BAB II
LANDASAN TEORI

2.1 Text Mining

Text mining adalah penemuan informasi yang baru yang oleh computer, sebelumnya belum di ketahui. Singkatnya, *test mining* bekerja dengan mengekstrak informasi dari berbagai macam sumber secara otomatis dan menggabungkan berbagai informasi yang berhasil di ekstrak [4].

Text mining berbeda dari apa yang kita kenal dalam berbeda dengan pencarian *website* biasa yang pengguna mendapatkan informasi yang sebelumnya sudah diketahui atau ditulis orang lain. Dalam *text mining*, tujuannya adalah untuk menemukan informasi yang sebelumnya tidak diketahui, sesuatu yang belum diketahui siapa pun dan belum dapat ditulis. Menurut Dean Abbot beberapa tipe *text mining*, yaitu[5]:

1) *Search and Information Retrieval (IR)*

   Mengambil dan menyimpan dokumen *text*, termasuk *search engines* atau mesin pencari dan kata kunci.

2) *Document Clustering*

   Menggunakan metode clustering, mengelompokan dan mengkategorikan paragraph atau dokumen.

3) *Document Classification*

   Menggunakan metode klasifikasi, potongan *paragraph* digabungkan dan dikategorikan menurut trained model pada contoh label.

4) *Web Mining*

   Data dan *text mining* pada internet dengan spesifikasi khusus yang berfokus pada skala dan keterkaitan terhadap suatu web.

5) *Information Extraction*

Proses membuat data yang terstruktur dari yang sebelumnya tidak terstruktur ataupun semi struktur atau mengidenfitikasi dan mengekstrak fakta yang terkait dan berhubungan dari data yang tidak terstruktur.

6) *Natural Language Processing*

Bahasa yang mengandung deskripsi dan memiliki tingkat rendah pada prosesingnya.

7) *Concept Extraction*

Penggabungan kata – kata dan frasa kedalam kelompok yang mirip.

2.2 Analisis Sentimen

Analisis Sentimen adalah suatu teknik mengekstrak data teks untuk mendapatkan informasi tentang sentimen bernilai positif, netral maupun negatif[6]. Analisa sentimen adalah gambaran dari opini penulis terhadap sebuah topik.

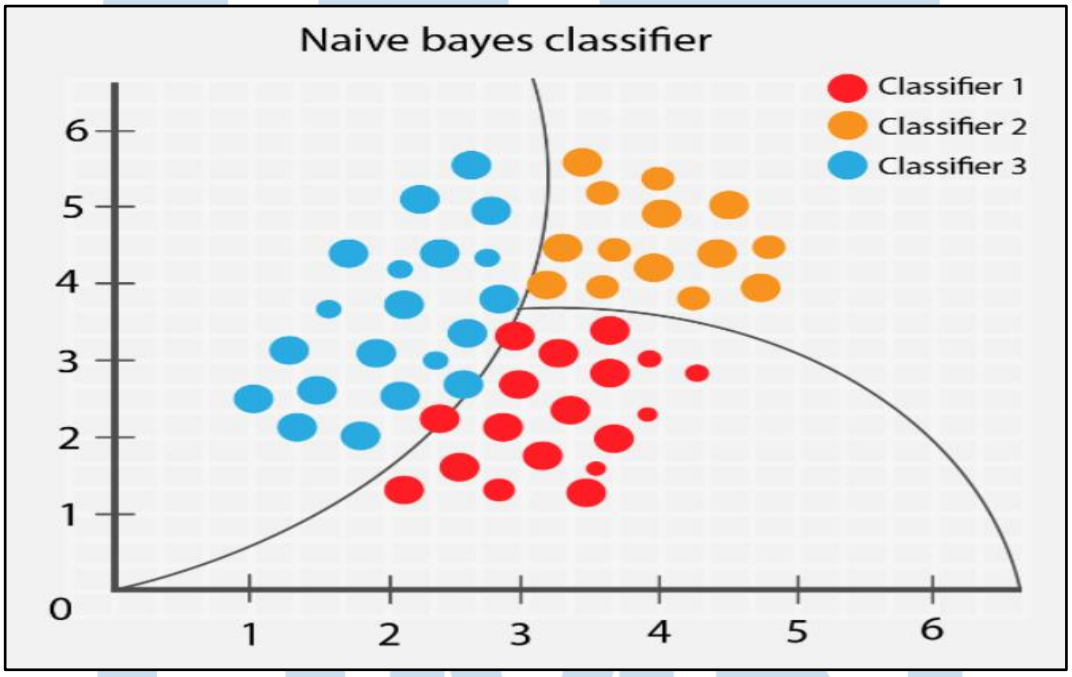
Sentiment analysis (analisis sentimen) atau sering disebut juga dengan *opinion mining* (penambangan opini) adalah studi komputasi untuk mengenali dan mengekspresikan opini, sentimen, evaluasi, sikap, emosi, subjektifitas, penilaian atau pandangan yang terdapat dalam suatu teks[3].

