latar Belakang

Evolusi pesat dunia digital telah membawa dampak besar dalam segala aspek kehidupan. Transformasi digital yang pesat ini telah menghasilkan akumulasi data yang sangat masif, seperti informasi pribadi, preferensi konsumen, data kesehatan, serta data finansial. Data-data ini tidak hanya menjadi aset berharga untuk inovasi dan efisiensi, tetapi juga menimbulkan tantangan besar terkait keamanan, terutama dalam hal perlindungan data nasional. Kebocoran data, penyalahgunaan informasi, dan serangan siber kini menjadi ancaman serius yang dapat berdampak negatif bagi individu, organisasi, dan bahkan kedaulatan negara. Isu ini semakin relevan dengan meningkatnya jumlah kasus kebocoran data, yang memicu perhatian publik dan mendorong pemerintah memperketat regulasi terkait perlindungan data [1].

Perlindungan data nasional merupakan upaya menyeluruh yang melibatkan aspek hukum, teknis, dan organisasional untuk menjaga keamanan dan kerahasiaan data. Upaya ini penting dalam menghadapi serangan siber dan kebocoran data yang kian marak. Namun, pelaksanaannya tidak tanpa tantangan, seperti kompleksitas regulasi yang sulit dipahami masyarakat serta kurangnya kesadaran publik tentang pentingnya melindungi data pribadi. Selain itu, sumber daya dan keahlian teknis yang terbatas, terutama di kalangan usaha kecil dan menengah, turut memperburuk keadaan. Lemahnya keamanan siber Indonesia tercermin dari peringkat negara dalam Cyber Security Index (NCIS), di mana Indonesia berada di posisi ke-83 dari 160 negara, jauh tertinggal dibandingkan negara tetangga seperti Malaysia dan Singapura [2]. Kondisi ini diperparah dengan data yang menunjukkan adanya 94 kasus kebocoran data di Indonesia antara tahun 2019 hingga 2023, yang sebagian besar melibatkan sektor privat. Kelemahan dalam keamanan siber ini menggarisbawahi urgensi untuk meningkatkan perlindungan data nasional di tengah pesatnya transformasi digital [3].

Kasus-kasus kebocoran data ini juga mendapat respons yang beragam dari masyarakat, terutama dari netizen di dunia maya. Platform media sosial seperti

Twitter telah menjadi ruang aktif bagi masyarakat untuk mengekspresikan opini dan emosi terkait isu-isu publik, termasuk perlindungan data. Melalui platform ini, respons masyarakat terhadap kebocoran data dan langkah-langkah perlindungan yang diambil oleh pemerintah menghasilkan berbagai sentimen yang menarik untuk dianalisis. Sentimen yang muncul di media sosial dapat mencerminkan kekhawatiran publik serta menjadi indikator sejauh mana kesadaran dan pemahaman masyarakat tentang pentingnya menjaga keamanan data pribadi mereka. Analisis terhadap respons dan sentimen ini dapat menjadi bahan evaluasi bagi pemerintah dan pihak terkait dalam merumuskan kebijakan perlindungan data yang lebih efektif dan dapat diterima oleh masyarakat luas.

Analisis sentimen menjadi penting untuk mengevaluasi efektivitas kebijakan dan strategi perlindungan data nasional. Teknik komputasi ini digunakan untuk mengidentifikasi dan mengkategorikan opini yang diekspresikan dalam teks, guna menentukan sikap penulis terhadap topik tertentu. Teknik tersebut dapat dimanfaatkan untuk memahami persepsi publik terhadap perlindungan data nasional, mengukur tingkat kepercayaan terhadap kebijakan pemerintah, serta mengidentifikasi potensi risiko dan kerentanan [4].

Meskipun sejumlah penelitian sebelumnya telah menggunakan analisis sentimen untuk menganalisis opini publik terkait isu sosial dan politik, penerapan analisis sentimen dalam konteks perlindungan data nasional masih terbatas. Pendekatan berbasis leksikon dan machine learning telah digunakan, tetapi akurasi dan efektivitasnya masih perlu ditingkatkan.

