BAB II

LANDASAN TEORI

2.1 Tinjauan Pustaka

Untuk tujuan penelitian, rujukan atau tinjauan pustaka harus diberikan sebagai acuan dasar, Di antaranya adalah penelitian yang dilakukan oleh Joshua Muliawan dan Erick Dazki, yang melihat hubungan antara pemindahan ibu kota Indonesia dan penggunaan tiga algoritma: Naive Bayes, Knn, dan Random Forest. Penelitian ini bertujuan untuk menganalisis pendapat publik tentang pemindahan ibu kota Indonesia. Data diambil dari komentar tweet yang dikumpulkan antara Juni dan September 2023. Sebelum algoritma pelabelan dan klasifikasi diterapkan pada data, penelitian ini mempersiapkan data dengan teknik pre-processing. Studi ini menyelidiki nilai keakuratan tiga algoritma pengklasifikasian: Naive Bayes Classifier, K-Nearest Neighbor, dan Random Forest. Hasil klasifikasi data menunjukkan bahwa sentimen positif sebesar 36.8%, sentimen netral sebesar 25%, dan sentimen negatif terkait pemindahan ibu kota sebesar 38.1%. Pengujian akurasi kemudian dilakukan terhadap metode Algoritma Naïve Bayes Classifier, yang menunjukkan nilai keakuratan sebesar 65.26%, Algoritma K-Nearest Neighbor sebesar 58.25%, dan Algoritma Random Forest sebesar 45.05%. Hasil ini menunjukkan bahwa metode Algoritma Naïve Bay Selain itu, penelitian ini menemukan jumlah kata kunci yang sering muncul dalam setiap sentimen, yang dapat berguna untuk melacak opini publik di media sosial.[12]

Dalam studi yang dilakukan oleh Andrew Kurniawan dan Sejati Waluyo berjudul "Penerapan Algoritma Naive Bayes Dalam Analisis Sentimen Pemindahan Ibukota Pada Twitter," peneliti menggunakan analisis sentimen untuk mengetahui opini dominan masyarakat terkait topik ini. Mereka menggunakan pendekatan naïve Bayes untuk menganalisis sentimen terkait pindah ibukota. Tujuan penelitian ini adalah untuk memahami sentimen yang berkembang di masyarakat serta seberapa akurat topik ini dipahami. Setelah tahap preprocessing, penelitian ini menggunakan 822 tweet yang terdiri dari 317 sentimen negatif, 298 positif, dan 207 sentimen netral. Dengan membagi data menjadi 80% untuk pelatihan dan 20% untuk pengujian, hasil akurasi tertinggi yang dicapai adalah sebesar 60,606%.

Studi Sri Lestari, Mupaat Mupaat, dan Adhitia Erfina menganalisis sentimen masyarakat Indonesia tentang pemindahan Ibu Kota Negara Indonesia, termasuk penggunaan nama Nusantara di media sosial Twitter. Studi ini membandingkan tiga algoritma: Support Vector Machine (SVM), Naïve Bayes, dan K-Nearest Neighbor (KNN). Penelitian ini menghasilkan 1.141 komentar positif dan 591 komentar negatif. Hal ini menunjukkan bahwa orang-orang di Indonesia menyambut Ibu Kota baru negara. Pengujian kalsifikasi dan model menggunakan validasi cross-cross sepuluh kali. Nilai akurasi algoritma SVM 85,71%, Naive Bayes 76,70%, dan KNN 52,74%, masing-masing. Penelitian ini menunjukkan bahwa algoritma SVM lebih baik daripada algoritma Naive Bayes dan KNN. Nilai akurasi algoritma KNN rendah karena sensitif terhadap fitur yang tidak penting.[14]