Opinion mining biasa dikenal juga dengan nama sentimen analisis biasanya digunakan untuk menggali emosi yang terkandung dibalik *feedback*/ulasan dari pelanngan suatu produk. Analisis sentimen biasa dilakukan untuk menarik tulisan-tulisan yang mengandung sentimen tertentu terhadap suatu topik untuk kemudian dilakukan klasifikasi sentimen.

Metode *text mining* yang biasa digunakan untuk menyelesaikan masalah *opinion mining* adalah *Naïve Bayes Classifier*, *Decision Tree* dan *Support Vector Machine (SVM)*. Metode-metode tersebut dapat dipakai untuk melakukan klasifikasi opini kedalam beberapa kelompok dan untuk mengklasifikasikan opini negatif dan positif yang terkandung sehingga hasil klasifikasi teks dapat menjadi maksimal[7].

2.3 Naïve Bayes Classifier

Naïve Bayes Classifier merupakan salah satu metode probabilitas sederhana yang bekerja dengan menjumlahkan frekuensi dan menghitung probabilitas serta kombinasi nilai dari dataset yang diberikan. Pengklasifikasi'an *Naïve Bayes* juga dilakukan dengan asumsi bahwa akibat dari suatu nilai atribut tertentu itu independen atau tidak tergantung dengan nilai atribut lainnya[5].



Gambar 2. 1 Naive Bayes Classifier

Naïve Bayes berkerja dengan mengasumsikan bahwa nilai atribut saling bebas apabila diberikan nilai outputnya. Probabilitas dari produk individu dapat diamati bersamaan dengan probabilitas dari produk layanan yang ingin diteliti. Pada situasi yang tidak tentu, biasanya *Naïve Bayes* akan bekerja lebih baik. Berikut pada rumus 2.1 adalah teorema dari *Naïve Bayes*.

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \quad (2.\ 1)$$

Keterangan:

X: Data dengan *class* yang belum diketahui.

H: Hipotesis data X merupakan suat *class* spesifik.

P(H|X): Probabilitas hipotesis H berdasarkan kondisi X (*posteriori probability)*.

