BAB II

TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Tinjauan Pustaka

Adapun beberapa Penelitian terdahulu yang sejenis sehingga digunakan sebagai acuan studi pustaka dalam Penelitian ini. Adapun tinjauan pustaka pada Penelitian terdahulu diantaranya adalah:

Dalam Klasifikasi Algoritma *K-Nearest Neighbor* dan *Naïve Bayes* Kelayakan Pemberian Kredit Perbankan. Penilitian ini melakukan klasifikasi kelayakan kredit pada kredit pemilikan rumah (KPR), menganalisis keberhasilan pengelolahan kredit bank untuk hasil yang akurat. Untuk mengetahui tingkat akurasi algoritma peneliti ini melalui tiga tahap pengujian, yaitu dilakukan beberapa tahapan *pre-processing* mulai dari pengecekan *duplicate*, menangani *outliers*, *missing value*, melakukan *label encoding*, mengatasi data *imbalance* menggunakan metode *smote* dan melakukan standarisasi menggunakan *standar scaler*.(Asyari, B., 2023).

Analisis Sentimen Pembangunan Kereta Cepat Jakarta-Bandung di Media Sosial *Twitter* Menggunakan Metode *Naïve* Bayes. Penelitian ini mengangkat masalah tentang pembangunan kereta cepat Jakarta-Bandung (KCJB) dengan membutuhkan dana yang cukup besar. Sehingga menimbulkan suatu permasalahan seperti perbedaan pendapat dan pernyataan setuju dan tidak setujunya pembangunan kereta cepat Jakarta-Bandung. Untuk metode yang digunakan yaitu *Naïve Bayes* karena mempunyai nilai probabilitas atau peluang tinggi untuk

pengklasifikasian data, untuk pembobotan menggunakan perhitungan TF-IDF, dan pengujian data menggunakan *Confusion Matrix* (Sianipar et al., 2023).

Penelitian yang dilakukan oleh Muttaqin & Kharisudin (2021) terkait Analisis Sentimen Aplikasi Gojek Menggunakan *Support Vector Machine* dan *K-Nearest Neighbor*" Penelitian yang dilakukan dengan tujuan melakukan klasifikasi pada ulasan Aplikasi Gojek menggunakan *metode Support Vector Machie (SVM) dan K Nearest Neighbor* (KNN). Dengan melakukan perbandingan tingkat akurasi antara metode *metode Support Vector Machie (SVM) dan K Nearest Neighbor* (KNN). Sumber data terbaik untuk mendapatkan dataset dalam penelitian analisis sentimen adalah *Google Playstore,* dikarenakan data yang diperoleh lebih bersih dan tidak mengandung unsur iklan atupun promosi.

Dalam Analisis Sentimen Pengguna Kereta Api Indonesia melalui Sosial Media *Twitter* dengan Algoritma *Naïve Bayes Classifier*. Dalam penelitian data diambil dari reaksi masyarakat terhaap pelayanan Kereta Api Indonesia yang dituangkan di media sosial *Twitter*. Penelitian dilakukan untuk mengklasifikasi sentimen pengguna layanan transportasi Kereta Api Indonesia menggunakan algoritma *Naïve Bayes Classifier*. Penelitian terdiri dari tahapan *Data Selection, pre-processing, Data transformation, Data Mining* dan *Evaluation.* (Azahri et al., 2023).

Menurut Mulya et al., (2023) Analisis Sentimen Masyarakat terhadap Pembangunan Kereta Cepat Jakarta-Bandung Menggunakan Algoritma *K-Nearest Neighbor* (KNN)". Penelitian dilakukan analisis sentimen masyarakat terhadap pembangunan kereta cepat Jakarta-Bandung. Penelitian ini menggunakan metode

deskriptif, karena bermaksud untuk mendapatkan fakta empiris, dan makna yang mendalam, mengamati, menangkap realitas dan memeriksa perilaku individu dan kelompok objek penelitian.