Saepul Aripiyanto, Tukino Ammar Sufyan, Riandi Nandaputra, Dengan relokasi ibu kota, muncul beragam pendapat dari masyarakat yang pro dan kontra. Untuk

memahami sentimen masyarakat terkait isu tersebut, dilakukan analisis menggunakan model Long Short-Term Memory (LSTM) dan lexicon based. Dua skenario digunakan: satu dengan dataset berisi 100 tweet, dan yang lainnya dengan dataset berisi 5112 tweet. Hasil dari skenario pertama menunjukkan akurasi sebesar 64%, presisi 40%, recall 64%, dan F1-Score 79%. Sedangkan pada skenario kedua, akurasi mencapai 79%, presisi 82%, recall 79%, dan F1-Score 79%. Analisis sentimen dari dataset 5112 tweet menunjukkan bahwa 44,8% sentiment positif, 36,2% sentiment negatif, dan 19,0% netral. Penelitian ini menunjukkan bahwa jumlah data yang digunakan mempengaruhi performa model deep learning yang dikembangkan menggunakan lexicon based dan Algoritma LSTM.

Studi yang dilakukan oleh Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, dan Fitri Nurapriani memfokuskan pada analisis sentimen masyarakat Indonesia terkait pemindahan Ibu Kota Nusantara menggunakan Algoritma Naive Bayes dan KNN. Tujuannya adalah untuk memahami perasaan masyarakat terkait perubahan tersebut. Metode Naive Bayes (NB) dan K-Nearest Neighbor (KNN) digunakan dalam penelitian ini. Hasil analisis menunjukkan bahwa metode Naive Bayes memiliki tingkat akurasi sebesar 82,27%, ketepatan sebesar 86,36%, dan nilai recall sebesar 76,93%. Di sisi lain, metode KNN menunjukkan tingkat akurasi sebesar 88,12%, ketepatan 93.98%, dan nilai recall 81,53%. Dari hasil ini, dapat disimpulkan bahwa metode KNN lebih unggul daripada metode NB dalam mengukur sentimen terhadap pemindahan Ibu Kota Nusantara.

2.2 Penelitian Terkait

Tabel 2. 1 Matriks literatur review dan posisi penelitian

No	Nama dan Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Hasil Penelitian	Kelemahan	Saran
1	Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN	Syahril Dwi Prasetyo, Shofa Shofiah Hilabi , Fitri Nurapriani 2023 Vol. 10 No. 1 Hal: 1-7 p-ISSN: 2356-0010, e-ISSN: 2502-8758	Studi ini menganalisis perasaan masyarakat Indonesia tentang pemindahan Ibu Kota Nusantara. Penelitian ini menggunakan metode Naive Bayes (NB) dan K-Nearest Neighbor (KNN).	Hasil penelitian menunjukkan bahwa metode Naive Bayes memiliki tingkat akurasi analisis sentimen 82,27%, nilai precision 86,36%, dan nilai recall 76,93%. Metode KNN juga memiliki tingkat akurasi 88,12%, precision 93,%98, dan recall 81,53%.	Keterbatasan Metode Klasifikasi, Keterbatasan Data, Keterbatasan Validasi, Keterbatasan Proses Cleaning	Penggunaan Metode Klasifikasi Lain: Disarankan untuk memperluas metode klasifikasi yang digunakan, seperti Decision Trees, Support Vector Machines, atau Neural Networks, untuk membandingkan dan memperoleh hasil yang lebih komprehensif . Penambahan Data: Penting untuk menambah jumlah data yang digunakan dalam penelitian untuk meningkatkan generalisasi dan keakuratan hasil. Proses pengumpulan data yang lebih luas dapat membantu dalam mengatasi keterbatasan variasi data .
2	Penerapan Algoritma Naive Bayes Dalam Analisis Sentimen Pemindahan Ibukota Pada Twitter	Andrew Kurniawan ,Sejati Waluyo Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI) Jakarta – Indonesia, 6 September 2022	Penelitian ini bertujuan untuk mempelajari metode preprocessing data, pengujian akurasi, dan rancangan menu aplikasi Naive Bayes. Selain itu, penelitian ini akan membahas penggunaan algoritma Naive Bayes dalam analisis sentimen untuk pemindahan ibukota Indonesia	Hasil penelitian ini mencakup penggunaan algoritma Naive Bayes untuk menganalisis sentimen tentang pemindahan ibu kota Indonesia dari Jakarta ke Kalimantan Timur, menggunakan data tweet dari Twitter. Studi ini menemukan bahwa masyarakat Indonesia memiliki perasaan paling	Salah satu kelemahan dari penelitian ini adalah terbatasnya jumlah data yang digunakan dalam analisis sentimen. Penelitian ini hanya menggunakan 822 tweet untuk mewakili sentimen masyarakat Indonesia terkait pemindahan ibukota. Jumlah data yang terbatas dapat	Berdasarkan kelemahan yang telah diidentifikasi dalam penelitian ini, beberapa saran dapat diberikan untuk penelitian selanjutnya. Pertama, disarankan untuk meningkatkan jumlah data yang digunakan dalam analisis sentimen agar model