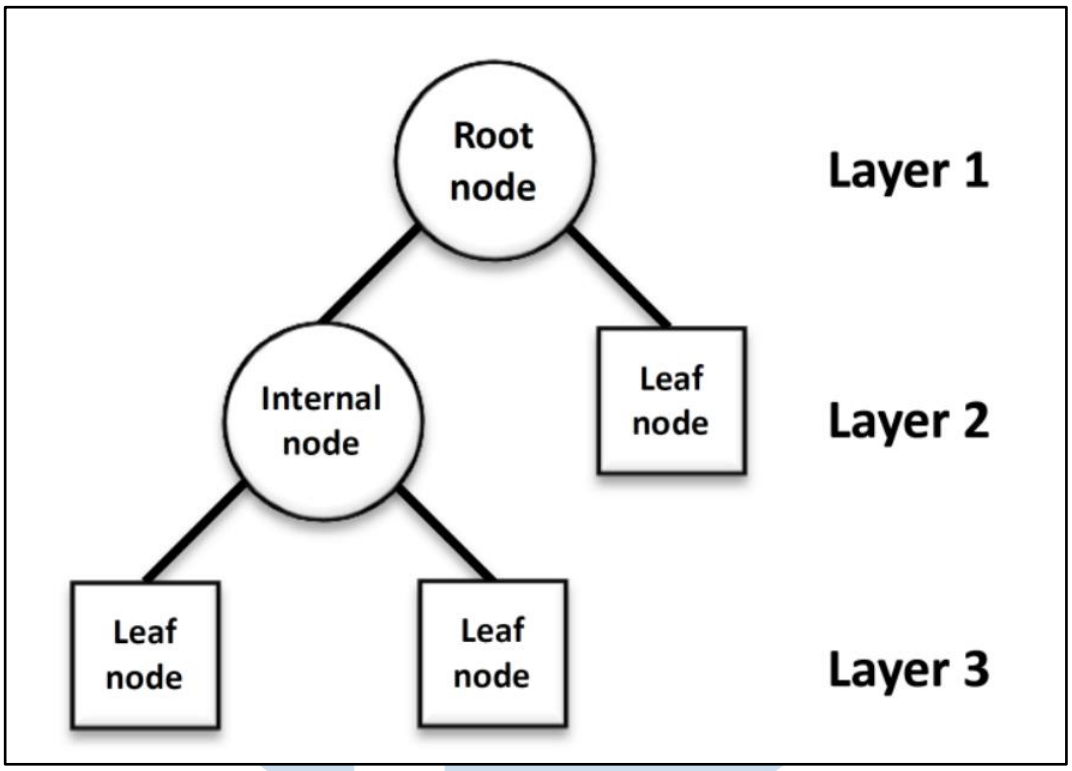
P(H): Probabilitas hipotesis H (*prior probability)*

P(X|H): Probabilitas X berdasarkan kondisi hipotesis H

P(X): Probabilitas dari X

2.4 Decision Tree

Decision tree adalah metode klasifikasi dengan struktur yang menyerupai pohon untuk memodelkan kemungkinan hasil serta konsekuensi. *Decision tree* menyajikan algoritma yang berisi Langkah-langkah pengambilan keputusan sehingga dapat mengarah kepada hasil yang paling menguntungkan. Dengan *control* bersyarat, *node internal* mewakili atribut setiap tahap dan setiap cabang mewakili hasil untuk atribut dan jalur dari daun ke akar mewakili aturan untuk klasifikasi[8].