Penelitian ini juga akan membahas analisis sentimen masyarakat terhadap adanya kereta cepat Jakarta-Bandung menggunakan algoritma KNN. Berbeda dengan penelitian Mulya et al., (2023), pada penelitian ini dilakukan skenario 3 skema perbandingan data latih dan data uji yaitu 90:10, 80:20 dan 70:30 dengan menggunakan variasi nilai K ganjil antara 1 sampai 10. Hasil dari klasifikasi divisualisasikan dalam bentuk kurva ROC, sedangkan untuk pengolahan teks serta implementasi algoritma digunakan RapidMiner. Keterangan perbandingan hasil dari penelitian sebelumnya dapat dilihat pada tabel 2.1.

Tabel 2. 1 Perbandingan Dengan Penelitian Sebelumnya

No	Penulis	Topik	Metode	Keterangan
1	Asyari, B., (2023).	Klasifikasi Algoritma *K-Nearest Neighbor* dan Naïve Bayes untuk Kelayakan Pemberian Kredit Perbankan.	*K-Nearest Neighbor* dan *Naïve Bayes*	Hasil dari algoritma *Naïve Bayes* dan KNN dengan tahapan model di evaluasi data terhadap kemampuan model dalam prediksi, matrik evaluasi yang digunakan berupa hasil *Confusion Matrix.* Terdapat hasil terbaik yaitu pada algoritma KNN di pengujian keiga dengan nilai K=10 dengan performa akurasi data latih 80.92% dan uji 78.86% dan mendapatkan score *Confusion matrix* True Positive 76 dan True Negative 21.
2	Sianipar et al., (2023).	Analisis Sentimen Pembangunan Kereta Cepat Jakarta-Bandung di Media Sosial *Twitter* Menggunakan Metode Naïve Bayes	*Naïve Bayes*	Setelah melalui pemrosesan dengan hasil sentimen negatif sebanyak 673, hasil sentimen positif sebanyak 668, dan hasil sentimen netral sebanyak 665 hasil accuracy 71%, precision 73%, recall 89%. Dari hasil penelitian mendapatkan hasil tanggapan tergolong negatif terhadap pembangunan kereta cepat Jakarta-Bandung.

Tabel 2. 1 Perbandingan Dengan Penelitian Sebelumnya (Lanjutan)

No	Penulis	Topik	Metode	Keterangan
3	Muttaqin & Kharisudin (2021)	Analisis Sentimen pada Aplikasi Gojek Menggunakan *Support Vector Machine* dan *K-Nearest Neighbor*	*Support Vector Machine* dan *K-Nearest Neighbor*	Hasil akurasi menggunakan Metode KNN dengan nilai K=22 memperoleh nilai akurasi, presisi, dan recall berturut-turut sebesar 82,14%, 82,28%, dan 95,43%, sedangkan metode SVM dengan kernel linear dan parameter C=1 memperoleh nilai akurasi, presisi, dan recall berturut-turut sebesar 87,98%, 88,55%, dan 95,43%.
4	Azahri et al., (2023)	Analisis Sentimen Pengguna Kereta Api Indonesia Melalui Media *Twitter* Dengan Algoritma *Naive Bayes Classifier*	*Naive Bayes Classifier*	Hasil penelitian menunjukkan bahwa performa algoritma Naïve Bayes Classifier dalam mengklasifikasikan sentimen pengguna layanan transportasi Kereta Api Indonesia dengan nilai akurasi sebesar 0.9215686274509803 atau setara dengan 92.15%. Sentimen pada opini pengguna layanan transportasi Kereta Api Indonesia ini diketahui lebih banyak mengandung sentimen negatif berdasarkan pelabelan manual.
5	Mulya et al., (2023)	Analisis Sentimen Masyarakat terhadap Pembangunan Kereta Cepat Jakarta-Bandung menggunakan Algoritma *K-Nearest Neighbor*	*K-Nearest Neighbor*	Dari hasil penelitian ini telah dilakukan dan disimpulkan dari hasil akurasi AUC (Area Under Curve) dengan algoritma KNN, nilai keakuratannya 82.70% dapat dikategorikan sebagai Excellent Classification. Kemudian penelitian ini menghasilkan opini negative yang lebih tinggi, hal ini membuktikan bahwa masyarakat yang tinggal di bandung dan Jakarta lebih pro terhadap pembangunan kereta cepat Jakarta dan Bandung.
6	Triantara (2024)	Analisis Sentimen Masyarakat Terhadap Kereta Cepat Jakarta-Bandung pada Media Sosial X Menggunakan *K-Nearest Neighbor*	*K-Nearest Neighbor*	Penelitian dilakukan dengan metode *K-Nearest Neighbor* dengan menghitung akurasi opini publik terkait Kereta Cepat Jakarta-Bandung. Klasifikasi akan menghasilkan sentimen positif dan negatif dengan metode K-*Nearest Neighbor* dilakukan menggunakan perhitungan nilai K ganjil 1 sampai 10 dengan pembagian data menjadi 3 skema yaitu 90:10, 80:20, 70:30.