			dari Jakarta ke Kalimantan Timur, menggunakan data tweet dari Twitter.	sering negatif tentang pemindahan ibukota. Dalam analisis sentimen ini, akurasi tertinggi adalah 60.606%. Proses preprocessing data digunakan untuk membersihkan data, termasuk menghapus data ganda, membersihkan data suara, dan melakukan pembobotan kata menggunakan metode TF-IDF. Selain itu, untuk klasifikasi sentimen pada data tweet, metode Naive Bayes digunakan untuk melakukan pengujian akurasi; hasilnya berbeda-beda tergantung pada pembagian data pelatihan dan uji. Semoga penelitian ini dapat membantu penelitian terkait analisis sentimen dengan data media sosial seperti Twitter	mempengaruhi akurasi dan generalisasi dari model yang dikembangkan. Selain itu, kelemahan lainnya adalah penggunaan metode Naive Bayes yang bersifat "naive" atau sederhana, sehingga tidak mempertimbangkan ketergantungan antar fitur dalam data.	yang dikembangkan dapat lebih akurat dan generalisasi. Kedua, untuk mengatasi keterbatasan metode Naive Bayes yang bersifat sederhana, penelitian selanjutnya dapat mempertimbangkan penggunaan metode klasifikasi yang lebih kompleks seperti Support Vector Machine (SVM) atau Neural Networks. Selain itu, penelitian juga dapat mempertimbangkan penggunaan teknik ensemble learning untuk meningkatkan akurasi model.
3	Analisis Sentimen Pemindahan Ibu Kota Negara Indonesia Menggunakan Tiga Algoritma: Naïve Bayes, Knn, Dan Random Forest	Joshua Muliawan , Erick Dazki Jurnal Teknik Informatika (JUTIF) Vol. 4, No. 5, October 2023, pp. 1227-1236	Tujuan dari penelitian ini adalah untuk mendapatkan pemahaman yang lebih baik tentang bagaimana masyarakat menanggapi pemindahan Ibu Kota Negara Indonesia dan untuk menemukan istilah yang sering muncul dengan sentuhan positif, negatif, dan netral.	Studi ini menguji nilai keakuratan dari tiga algoritma pengklasifikasian: Naïve Bayes Classifier, K-Nearest Neighbor, dan Random Forest. Hasil klasifikasi data menunjukkan bahwa sentimen positif sebesar 36.8%, sentimen netral sebesar 25%, dan sentimen negatif terkait pemindahan ibu kota sebesar 38.1%.	Keterbatasan dalam pengumpulan data dari media sosial Twitter dapat mempengaruhi representasi sentimen masyarakat secara keseluruhan [5]. Penggunaan algoritma Naïve Bayes, K-Nearest Neighbor, dan Random Forest dalam analisis sentimen dapat memiliki tingkat akurasi yang berbeda-beda, sehingga hasil penelitian dapat dipengaruhi oleh pilihan	Berdasarkan kelemahan yang telah diidentifikasi, beberapa saran untuk penelitian ini adalah: 1. Memperluas sumber data selain dari media sosial Twitter untuk mendapatkan representasi yang lebih luas dari respons masyarakat terhadap pemindahan Ibu