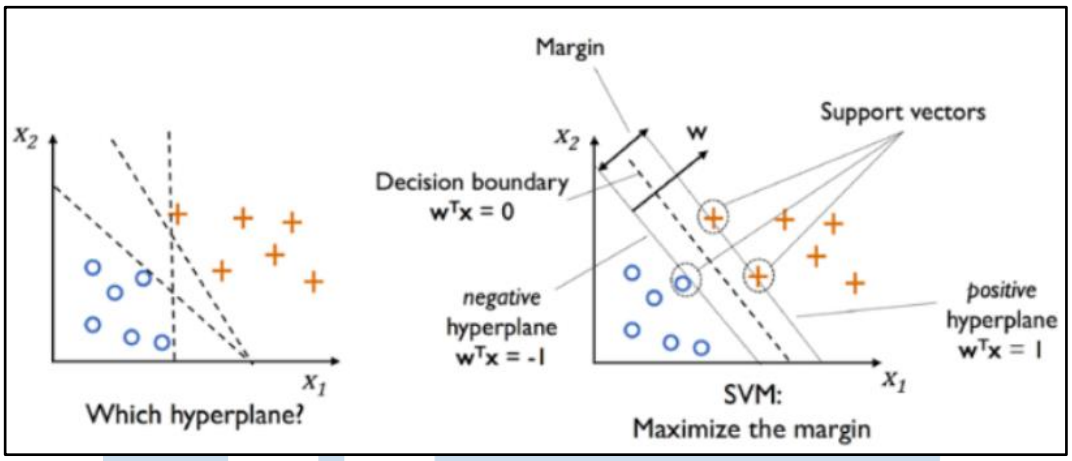


Gambar 2. 2 Struktur Dasar Decision Tree

Disebut *Decision tree* atau pohon keputusan dikarenakan bentuk dari pilihannya bercabang dan membentuk struktur menyerupai pohon. *Decision tree* meningkatkan model prediktif dengan akurasi dan kemudahan dalam interpretasi serta stabilitas. Metode ini salah satu bentuk algoritma pembelajaran terbaik. *Decision tree* juga efektif dalam menyesuaikan hubungan *non-linier* karna mampu melakukan penyesuaian data seperti klasifikasi dan regresi.

2.5 Support Vector Machine

Algoritma *Support Vector Machine* merupakan salah satu algoritma yang termasuk dalam kategori *Supervised Learning*, yang artinya data yang digunakan untuk belajar oleh mesin merupakan data yang telah memiliki label sebelumnya. Sehingga dalam proses penentuan keputusan, mesin akan mengkategorikan data testing ke dalam label yang sesuai dengan karakteristik yang dimiliki nya[9].



Gambar 2. 3 Contoh Hyperlane Pada SVM

Cara kerja dari metode *Support Vector Machine* khususnya pada masalah *non-linear* adalah dengan memasukkan konsep kernel ke dalam ruang berdimensi tinggi. Tujuannya adalah untuk mencari *hyperplane* atau pemisah yang dapat memaksimalkan jarak (*margin*) antar kelas data. Untuk menemukan *hyperplane* terbaik, kita dapat mengukur margin kemudian mencari titik maksimalnya[10]. Proses pencarian *hyperplane* yang terbaik ini adalah ini dari metode *Support Vector Machine ini.*

2.6 Confusion Matrix

Confusion Matrix biasa digunakan sebagai pengukur performa didalam masalah klasifikasi *machine learning* yang mana keluaran dapat berupa dua kelas atau lebih. *Confusion matrix* adalah suatu alat yang berfungsi untuk menganalisa sebuah *classifier* sehingga diketahui apakah *classifier* tersebut baik atau tidak didalam mengenali *tuple* dari tiap kelas yang berbeda[11].

Tabel 2. 1 Tabel Confusion Matrix

	Actual		
Predicted		*Positive "+"*	*Negative "-"*
	Positive (+)	*TP*	*FP*
	Negative (-)	*FN*	*TN*

Dapat dilihat pada tabel 2.1. terdapat 4 istilah sebagai representasi dari hasil proses klasifikasi pada *confusion matrix*, yaitu:

Deskripsi Table:

1) TP (*True Positive*): True Positive: Nilai yang diprediksi dengan benar diprediksi sebagai positif yang sebenarnya.
2) TP (*False Positive*): Nilai yang diprediksi salah memprediksi positif yang sebenarnya. yaitu, Nilai negatif diprediksi sebagai positif.
3) FN (*False Negative*): Nilai positif diprediksi sebagai negatif.
4) TN (*True Negative*): Nilai yang diprediksi dengan benar diprediksi sebagai negatif yang sebenarnya.

Sedangkan untuk melakukan perhitungan akurasi, presisi dan recall dilakukan dengan menggunakan persamaan *confusion matrix* seperti pada tabel 2.2:

Tabel 2. 2 Rumus Confusion Matrix

Accuracy	(TP+TN)/(TP+FP+FN+TN)
Precision	TP/(TP+FP)
Recall	TP/(TP+FN)

Nilai rasio prediksi yang mendekati hasil sebenarnya ditentukan dari nilai accuracy, nilai yang memiliki kesamaan dan ketepatan proses klasifikasi ditunjukan pada nilai precision dan nilai yang menunjukan rasio klasifikasi benar positif dibandingkan dengan keseluruhan dokumen saat klasifikasi adalah recall[12].

2.7 Term Frequency Inverse Document Frequency (TF-IDF)

Metode TF-IDF adalah metode yang biasa digunakan untuk menghitung bobot setiap kata yang sangat umum digunakan pada sistem temu balik informasi atau *information retrieval*. Karena memiliki kelebihan efisien, akurat dan lebih

mudah digunakan, TF-IDF biasa menjadi pilihan untuk memberikan bobot terhadap dokumen.