Oleh karena itu, penelitian ini mengusulkan penggunaan model Word2Vec untuk representasi kata serta TF-IDF (Term Frequency-Inverse Document Frequency) sebagai metode representasi berbasis frekuensi. TF-IDF mengukur seberapa penting suatu kata dalam dokumen, sedangkan Word2Vec mampu menangkap hubungan semantik antar kata dalam ruang vektor. Keduanya memiliki keunggulan tersendiri, namun efektivitasnya dalam konteks isu perlindungan data nasional masih perlu diuji lebih lanjut [5]. Sedangkan Logistic Regression merupakan algoritma klasifikasi statistik yang sederhana namun sangat efektif dalam memodelkan hubungan antara fitur dan label kelas biner, seperti sentimen positif dan negatif. Kelebihan utama Logistic Regression adalah kemampuannya

untuk memberikan interpretasi probabilistik, efisiensi komputasi, dan kinerja yang kompetitif dalam berbagai tugas klasifikasi teks.

Untuk klasifikasi sentimen, algoritma Logistic Regression dipilih karena kemampuannya dalam menangani data teks dan memberikan interpretasi probabilistik yang jelas [6]. Namun dalam praktiknya, distribusi data sentimen sering kali tidak seimbang (imbalance), di mana data sentimen positif atau negatif jumlahnya jauh lebih dominan. Hal ini dapat mengurangi akurasi model klasifikasi. Oleh karena itu, metode oversampling seperti SMOTE (Synthetic Minority Over-sampling Technique) digunakan untuk menyeimbangkan distribusi kelas dan diharapkan dapat meningkatkan kinerja model klasifikasi [7].

Hipotesis dalam penelitian ini menyatakan bahwa penggunaan metode representasi teks TF-IDF akan memberikan hasil klasifikasi sentimen yang lebih baik dibandingkan dengan Word2Vec ketika dikombinasikan dengan algoritma Logistic Regression. TF-IDF diperkirakan lebih unggul karena kesederhanaannya dalam merepresentasikan frekuensi kata yang relevan secara langsung dengan konteks sentimen, terutama pada data yang tidak terlalu besar. Selain itu, penerapan metode SMOTE diprediksi dapat meningkatkan kinerja model klasifikasi sentimen yang menggunakan kombinasi representasi teks dan Logistic Regression, khususnya dalam hal akurasi, precision, recall, dan F1-score. Dengan demikian, pemilihan metode representasi yang sesuai dan penanganan ketidakseimbangan data diupayakan mampu menghasilkan model analisis sentimen yang akurat dalam mengidentifikasi persepsi publik terhadap isu perlindungan data nasional.

Dengan demikian, penelitian ini tidak hanya berkontribusi pada peningkatan pemahaman tentang sentimen publik terhadap perlindungan data, tetapi juga memberikan dasar bagi pembuatan kebijakan yang lebih responsif dan efektif dalam menghadapi tantangan perlindungan data di era digital yang terus berkembang.

1.2 Perumusan Masalah

1.2.1 Permasalahan

Permasalahan utama dalam penelitian ini adalah bagaimana memahami secara mendalam persepsi dan sentimen publik terhadap perlindungan data

nasional. Perlindungan data telah menjadi prioritas penting di tengah meningkatnya kasus kebocoran data, namun respons publik terhadap kebijakan pemerintah masih bervariasi dan tidak selalu terdokumentasi dengan baik. Maka dari itu, untuk memahami persepsi publik secara menyeluruh, diperlukan pendekatan yang lebih fleksibel dan akurat guna mendapatkan hasil yang lebih mendalam dan representatif.

1.2.2 Metode Usulan

Pendekatan yang diusulkan dalam penelitian ini meliputi penggunaan Word2Vec untuk merepresentasikan kata-kata dalam bentuk vektor numerik yang bermakna secara semantik serta TF-IDF sebagai metode representasi berbasis frekuensi, dan algoritma Logistic Regression (LR) sebagai metode klasifikasi sentimen. Word2Vec memungkinkan representasi kata berdasarkan konteks, sedangkan Logistic Regression dipilih karena kemampuannya yang baik dalam menangani masalah klasifikasi biner seperti sentimen positif dan negatif, dengan hasil yang interpretatif dan efisien dalam skenario teks berdimensi tinggi.