					algoritma yang digunakan [4]. Proses pre-processing data seperti case folding, stopword removal, word normalizer, stemming, dan tokenizing dapat mempengaruhi hasil analisis sentimen jika tidak dilakukan dengan benar [5]. Pelabelan data secara manual dapat menimbulkan bias subjektif dari peneliti dalam menentukan sentimen dari setiap data [7].	Kota Negara Indonesia [9]. 2. Melakukan uji coba dengan berbagai algoritma analisis sentimen selain Naïve Bayes, K-Nearest Neighbor, dan Random Forest untuk memastikan hasil yang lebih konsisten dan akurat [8]. 3. Memperhatikan dengan seksama proses pre-processing data seperti case folding, stopword removal, word normalizer, stemming, dan tokenizing agar tidak terjadi bias dalam analisis sentimen [5]. 4. Menggunakan metode pelabelan data yang lebih objektif dan terstruktur untuk menghindari bias subjektif dalam penentuan sentimen dari setiap data [6].

4	Sentimen Analisis Twitter Ibu Kota Negara Nusantara Menggunakan Long Short-Term Memory dan Lexicon Based	Jurnal Manajemen system informasi dan teknologi Saepul Aripiyanto *, Tukino , Ammar Sufyan , Riandi Nandaputra	Penelitian ini bertujuan untuk mengkaji sentimen yang muncul di Twitter terkait IKN Nusantara dengan menggunakan algoritma LSTM dan metode berbasis leksikon. Fokus penelitian ini adalah untuk menilai kinerja model deep learning menggunakan dataset yang beragam dan untuk mengidentifikasi apakah sentimen yang terungkap terhadap IKN Nusantara bersifat positif, negatif, atau netral.	Dalam penelitian ini, model yang digunakan untuk melakukan analisis sentimen adalah algoritma Long Short-Term Memory (LSTM) dan metode berbasis leksikon, dengan dua skenario yang berbeda, yaitu menggunakan dataset 100 tweet dan 5112 tweet. Hasil dari skenario pertama menunjukkan akurasi sebesar 64%, presisi 40%, recall 64%, dan F1-Score 79%. Sedangkan pada skenario kedua dengan 5112 data tweet, diperoleh akurasi 79%, presisi 82%, recall 79%, dan F1-Score 79%. Analisis sentimen dari data 5112 tweet menunjukkan bahwa 44,8% sentimen positif, 36,2% sentimen negatif, dan 19,0% sentimen netral. Berdasarkan penelitian ini, dapat disimpulkan bahwa jumlah dataset berpengaruh pada kinerja model deep learning yang menggunakan metode berbasis leksikon dan Algoritma LSTM.	Kekurangan dari penelitian ini adalah keterbatasan dataset yang digunakan, yang dapat membatasi generalisasi hasil analisis sentimen. Selain itu, penggunaan pendekatan berbasis leksikon dalam menilai sentimen dapat menghasilkan hasil yang kurang akurat karena leksikon yang digunakan mungkin memiliki keterbatasan. Penelitian ini juga tidak memperhitungkan faktor-faktor eksternal seperti konteks sosial dan politik yang mungkin memengaruhi sentimen publik di Twitter terhadap IKN Nusantara.	Saran untuk penelitian ini adalah untuk meningkatkan cakupan dataset yang digunakan guna memastikan hasil analisis sentimen lebih mewakili variasi opini secara keseluruhan. Selain itu, disarankan untuk menimbang penggunaan metode analisis sentimen yang lebih canggih dan akurat serta untuk memperhitungkan faktor-faktor eksternal seperti konteks sosial dan politik yang mungkin memengaruhi sentimen publik.
5	Analisis Sentimen Masyarakat Indonesia terhadap Pemindahan Ibu Kota Negara Indonesia pada Twitter	Sri Lestari, Mupaat Mupaat, Adhitia Erfina* Jusifo p-ISSN: 2460-092X, e-ISSN: 2623-1662	Tujuan dari penelitian ini adalah untuk menganalisis perasaan masyarakat Indonesia tentang pemindahan Ibu Kota Negara Indonesia, termasuk nama yang dipilih,	Penelitian ini menunjukkan bahwa algoritma SVM dapat mencapai nilai akurasi 85,71%, algoritma Naive Bayes 76,70%, dan algoritma KNN 52,74%.	Salah satu kelemahan dari proses filter stopword adalah kemungkinan terjadi penghapusan kata-kata yang sebenarnya penting namun termasuk dalam kategori stopword. Hal ini dapat	Saran untuk mengatasi kelemahan dalam proses analisis teks adalah dengan melakukan validasi dan evaluasi yang cermat terhadap setiap tahapan preprocessing data. Hal ini

