



## **ANALISIS SENTIMEN MASYARAKAT TERHADAP TRANSPORTASI TRANSJAKARTA MENGGUNAKAN MAXIMUM ENTROPY, NAIVE BAYES DAN SUPPORT VECTOR MACHINE**

**IMADUDDIN ABDURRAHMAN**



**PROGRAM SARJANA ILMU KOMPUTER  
SEKOLAH SAINS DATA MATEMATIKA DAN INFORMATIKA  
INSTITUT PERTANIAN BOGOR  
BOGOR  
2025**



## **PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA**

Dengan ini saya menyatakan bahwa skripsi dengan judul “Analisis Sentimen Masyarakat terhadap Transportasi Transjakarta Menggunakan Maximum Entropy, Naive Bayes dan Support Vector Machine” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Juli 2025

Imaduddin Abdurrahman  
G64190023

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
  - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :  
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah  
b. Pengutipan tidak merugikan kepentingan wajar IPB University.

IMADUDDIN ABDURRAHMAN. Analisis Sentimen Masyarakat Terhadap Transportasi Transjakarta Menggunakan Maximum Entropy, Naive Bayes dan Support Vector Machine. Dibimbing oleh MEDRIA KUSUMA DEWI HARDHIENATA dan HARI AGUNG ANDRIANTO.

Moda transportasi Transjakarta mengalami peningkatan penggunaan secara signifikan pada awal tahun 2023. Peningkatan ini menyebabkan opini yang beredar mengenai Transjakarta di masyarakat bermunculan. Opini masyarakat terhadap moda transportasi Transjakarta dapat diteliti dengan analisis sentimen. Penelitian bertujuan untuk melakukan analisis sentimen opini masyarakat terhadap Transportasi Transjakarta menggunakan pendekatan machine learning. Data yang digunakan adalah tweet media sosial X dengan kata kunci "Transjakarta" pada rentang 1 Januari 2023 - 14 Juni 2023 sebanyak 238.734 tweet. Tahapan penelitian terdiri dari pengambilan data, praproses data, pembobotan kata, *resampling data*, pembagian data, klasifikasi menggunakan algoritma *machine learning* serta evaluasi kinerja algoritma. Dalam penelitian ini akan diuji tiga algoritma *machine learning*, yaitu algoritma *maximum entropy*, *naive bayes* dan *support vector machine (SVM)* untuk menganalisis sentimen masyarakat terhadap moda transportasi TransJakarta. Sentimen dibagi menjadi positif, netral dan negatif yang sudah diberi label menggunakan mesin dan diperiksa secara manual. Hasil simulasi menunjukkan bahwa model terbaik diperoleh dengan metode *support vector machine* dengan *precision* sebesar 72%, *recall* sebesar 71%, *F1-score* sebesar 71% dan akurasi sebesar 71%.

Kata Kunci: Analisis Sentimen, *Machine Learning*, *Maximum Entropy*, *Naive Bayes*, *Support Vector Machine (SVM)*, Transjakarta.

## ABSTRACT

IMADUDDIN ABDURRAHMAN. Analysis of Public Sentiment on Transjakarta Transportation Using Maximum Entropy, Naïve Bayes and Support Vector Machine. Supervised by MEDRIA KUSUMA DEWI HARDHIENATA and HARI AGUNG ANDRIANTO.

*Transjakarta's transportation mode experienced a significant increase in usage in early 2023. This increase caused opinions circulating about Transjakarta in the community to emerge. Public opinion on Transjakarta's mode of transportation can be examined by sentiment analysis. This study aims to analyze public opinion sentiment towards Transjakarta Transportation using a machine learning approach. The data used was 238,734 tweets from social media X with the keyword "Transjakarta" in the range of January 1, 2023 - June 14, 2023. The research stages consist of data collection, data preprocessing, word weighting, data resampling, data sharing, classification using machine learning algorithms and algorithm performance evaluation. In this study, three machine learning algorithms will be tested, namely maximum entropy, naive bayes and support vector machine (SVM) algorithms to analyze public sentiment towards TransJakarta*



*transportation modes. Sentiment is divided into positive, neutral and negative which have been labeled using a machine and checked manually. The simulation results showed that the best model was obtained by the support vector machine method with a precision of 72%, recall of 71%, F1-score of 71% and accuracy of 71%.*

**Keywords:** *Sentiment Analysis, Machine Learning, Maximum Entropy, Naive Bayes, Support Vector Machine (SVM), Transjakarta.*

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
- b. Pengutipan tidak merugikan kepentingan yang wajar IPB University

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



Hak Cipta Dilindungi Undang-undang  
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :  
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah  
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.  
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2025  
Hak Cipta dilindungi Undang-Undang

*Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.*

*Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.*



**PROGRAM STUDI ILMU KOMPUTER  
SEKOLAH SAINS DATA MATEMATIKA DAN INFORMATIKA  
INSTITUT PERTANIAN BOGOR  
BOGOR  
2025**

Hak Cipta Dilindungi Undang-undang  
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :  
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah  
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.  
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
- b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Tim Pengaji pada Ujian Skripsi:  
Dr. Sony Hartono Wijaya, S.Kom., M.Kom.