2.2 Landasan Teori

2.2.1 Analisis Sentimen

Analisis Sentimen adalah proses teks digital yang dianalisis untuk menentukan opini dari objek yang berujung pada sentimen positif, dan negatif. Analisis sentimen merupakan teknik untuk mengumpulkan informasi dari berbagai platform online. Fokusnya adalah pada menganalisis dan memahami perasaan yang terkandung

dalam teks ulasan dengan tujuan untuk memprediksi, menganalisis opini publik, suasana hati, dan gambaran perasaan secara otomatis dari pengguna internet terhadap suatu topik atau kasus (Que et al., 2020).

2.2.2 X (*Twitter*)

X merupakan *platform* media sosial *microblog* yang memungkinkan penggunannya untuk saling bertukar pesan blog namun platform ini membatasi penggunanya dengan batasan hingga 140 karakter yang dapat diposting pada halaman profil tiap penggunanya. Opini pengguna pada *twitter* ini dapat dijadikan sebagai sumber data untuk sebuah penelitian yang diambil dengan teknik *crawling* (Wahyu, 2023).

2.2.3 *Python*

Python merupakan salah satu dari bahasa pemrograman yang sering digunakan oleh programmer atau pembuat program dalam membuat program mereka. *Python* memiliki karakteristik sintaks yang tidak terlalu rumit. Sehingga *Python* menjadi salah satu Bahasa pemrograman tingkat tinggi yang mudah untuk digunakan. Dalam menulis sebuah kode program menggunakan bahasa pemrograman *Python*, terdapat beberapa aturan yang harus dipenuhi. Hal ini untuk mengantisipasi terjadinya error atau masalah pada program yang dibuat. Aturan sintaks *Python* yang pertama adalah dalam penulisan *Statement* atau perintah. (Mahawardana et al., 2022). *Python* juga menyediakan dukungan untuk menangani tipe data dinamis. Secara eksplisit dan dapat mengubah tipe datanya saat *runtime*. Dengan komunitas pengembang yang besar dan aktif, terus berkembang dan menjadi salah satu bahasa

pemrograman terpopuler di berbagai bidang seperti pengembangan perangkat lunak, analisis data, kecerdasan buatan, dan pengembangan *web*.

2.2.4 *Tweet Harvest*

Tweet Harvest merupakan sebuah alat baris perintah yang menggunakan *Playwright* untuk mengumpulkan *tweet* dari hasil pencarian X berdasarkan kata kunci dan rentan waktu tertentu. Untuk menggunakan *Tweet Harvest*, perlu memiliki *authorization token* yang diperoleh dengan login ke akun X melalui browser dan mengekstrak *authorization token* (Vincent, 2023).