		Vol. 8, No. 1, Juni 2022	Nusantara, di media sosial Twitter.	bekerja lebih baik daripada algoritma Naive Bayes dan KNN. Nilai akurasi KNN rendah karena sensitif terhadap fitur yang tidak penting.	mengakibatkan hilangnya informasi yang relevan dalam analisis teks . Pada tahap tokenization, kelemahannya terletak pada kemungkinan kesalahan pemisahan kata-kata yang kompleks atau kata-kata yang seharusnya tetap tergabung. Hal ini dapat memengaruhi hasil analisis teks secara keseluruhan . Dalam proses data preprocessing, salah satu kelemahannya adalah jika tidak dilakukan dengan benar, dapat menghasilkan data yang tidak akurat atau tidak representatif. Misalnya, jika langkah-langkah preprocessing tidak mempertimbangkan konteks data dengan baik, hasil analisis dapat menjadi bias atau tidak valid .	dapat dilakukan dengan membandingkan hasil analisis sebelum dan sesudah preprocessing untuk memastikan tidak terjadi kehilangan informasi yang penting . Selain itu, penggunaan metode analisis teks yang lebih kompleks dan canggih seperti deep learning atau natural language processing dapat membantu meningkatkan akurasi dan validitas hasil analisis teks .

2.3 Landasan Teori

2.3.1 Analisis Sentimen

Analisis sentimen adalah bidang penelitian yang menganalisis pendapat, perasaan, penilaian, penilaian, sikap, dan perasaan terhadap suatu objek, organisasi, individu, benda, atau peristiwa tertentu.[19] Penelitian ini memiliki beberapa nama. Yaitu, analisis sentimen, penambangan opini, Eksperimen opini, dan penambangan emosi. Semua ini termasuk dalam domain analisis sentimen atau penambangan opini. Istilah "analisis sentimen" sering digunakan dalam industri, tetapi istilah "analisis sentimen" dan "penambangan opini" sering digunakan dalam dunia pendidikan. Kata "analisis sentimen" pertama kali diperkenalkan oleh Nasukawa dan Lee pada tahun 2003, dan kata "penambangan opini" diperkenalkan oleh Dave pada tahun 2003. *Lawrence dan Pennock*. [20]

Linguistik dan Pemrosesan Bahasa Alami (disingkat NLP). Sejarah panjang, tetapi sedikit penelitian dilakukan sebelum tahun 2000 tentang pendapat dan perasaan orang. Sejak itu, bidang ini menjadi bidang penelitian yang sangat aktif. Ada beberapa alasan untuk ini. Pertama, ini banyak digunakan di hampir setiap bidang. Salah satu motivasi kuat untuk penelitian adalah pengembangan aplikasi komersial analisis sentimen di lingkungan industri. Kedua, menampilkan sejumlah besar masalah penelitian yang sulit yang belum pernah diteliti sebelumnya.