Judul Skripsi : Analisis Sentimen Masyarakat terhadap Transportasi Transjakarta Menggunakan Maximum Entropy, Naive Bayes dan Support Vector Machine

Nama : Imaduddin Abdurrahman  
NIM : G64190023

Pembimbing 1:  
Medria Kusuma Dewi Hardhienata, S.Komp., Ph.D

Disetujui oleh



Pembimbing 2:  
Hari Agung Adrianto, S.Kom, M.Si., Ph.D.



Diketahui oleh

Ketua Program Sarjana Ilmu Komputer

Dr. Sony Hartono Wijaya, S.Kom., M.Kom.  
19810809 200812 1 002



Tanggal Ujian:  
21 Juli 2025

Tanggal Lulus:  
(tanggal penandatanganan oleh Dekan  
Fakultas/Sekolah ...)



- Hak Cipta Dilindungi Undang-undang  
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :  
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah  
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University  
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

## PRAKATA

Rasa syukur paling sempurna penulis panjatkan kepada Allah subhanaahu wa ta'ala, Tuhan yang Maha Esa karena dengan segala karunia yang diberikan oleh-Nya, karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian yang dilaksanakan sejak bulan Januari 2024 sampai bulan Juli 2025 ini ialah Analisis Sentimen, dengan judul “Analisis Sentimen Masyarakat terhadap Transportasi Transjakarta Menggunakan Maximum Entropy, Naive Bayes dan Support Vector Machine”.

Terima kasih banyak penulis ucapkan kepada para pembimbing yang sangat berjasa, membimbing dengan sabar dan banyak sekali memberi saran, Ibu Medria Kusuma Dewi Hardhienata, S.Komp., Ph.D serta Bapak Hari Agung Adrianto, S.Kom, M.Si., Ph.D. Ucapan terima kasih juga saya sampaikan kepada pembimbing akademik, moderator seminar, dan penguji luar komisi pembimbing. Di samping itu, penghargaan penulis sampaikan kepada Pihak Ivosight yang telah membantu memberikan data sebagai bahan penelitian. Ungkapan terima kasih juga disampaikan kepada Abi, Umi, seluruh keluarga dan juga teman-teman semua yang tak henti-hentinya memberikan dukungan, doa, dan kasih sayangnya sehingga karya ilmiah ini bisa diselesaikan.

Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Juli 2025

*Imaduddin Abdurrahman*



## DAFTAR TABEL

## DAFTAR GAMBAR

	PENDAHULUAN	xii
	1.1 Latar Belakang	1
	1.2 Rumusan Masalah	3
	1.3 Tujuan	3
	1.4 Manfaat	3
	1.5 Ruang Lingkup	4
II	TINJAUAN PUSTAKA	5
	2.1 Transjakarta	5
	2.2 Analisis Sentimen	5
	2.3 <i>Maximum Entropy</i>	6
	2.4 <i>Naive Bayes</i>	6
	2.5 <i>Support Vector Machine</i>	7
	2.6 <i>Data Cleansing</i>	7
	2.7 <i>Casefolding</i>	8
	2.8 <i>Stopwords Removal</i>	8
	2.9 Normalisasi	8
	2.10 Tokenisasi	8
	2.11 <i>Stemming</i>	9
	2.12 <i>Term Frequency-Inverse Document Frequency (TF-IDF)</i>	9
	2.13 <i>K-Fold Cross Validation</i>	9
	2.14 <i>Random Under Sampling</i>	10
	2.15 <i>Confusion Matrix</i>	10
III	METODE	12
	3.1 Lingkungan Pengembangan	12
	3.2 Tahapan Penelitian	12
	3.3 Pengambilan Data	12
	3.4 Pelabelan Data	13
	3.5 Praproses Data	13
	3.6 <i>Data Cleansing</i>	14
	3.7 <i>Casefolding</i>	14
	3.8 Normalisasi	14
	3.9 Tokenisasi	14
	3.10 <i>Stopwords Removal</i>	14
	3.11 <i>Stemming</i>	14
	3.12 Pembobotan Kata	15
	3.13 Resampling Data	15
	3.14 Pembagian Data	15
	3.15 Analisis Sentimen	15

Hak Cipta Dilindungi Undang-undang  
 1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
- b. Pengutipan tidak merugikan kepentingan yang wajar IPB University