2.2.5 Seleksi Fitur

Seleksi Fitur adalah salah satu tahapan dalam *text mining*. Dalam tahapan ini dilakukan proses pembuangan beberapa term atau kata yang tidak terkait sehingga mendapatkan term atau kata penting sebagai kumpulan dokumen yang di analisis. Dalam seleksi fitur terdapat beberapa metode yang digunakan yaitu *Term Frequency-Inverse Document Frequency* (TF-IDF), *N-Gram*, dan *Chi Square* (Razaq et al., 2023). *Term frequency* merupakan cara untuk mengukur seberapa sering sebuah kata muncul dalam sebuah dokumen. Bobot *term* dihitung dengan menggunakan rumus yang menjelaskan hubungan antara frekuensi kata tersebut dengan dokumen yang bersangkutan (Soleqah, 2023). Rumus 2.1 merupakan *Term Frequency*.

$$TF_{t.d} = \frac{ft,d}{\sum t' \in d ft',d} \quad (2.1)$$

Keterangan:

$TF_{t.d}$ = nilai *term frequency* dari *term* t dalam dokumen d.

ft, d= frekuensi kemunculan *term* t dalam dokumen d.

$\sum t' \in dft', d$ = total frekuensi kemunculan semua *term* dalam dokumen d.

2.2.6 *SMOTE Upsampling*

SMOTE Upsampling adalah teknik dalam pengolahan data yang digunakan untuk menangani ketidakseimbangan kelas dalam sebuah set data, terutama ketika kelas minoritas memiliki representasi yang jauh lebih rendah daripada kelas mayoritas. Dengan menerapkan *SMOTE Upsampling*, model pembelajaran mesin cenderung menjadi lebih baik dalam mengenali pola dalam kelas minoritas, yang menghasilkan kinerja yang lebih baik dalam tugas seperti klasifikasi di mana ketidakseimbangan kelas menjadi masalah (Irawan & Bahtiar, 2023).

2.2.7 *RapidMiner*

Rapid Miner sebelumnya Bernama YALE (*Yet Another Learning Environment*), versi awalnya mulai dikembangkan pada tahun 2001 oleh Ralf Klinkenverg, Ingo Mierswa, dan Simon Fischer di *Artificial Intelligence Unit* dari *University of Dortmund. RapidMiner* didistribusikan di bawah lisensi AGPL (*GNU Affero General Public License*) versi 3. *RapidMiner* menyediakan antarmuka grafis (GUI) yang memungkinkan pengguna merancang *pipeline* analitik. Melalui GUI ini, pengguna dapat menentukan proses analitik yang diinginkan untuk diterapkan pada data. Setelah selesai, GUI akan menghasilkan file XML (*Extensible Markup Language*) yang memuat definisi proses analitik tersebut. File XML ini kemudian dapat dibaca oleh *RapidMiner* untuk menjalankan analisis secara otomatis sesuai dengan yang diinginkan oleh pengguna. *RapidMiner* adalah perangkat lunak sumber terbuka yang digunakan untuk melakukan berbagai jenis analisis data,

seperti data mining, text mining, dan analisis prediksi. Ini merupakan solusi yang komprehensif untuk kebutuhan analisis data (Utami & Erfina, 2021).

2.2.8 *K-Nearest Neighbor*

Metode *K-Nearest Neighbor* (K-NN) adalah pendekatan yang digunakan untuk mengelompokkan objek berdasarkan data pelatihan yang memiliki jarak terdekat dengan objek yang ingin diklasifikasikan. Algoritma K-NN adalah kelas yang paling banyak muncul merupakan kelas yang ditentukan pada hasil klasifikasi. Kedekatan didefinisikan dalam jarak matrix, seperti jarak *Euclidean.* Teknik ini mengelompokkan objek berdasarkan data yang paling dekat dengan objek tersebut. Metode K-NN menggunakan klasifikasi berbasis tetangga terdekat sebagai 24 sentimen untuk menanyakan *instance* baru (Rivita et al., 2023).

Berikut ini adalah rumus metode *K-Nearest Neighbor* menghitung tingkat kesamaan dalam data *Euclidean distance.* Rumus 2.2 merupakan persamaan yang digunakan untuk menghitung *Euclidean distance*.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_{training}^{i} - y_{testing}^{i})^2} \qquad (2.2)$$

Keterangan:

D(x,y) = jarak antara dua titik x dan y

$x_{training}^{i}$ = Data Latih

$y_{testing}^{i}$ = Data Uji

n = Dimensi Data

i = Variabel Data

Nilai k pada umumnya dilakukan penentuan dalam jumlah ganjil (3, 5, 7, 9) untuk menghindari timbulnya jumlah jarak yang sama pada proses klasifikasi. Jika terjadi dua atau lebih jumlah kelas yang sering muncul sama maka nilai k menjadi k-1 (satu tetangga kurang), jika ada yang sama lagi maka nilai k menjadi k-2, maka seterusnya sampai tidak ada kelas yang sama banyak (Febrealti, 2011).