2.3.2 *Naïve Bayes*

Naive Bayes adalah naif dengan dua interpretasi yang berbeda. Dalam interpretasi *Bayesian*, naif ini menunjukkan seberapa besar tingkat kepercayaan subjektif perlu diubah secara wajar jika terjadi tanda-tanda baruKesederhanaan ini menjelaskan representasi terbalik dari probabilitas dari dua peristiwa dalam interpretasi frekuensi.[21]

Naive Bayes ditemukan oleh Pendeta *Thomas Bayes* pada abad ke-18. Ini adalah dasar dari statistik *Bayesian* dan dapat diterapkan dalam banyak bidang, seperti sains, teknik, ekonomi, teori permainan kesehatan, dan hukum. Perhitungan Bayesian rincinya adalah :

pembuktian tunggal (E) dan hipotesis tunggal (H) pada persamaan (1),

perhitungan pembuktian tunggal (E) pada persamaan (1) dan hipotesis ganda (H1, H2, ... Hn). Hitung 2), banyak bukti dan beberapa hipotesis persamaan (3), dan probabilitas bersyarat dari semua kombinasi, dan ganti persamaan (4).

Envidence Tunggal (E) dan hipotesis tunggal (H)

$$\mathbf{P(H|E)} = \frac{\mathbf{P(E|H)xP(H)}}{\mathbf{P(E)}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Keterangan :

P(H|E) = probabilitas hipotesis H terjadi jika envidence E terjadi

P(E|H) = probabilitas munculnya envidence E jika hipotesis H terjadi

P(H) = probabilitas hipotesis H tanpa memandang envidence apapun

P(E) = probabilitas envidence E tanpa memandang apapun

1. Envidence tunggal (E) dan hipotesis ganda (H1, H2,...Hn)

$$\mathbf{P(H_i|E) = \frac{P(E|H) \times P(H)}{\sum_{k=1}^{n} P(E|H) \times P(H_k)}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Keterangan

P(Hi|E) = probabilitas hipotesis Hi benar terjadi jika diberikan envidence E

P(Ei|H) = probabilitas munculnya envidence E jika diketahui hipotesis Hi Benar

P(Ei|H) = hipotesis Hi benar terjadi Hi (menurut hasil sebelumnya tanpa memandang envidence apapun

n = jumlah hipotesis yang mungkin

2. Envidence ganda dan hipotesis ganda

$$\mathbf{P(H_i|E_1E_2 \ldots E_m) = \frac{P(E_1|H_i) \times P(E_2|H_i) \times \ldots P(E_1|H_i) \times P(H)}{\sum_{k=1}^{n} P(E_1E_2 \ldots E_m|H_k) \times P(H_k)}} \ldots\ldots\ldots (3)$$

Namun, karena kita perlu mengetahui semua kemungkinan bersyarat dari setiap kombinasi, pengeplikasian ini tidak mungkin. Akibatnya, persamaan ini digantikan dengan persamaan :

$$\mathbf{P(H_i|E_1E_2 \ldots E_m) = \frac{P(E_1|H_i) \times P(E_2|H_i) \times \ldots \times P(E_m|H_i) \times P(H)}{\sum_{k=1}^{n} P(E_1|H_i) \times P(E_2|H_k) \times \ldots \times P(E_m|H_k) \times P(H)}} \ldots\ldots\ldots (4)$$

2.3.2 *Support Vector Machine (SVM)*

Diciptakan oleh Boser, Guyon, dan Vapnik, *Support Vector Machine* (SVM) pertama kali disebutkan pada tahun 1992 di Annual Workshop on *Computational Learning Theory*. Ide dasar metode SVM sebenarnya berasal dari teori komputasi yang telah ada pada tahun sebelumnya. Metode solver variabel (SVM) digunakan untuk memprediksi kedua prediksi dalam kasus regresi dan klasifikasi.[22] Teknologi SVM untuk mendapatkan fungsi pemisahan yang optimal *(hyperplane)* untuk pemisahan Pengamatan dengan nilai variabel target yang berbeda.[23] *Hyperplane* ini dapat berupa line pada two dimension dan dapat berupa flat plane pada multiple dimension.