Metode TF-IDF berfungsi untuk menentukan nilai dari tiap bobot yang terkandung dalam suatu kata, singkatnya untuk mengevaluasi seberapa penting sebuah kata yang terkandung di dalam dokumen. Frekuensi kemunculan dokumen yang mengandung kata tersebut menunjukan seberapa umum kata tersebut. Perhitungan bobot (W) pada algoritma TF-IDF menggunakan rumus di masing-masing dokumen yang dapat dilakukan dengan rumus 2.2[13]:

$$WDT = TFDT * IDFT \quad (2.\ 2)$$

Deskripsi:

1) WDT: Bobot dokumen ked terhadap kata ket.
2) TFDT: Banyaknya kata yang dicari pada sebuah dokumen.
3) IDFT: *Inversed Document Frequency* (log (N/df)).
4) N: Total dokumen.
5) D: Banyak dokumen yang mengandung kata yang dicari.

2.8 Area Under The Curved (AUC)

AUC atau *Area Under Curve* adalah nilai yang berada pada *ROC* (*Receiver Operating Characterictic*). AUC mencangkup area luas yang terdapat dibawah ROC atau merupakan keseluruhan fungsi dari ROC. AUC adalah kurva yang diperoleh diantara sensitivias dan spesifitas pada suatu titik potong.

Nilai pada AUC biasanya adalah 0 hingga 1. Nilai yang dihasilkan inilah yang dapat dijadikan pendukung akan pengukuran yang dilakukan oleh suatu model/metode. Semakin mendekati nilai *AUC* ke angka 1, maka semakin baik model yang digunakan untuk melakukan prediksi[14]. Berikut ini adalah nilai pada *AUC*:

1) *Excellent Classification* apabila nilainya 0.90 hingga 1.00.

2) *Good Classification* apabila nilainya 0.80 hingga 0.90.
3) *Fair Classification* apabila nilainya 0.70 hingga 0.80.
4) *Poor Classification* apabila nilainya 0.60 hingga 0.70.
5) *Failure Classification* apabila nilainya 0.50 hingga 0.60.

2.9 Google Playstore

Google Playstore adalah sebuah aplikasi/program yang berfungsi untuk mendistribusikan berbagai konten dan aplikasi digital yang di perasikan dan dikembangkan oleh Google. Google Playstore juga merupakan toko aplikasi resmi untuk sistem operasi Android. diresmikan pada 22 Oktober 2021, kini Google Playstore adalah toko aplikasi terbesar di dunia[15].

Tidak hanya sebagai toko aplikasi, Google *Playstore* memiliki berbagai macam fitur mulai dari Google *wallet*, streaming film serta review pada tiap aplikasi yang terdaftar di dalamnya, pada fitur *review* inilah para pengguna bisa memberikan penilaian dan kesannya terhadap aplikasi yang telah diinstallnya.

2.10 R-Studio

RStudio merupakan perangkat lunak/*software* yang digunakan untuk mempermudah penulisan menggunakan bahasa R. pada perangkat lunak ini terdapat beberapa fitur yaitu konsol, editor, pendukung eksekusi code, Riwayat, debugging serta manahemen ruang kerja[16].

RStudio tersedia dapat digunakan dan berjalan pada *desktop* (Windows, Linux dan MacOS) atau sistem operasi yang dapat terhubung di RStudio server (Debian, Ubuntu, Linux, CentOS dan SUSE). RStudio tersedia dalam 2 edisi yaitu *open source* dan komersial.

2.11 Python

Python adalah Bahasa pemograman tingkat tinggi yang memiliki banyak fungsi yang interaktif dan beorientasi objek. Python bertujuan untuk membantu

programer dalam merancang program dengan penulisan yang jelasi dan logis di dalam proyek kecil maupun besar. Bahasa Python adalah Bahasa pemograman yang formal disertai aturan-aturan dan formatnya sendiri[17].