2.2.9 Model Evaluasi

Model evaluasi merupakan tahapan yang digunakan untuk mengukur performa dari setiap model dengan mencari model untuk mengetahui performa tertinggi dengan melakukan perhitungan dari akurasi, presisi, dan *recall*. Salah satu cara yang digunakan untuk menghitung akurasi tersebut adalah dengan menggunakan *confusion matrix*.

Confusion matrix merupakan salah satu cara untuk mengukur akurasi. *Confusion matrix* adalah alat yang bermanfaat dalam menganalisis seberapa baik pengklasifikasi mengidentifikasi data uji. Ketika dataset uji terdiri dari dua kelas, satu kelas dianggap sebagai positif sedangkan yang lainnya sebagai negatif (Mulya et al., 2023).

Menurut Wahyu (2023) pada pengujian akan menghasilkan data dari klasifikasi yang telah menentukan jumlah sentimen positif dan negatif yang diuji dalam metode *Confusion Matrix.* Untuk menghitung performa dari model melalui akurasi, presisi, *recall*, menggunakan persamaan sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.3)$$

$$Precision = \frac{TP}{TP+FP} \quad (2.4)$$

$$Recall = \frac{TP}{TP+FN} \quad (2.5)$$

Keterangan:

TP = *True Positif*

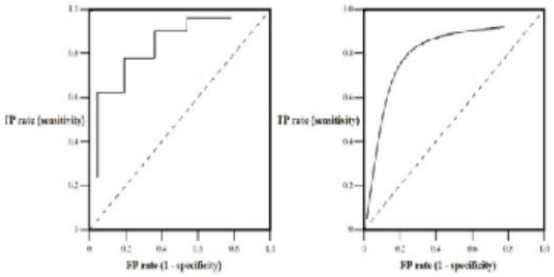
TN = *True Negatif*

FP = *False Positif*

FN = *False Negatif*

2.2.10 Pengukuran Menggunakan Kurva ROC

Hasil pengukuran validasi menggunakan kurva ROC dan Confusion Matrix hingga menghasilkan tingkat akuras tertinggi, kurva ROC merupakan kurva yang banyak dipakai untuk melakukan penilaian hasil dari prediksi, kurva ROC memvisualisasikan kinerja pengklasifikasian yang tidak memperhatikan distribusi kelas ataupun kesalahan, di sumbu *vertical* menunjukkan nilai positif (TP) dan sumbu horizontal merupakan nilai negative (FP).



Gambar 2. 1 Kurva Roc (Sumber: Sucipto,2012)

Sebuah garis diagonal yang memisahkan ruang ROC menggambarkan sebuah ruang diatas garis diagonal menunjukkan bahwa klasifikasi baik sedangkan ruang yang dibawah garis diagonal merupakan klasifikasi buruk, untuk tebakan benar-benar dilakukan acak pada sepanjang garis diagonal yang dimulai dari kiri bawah sampai kanan atas. Metode umum untuk melakukan perhitungan daerah bawah

kurva ROC yaitu *Area Under Curve (AUC)* yang mana bidang dibawah kurva memiliki nilai yang selalu berada dibawah nilai 0,0 dan 1,0, tetapi hal yang menarik untuk dilakukan perhitungan adalah memiliki luas diatas 0,5 sehingga jika semakin tinggi luasnya maka semakin baik. Seperti petunjuk yang disajjikan seperti berikut (Sucipto, 2012):

- 0,9 – 1.00 = klasifikasi sangat baik
- 0,8 - 0,9 = klasifikasi baik
- 0,7 – 0,8 = klasifikasi rata-rata
- 0,6 – 0,7 = klasifikasi rendah
- 0,5 – 0,6 = kegagalan