Nugroho (2003) menguraikan beberapa fitur SVM sebagai berikut: pertama, SVM adalah classifier linear secara prinsip; kedua, pengidentifikasian pola dilakukan dengan mengubah data dari ruang input (input space) ke ruang feature (feature space), dan ketiga, optimisasi dilakukan pada ruang vektor yang baru diciptakan. Ini membedakan SVM dari solusi pengidentifikasian pola yang umumnya melakukan optimisasi parameter pada hasil transformasi nya, Menerapkan strategi Minimisasi Risiko Struktural (SRM). Prinsip kerja SVM pada dasarnya hanya mampu menangani klasifikasi dua kelas, tetapi telah dikembangkan untuk menangani klasifikasi lebih dari dua kelas dengan pengakuan pola.

2.3.4 *K-Fold Cross Validation*

Metode cross-validation adalah metode pengujian yang dilakukan dengan membagi data menjadi dua bagian: instruksi data dan pengujian data.[24] Dalam validasi cross-K-fold, data awal dibagi secara acak menjadi subset k yang terpisah dari yang

lain, atau "kelipatannya". Subset D1, D2, dll., masing-masing memiliki ukuran yang sama. Data latihan dan tes dilakukan n kali. Dalam iterasi pertama, subset-subset D2 sampai Dk digunakan sebagai set pelatihan untuk mendapatkan model pertama yang akan diuji pada subset di. Dalam iterasi kedua, pardisi Di diperuntutkan sebagai set uji, dan partisi sisanya digunakan secara kolektif untuk pelatihan model. Berbeda dengan metode Holdout dan Subsampling Secara Acak, Metode ini menggunakan jumlah yang sama untuk setiap pelatihan data dan sekali untuk pengujian untuk setiap sempel. Untuk rekomendasi, estimasi akurasi adalah jumlah total rekomendasi yang benar dari n iterasi dibagi dengan jumlah total tupel atau rekaman data awal.[25]

2.3.5 Metode Pengujian *Confusion Matrix*

Confusion matrix digunakan untuk mengevaluasi kinerja algoritme pembelajaran mesin (ML). Confusion matrix menunjukkan prediksi dan kondisi sebenarnya (aktual) dari data yang dihasilkan ML. Dengan menggunakan Confusion matrix, kita dapat menentukan akurasi, ketepatan, pengenalan, dan waktu. Dalam tulisan ini, saya membahas pengukuran kinerja hasil ML. Berdasarkan tulisan ini, saya kemudian menulis

Tabel 2. 2 Pengujian Confusion Matrix

Pengukuran *Performance Accuracy*	merupakan rasio prediksi Benar dengan data lengkap. Akibat kesuluruhan pengguna Twitter, Akurasi menjawab pertanyaan, "Berapa persen masyarakat yang benar diprediksi Respon Positif dan Respon Negatif terhadap pengesahan UU Ibu Kota Nusantara?"” Akurasi = (TP + TN) / (TP+FP+FN+TN).

	Dalam contoh sebelumnya, akurasi = (4+3) / (4+2+1+3) = 7/10 = 70%.
Precission	adalah rasio prediksi benar yang positif dibandingkan dengan total hasil prediksi yang positif. Precission menjawab pertanyaan, "Berapa persen masyarakat yang benar diprediksi Respon Positif dan Respon Negatif terhadap pengesahan UU Ibu Kota Nusantara?""dari kesuluruhan pengguna Twitter?"" Presisi = (TP) / (TP+FP) Pada contoh di atas, ketepatan = 4/(4+2) = 4/6 = 67%.
Recall **(Sensitifitas)**	merupakan rasio prediksi benar positif dibandingkan dengan jumlah data yang benar positif secara keseluruhan. Recall memberikan jawaban atas pertanyaan, "Berapa persen masyarakat yang benar diprediksi Respon Positif dan Respon Negatif terhadap pengesahan UU Ibu Kota Nusantara dari kesuluruhan pengguna Twitter?"" Recall = (TP) / (TP + FN) pada contoh kasus di atas Recall = 4/(4+1) = 4/5 =80%.
Waktu Training ***(Times)***	merupakan rasio waktu dari awal pelatihan hingga akhir pelatihan. Jumlah waktu yang dihabiskan untuk pelatihan digunakan untuk mengukur seberapa cepat alogoritma dapat menyelesaikan masalah perhitungan. Waktu latihan = Mulai (t start) - Tutup (t end).