2.12 Rapid Miner

RapidMiner adalah perangkat lunak yang biasa digunakan untuk menganalisis hal yang terkait prediksi serta text dan data mining. RapidMiner merupakan *software* yang bersifat *open source*. RapidMiner menggunakan berbagai Teknik deskriptif dan prediktif sehingga dapat membantu pengguna membuat keputusan yang paling baik[18].

Terdapat lebih dari 500 operator data mining yang terdapat pada RapidMiner, diantaranya adalah operator untuk *input, output, data pre-processing* dan visualisasi. RapidMiner ditulis menggunakan Bahasa pemograman java dan dapat berjalan disemua sistem operasi. RapidMiner merupakan *software* untuk menganalisa data dan sebagai mesin *data mining* yang dapat diintegrasikan langsung pada produknya sendiri.

Terdapat juga berbagai manfaat dari RapidMiner yaitu[19]:

1) Sistem operasinya menggunakan Bahasa pemgograman Java.
2) Pemodelannya dimodelkan dalam bentuk *Trees* sehingga mudah dipahami.
3) Representasi XML *internal* untuk memastikan format standar pertukaran data.
4) Multi-layer konsep yang merubah tampilan data menjadi lebih simple.
5) Dilengkapi dengan GUI, Command Line, dan JAVA API yang dapat terintegrasikan dengan aplikasi dan program lainnya.

Terdapat beberapa fitur didalam RapidMiner yaitu[20]:

1) Terdapat berbagai macam algoritma data mining.
2) Memiliki grafis dan model yang mudah dipahami seperti *diagram histogram, Tree Chart, 3D Scatter Plots*.

3) Variasi pluginnya *variative*.
4) Sudah menyediakan prosedur *machine learning* dan *data mining*.
5) *Extraction, Transformation, Loading* (ETL) telah tersedia pada RapidMiner serta *data preprocessing*, *visualisasi*, *modelling* dan evaluasi.
6) Dapat terkoneksi dengan berbagai program *data mining* seperti R.

2.13 Jupyter Notebook

Jupyter adalah Yayasan swasta *non-profit* yang mengembangkan *software* interaktif yang dapat menjembatani berbagai bahasa pemograman. *Notebook* adalah salah satu perangkat lunak persembahan Jupyter, yaitu aplikasi web terbuka/*open-source* yang dapat merancang dan berbagi berbagai dokumen interaktif yang berisi code, *live*, visualisasi dan *text* naratif.

Singkatnya, Jupyter notebook menyatukan kode-kode dari dokumen yang terpisah-pisah kedalam sebuah *library*/aplikasi/proyek (Visual Studio, Eclipse, dan lainnya). Sehingga di dalam dokumen bisa tampilkan cuplikan kode, tampilan hasil, dan visualisasi lainnya dari program yang telah dibuat[21].

2.14 Anaconda Navigator

Anaconda adalah *package* distribusi Bahasa pemograman Python dari perusahaan teknologi *Continuum analytic* yang berisi paket Bahasa pemograman Python yang telah ditambahkan beberapa tambahan paket untuk keperluan pemrograman *data science* dan matematika. Anaconda juga merupakan *software* terbuka yang dapat membantu merancang program hingga teknik code dalam satu distribusi *platform* yang *user friendly*. Terdapat beberapa paket lain seperti, cmd, R-studio, Spyder, dan lain sebagainya[21].

2.15 Penelitian Terdahulu

Setelah membaca dan mempelajari topik tentang analisis sentiment, maka ditemukan beberapa jurnal-jurnal dan penelitian terdahulu yang memiliki kaitan dengan penelitian ini, yaitu:

Tabel 2. 3 Penelitian Terdahulu

Judul Penelitian, Peneliti dan Tahun Publikasi	**Objek Penelitian**	**Teknik Penelitian**	**Hasil Penelitian**
Nico Nathanael Wilim, Raymond Sunardi Oetama, "*Sentiment Analysis about Indonesian Lawyers Club Television Program Using K-Nearest Neighbor, Naïve Bayes Classifier, and Decision Tree"*, IJNMT (*International Journal of New Media Technology*), Vol. 8, No. 1, June 2021, Page 50-56[22], ISSN 2355-0082	Analisis sentimen opini publik terhadap program televisi *Indonesian Lawyers Club*	-Twitter -*Naïve Bayes* -*K-NN* -*Decision Tree* -Analisis Sentimen	-Dengan data tahun 2018, *Naïve Bayes* memiliki tingkat akurasi terbaik sedangkan pada 2019, K-NN yang memiliki tingkat akurasi yang terbaik - Tidak ada algoritma yang menunjukan hasil yang terbaik pada semua data
Muhamad Fani Al-shufi, Adhitia Erfina, "Sentimen Analisis Mengenai Aplikasi Streaming Film Menggunakan Algoritma *Support Vector Machine* Di Playstore", SISMATIK (Seminar Nasional Sistem Informasi dan Manajemen Informatika) Universitas Nusa Putra,)7 Agustus 2021, Page 26-29[23], E-ISSN 2775-6734	Sentimen analisis terhadap layanan Netflix, Iflix, Disney, Wetv, Vidio	-Playstore -*Support Vecor Machine* -Analisis Sentimen	-Tingkat keakurasian dengan *SVM* adalah Iflix 92,67%, Netflix 81,33%, Disney 69,33%, Wetv 64,67% dan Vidio 62%. -Nilai akurasi minimal 58,78% dan maksimal 65,22% -Hasil dari 1000 data (200 masing-masing platform)
Ragil Dimas Himawan, Eliyani,	Sentimen analisis terhadap pemerintah	-Twitter -*SVM*	-*SVM* memiliki akurasi tertinggi

"Perbandingan Akurasi Analisis Sentimen Tweet terhadap Pemerintah Provinsi DKI Jakarta di Masa Pandemi", JEPIN (Jurnal Edukasi dan Penelitian Informatika), April 2021, Page 58-63[24], ISSN(e): 2548-9364	provinsi DKI- Jakarta di masa pandemi Covid19	-*Naïve Bayes* -*Random Forrest*	disbanding algoritma lain -Hasil akurasi *SVM* 77,58%, *Random Forrest* 75,81% dan *Naïve Bayes* 75,22%
Khoirul Abbi Rokhman, Berlilana, Primandani Arsi "Perbandingan Metode *Support Vector Machine* dan *Decision Tree* Untuk Analisis Sentimen *Review* Komentar Pada Aplikasi Transportasi *Online*", JOISM (*Journal Of Information System Management*), 2021, Page 1-7[25], e-ISSN: 2715-3088	Sentimen analisis layanan Gojek pada Google Playstore	-Google Playstore -SVM -*Decision Tree*	-SVM lebih baik dalam tingkat akurasinya -*Decision Tree* mendapat tingkat akurasi 89,80% sedangkan SVM sebesar 90,20%

Pada penelitian sebelumnya di tabel 2.3. terdapat beberapa penelitian yang dijadikan refrensi di dalam penelitian ini. Penelitian dengan judul "*Sentiment Analysis about Indonesian Lawyers Club Television Program Using K-Nearest Neighbor, Naïve Bayes Classifier, and Decision Tree*" menghasilkan *K-NN* yang memiliki tingkat akurasi tertinggi, pada "Perbandingan Akurasi Analisis Sentimen Tweet terhadap Pemerintah Provinsi DKI Jakarta di Masa Pandemi" didapatkan SVM sebagai algoritma terbaik dengan akurasi 90,2% dan pada penelitian "Perbandingan Metode *Support Vector Machine* dan *Decision Tree* Untuk Analisis Sentimen *Review* Komentar Pada Aplikasi Transportasi *Online*" yang mendapatkan hasil tertinggi adalah *SVM* dengan 89.8% disusul dengan *Decision Tree* yakni

89.8%. Penelitian ini berfokus pada perbandingan analisis sentiment diantara 3 algoritma yaitu *Naïve Naïve Bayes, Decision Tree* dan *Support Vector Machine* dengan objek penelitian aplikasi MolaTV pada Google playstore.