1.2.3 Pertanyaan Penelitian

1. Bagaimana pengaruh penggunaan TF-IDF dengan Word2vec terhadap kinerja algoritma Logistic Regression dalam mengklasifikasikan sentimen publik terhadap isu perlindungan data nasional?
2. Bagaimana pengaruh penggunaan metode SMOTE terhadap akurasi klasifikasi sentimen menggunakan kombinasi Word2Vec dan Logistic Regression?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan Penelitian

Adapun Tujuan yang didapat dari hasil penelitian ini ialah :

1. Menganalisis pengaruh penggunaan metode TF-IDF dan Word2Vec terhadap kinerja algoritma Logistic Regression dalam klasifikasi sentimen publik terhadap isu perlindungan data nasional.
2. Mengukur pengaruh penerapan metode SMOTE terhadap peningkatan akurasi klasifikasi sentimen yang dilakukan menggunakan representasi Word2Vec dan algoritma Logistic Regression.

1.3.2 Manfaat Penelitian

Hasil penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Memberikan pemahaman komparatif antara metode representasi teks (TF-IDF dan Word2Vec) dalam meningkatkan kinerja algoritma Logistic Regression untuk analisis sentimen.
2. Memberikan wawasan empiris mengenai pengaruh metode SMOTE dalam menangani ketidakseimbangan data pada analisis sentimen berbasis teks.
3. Menghasilkan model analisis sentimen yang optimal, yang dapat digunakan untuk memahami opini masyarakat terhadap isu perlindungan data nasional secara lebih akurat.

1.4 Batasan Masalah

Untuk memperjelas ruang lingkup penelitian, perlu ditetapkan beberapa batasan masalah sebagai berikut:

1. Penelitian ini fokus pada analisis sentimen terkait isu perlindungan data nasional di Indonesia, tidak mencakup isu di luar negeri.
2. Sumber data terbatas pada platform media sosial Twitter, yang umum digunakan untuk menyuarakan opini publik.
3. Metode yang digunakan dalam penelitian ini adalah Word2Vec, TF-IDF untuk representasi kata dan Logistic Regression untuk klasifikasi sentimen.

1.5 Sistematika Penulisan

Sistematika penulisan penelitian ini disusun untuk memberikan gambaran struktur dan alur penelitian, mulai dari latar belakang, kajian teoritis, hingga hasil dan kesimpulan. Setiap bab dijelaskan sebagai berikut:

BAB I Pendahuluan

Bab ini menjelaskan latar belakang yang mendasari penelitian, diikuti dengan perumusan masalah yang ingin diselesaikan. Selain itu, tujuan dan manfaat penelitian ini diuraikan untuk menunjukkan kontribusi penelitian ini, baik dari segi akademis maupun praktis. Batasan masalah disertakan untuk memperjelas ruang lingkup penelitian, dan diakhiri dengan metodologi penelitian serta sistematika

penulisan sebagai panduan pembaca untuk memahami alur penyusunan penelitian ini.

BAB II Tinjauan Pustaka

Tinjauan pustaka memuat landasan teoritis, konsep-konsep penting, serta hasil penelitian terdahulu yang mendukung dan berkaitan dengan topik yang dikaji. Bagian ini berfungsi sebagai acuan untuk memperkuat dasar metodologis dan arah penelitian yang dilakukan.

BAB III Metodologi Penelitian

Bab ini menjelaskan tahapan metodologi yang digunakan dalam penelitian, mulai dari pengumpulan data di media sosial, preprocessing, hingga penerapan Word2Vec dan TF-IDF. Proses klasifikasi sentimen dilakukan menggunakan algoritma Logistic Regression. Evaluasi model dilakukan dengan menggunakan metrik seperti precision, recall, dan F1-score.

BAB IV Hasil dan Pembahasan

Hasil Dan Pembahasan Hasil dan Pembahasan menjelaskan temuan yang diperoleh dari proses penelitian yang telah dilakukan. Hasil tersebut dianalisis dan dibahas dengan mengaitkannya pada teori atau penelitian sebelumnya, untuk melihat sejauh mana temuan penelitian mendukung atau berbeda dari referensi yang ada.

BAB V Penutup

Bab ini menyajikan kesimpulan yang diperoleh dari hasil penelitian dan analisis yang telah dilakukan. Selain itu, disampaikan pula beberapa saran sebagai masukan untuk penelitian selanjutnya maupun pemanfaatan hasil oleh pihak yang berkepentingan.