2.3.6 Pemodelan Bahasa TF-IDF

Pemodelan Bahasa TF-IDF (Term Frequency-Inverse Document Frequency) adalah salah satu teknik yang digunakan dalam pengolahan bahasa alami dan analisis teks untuk mengekstraksi dan memahami makna dari suatu dokumen atau teks.[26][27] Konsep dasar di balik TF-IDF adalah untuk mengevaluasi seberapa penting suatu kata dalam suatu dokumen dalam kumpulan dokumen. Hal ini dilakukan dengan menghitung frekuensi kata dalam dokumen

(TF) dan kemudian menyesuaikannya dengan frekuensi kemunculan kata tersebut di seluruh dokumen dalam kumpulan (IDF).

Pada tahap pertama, TF-IDF menghitung frekuensi kata dalam suatu dokumen. Ini dilakukan dengan menghitung berapa kali sebuah kata muncul dalam dokumen tersebut. Kata-kata yang muncul lebih sering cenderung memiliki bobot yang lebih tinggi dalam analisis ini karena kemungkinan besar kata-kata tersebut memiliki relevansi yang lebih besar dengan topik dokumen.[28][29]

Selanjutnya, IDF digunakan untuk menyesuaikan bobot kata berdasarkan frekuensi kemunculan kata tersebut di seluruh dokumen dalam kumpulan. Kata-kata yang muncul di seluruh dokumen dengan frekuensi yang tinggi diberi bobot yang lebih rendah, sementara kata-kata yang muncul hanya dalam beberapa dokumen memiliki bobot yang lebih tinggi karena kemungkinan relevansi yang lebih besar dengan topik spesifik.[30]

Gabungan antara TF dan IDF memberikan nilai yang menunjukkan seberapa penting suatu kata dalam suatu dokumen dalam konteks kumpulan dokumen. Kata-kata yang memiliki nilai TF-IDF tinggi cenderung menjadi kata-kata kunci yang mewakili makna atau topik utama dari dokumen tersebut. Dengan demikian, pemodelan bahasa TF-IDF memungkinkan untuk mengekstrak informasi yang relevan dan memahami konten suatu dokumen secara lebih baik.[31] Teknik ini sering digunakan dalam berbagai aplikasi seperti klasifikasi teks, pencarian informasi, dan analisis sentimen.

2.3.7 Pemodelan Bahasa Word Embedding

Pemodelan Bahasa Word Embedding adalah teknik dalam pengolahan bahasa alami yang digunakan untuk merepresentasikan kata-kata dalam bentuk vektor dalam ruang multidimensi.[32][33] Konsep dasar dari word embedding adalah mengonversi kata-kata dalam suatu teks menjadi vektor numerik, di mana setiap dimensi dalam vektor merepresentasikan makna atau karakteristik tertentu dari kata tersebut. Representasi ini memungkinkan komputer untuk memahami makna kata-kata dalam konteks yang lebih dalam, bahkan dalam kasus-kasus di mana kata-kata tersebut tidak secara langsung terkait secara semantik.

Salah satu teknik word embedding yang paling populer adalah Word2Vec, yang dikembangkan oleh Google. Word2Vec bekerja dengan cara melatih model neural network pada sejumlah besar teks untuk mempelajari representasi vektor kata-kata yang berarti.[34] Dengan menggunakan teknik ini, kata-kata yang sering muncul bersamaan dalam konteks yang serupa akan memiliki representasi vektor yang lebih dekat satu sama lain dalam ruang embedding. Word embedding memungkinkan untuk melakukan berbagai jenis analisis dan tugas pengolahan bahasa alami. Misalnya, dengan menggunakan representasi vektor kata-kata, kita dapat melakukan klasifikasi teks, pencarian semantik, pemrosesan bahasa alami, dan bahkan penerjemahan mesin.[35] Selain itu, word embedding juga memungkinkan komputer untuk memahami hubungan antara kata-kata, seperti sinonim, antonim, dan hubungan semantik lainnya.