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



3.16	Evaluasi	17
<b>IV</b>	<b>HASIL DAN PEMBAHASAN</b>	18
4.1	Pengambilan Data	18
4.2	Pelabelan Data	18
4.3	Praproses Data	18
4.4	Pembobotan Kata	20
4.5	Resampling Data	20
4.6	Pembagian Data	21
4.7	Analisis Sentimen	21
4.8	Evaluasi	23
<b>V</b>	<b>SIMPULAN DAN SARAN</b>	27
5.1	Simpulan	27
5.2	Saran	28
<b>DAFTAR PUSTAKA</b>		29
<b>LAMPIRAN</b>		32
<b>RIWAYAT HIDUP</b>		64

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah

b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



1	<i>Confusion matrix</i>	10
2	Atribut data tweet	13
3	Model <i>confusion matrix</i>	17
4	Tahapan <i>data cleansing</i> para praproses data	18
5	Tahapan <i>casefolding</i> para praproses data	19
6	Tahapan normalisasi para praproses data	19
7	Tahapan tokenisasi para praproses data	19
8	Tahapan <i>stopwords removal</i> para praproses data	20
9	Contoh tahapan praproses data	20
10	Rata-rata <i>confusion matrix maximum entropy</i>	24
11	Rata-rata hasil klasifikasi <i>maximum entropy</i>	24
12	Rata-rata <i>confusion matrix naive bayes</i>	24
13	Rata-rata hasil klasifikasi <i>naive bayes</i>	24
14	Rata-rata <i>confusion matrix support vector machine</i>	25
15	Rata-rata hasil klasifikasi <i>support vector machine</i>	25
16	Perbandingan hasil klasifikasi tiga algoritma <i>machine learning</i>	25

## DAFTAR GAMBAR

1	Ilustrasi tokenisasi	8
2	Ilustrasi <i>k-fold cross validation</i>	10
3	Tahapan penelitian	12
4	<i>Wordcloud</i> sentimen kelas positif dan sentimen kelas negatif	21

Hak Cipta Dilindungi Undang-undang  
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah  
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

# I PENDAHULUAN

## 1.1 Latar Belakang

Transjakarta merupakan moda transportasi yang banyak digunakan oleh masyarakat Jakarta dan sekitarnya. Data Badan Pusat Statistik (BPS) Provinsi DKI Jakarta menyebutkan jumlah penumpang TransJakarta mencapai 180 Juta orang sepanjang Januari-Juni 2024. Data dari BPS Provinsi DKI Jakarta (2024) ini menunjukkan terjadi peningkatan jumlah penumpang sebesar 48,1% persen dibanding dengan jumlah penumpang berjumlah 121 juta orang sepanjang Januari-Juni 2023. Jumlah penumpang yang besar ini membuat Transjakarta menjadi moda transportasi yang paling banyak digunakan dibandingkan MRT (*mass rapid transit*) yang berjumlah 17,86 juta orang dan LRT (*light rail transit*) yang berjumlah 559,40 ribu orang pada rentang waktu Januari-Juni 2024. Peningkatan ini juga didukung dengan banyaknya jumlah bus yang beroperasi tercatat sejumlah 4.487 unit bus pada Juni 2024.

Berdasarkan data dari BPS Provinsi DKI Jakarta, Transjakarta merupakan moda transportasi publik yang paling banyak digunakan oleh masyarakat. Beberapa penelitian-penelitian sebelumnya juga menyebutkan bahwa Transjakarta dipilih karena berbagai macam pertimbangan. Transjakarta dari segi harga cukup terjangkau untuk semua segmen masyarakat (Rahadiano *et al.* 2019), terintegrasi baik dengan layanan moda transportasi lain (mikrotrans) (Dharmawan 2022), serta kualitas produk dan pelayanan yang baik (Prabantari 2020). Menurut Sahara dan Ferdiansyah (2023) penggunaan Transjakarta juga membuat pelanggan juga merasa aman di halte bus tersebut karena adanya petugas keamanan serta cctv dalam halte. Namun, beberapa halte Transjakarta memiliki kondisi yang perlu ditingkatkan karena pengelolaan yang belum merata (Pristanto *et al.* 2023).

Setiap tahun, semakin banyak masyarakat Indonesia yang mengakses internet. Masyarakat yang menggunakan moda transportasi Transjakarta yang tinggal di daerah perkotaan juga merupakan pengguna aktif internet. Hasil pendataan Survei Susenas BPS (2023) menunjukkan bahwa 66,48 persen penduduk Indonesia telah mengakses internet di tahun 2022. Penduduk Indonesia yang mengakses Internet untuk tujuan media sosial sebesar 73,94 % dari total populasi. We Are Social (2024) melaporkan bahwa terdapat 24,69 juta pengguna aktif media sosial X (sebelumnya twitter) di Indonesia. Dengan data tersebut, Indonesia menempati urutan ke-5 dunia dari negara-negara pengguna media sosial X.

Menurut data BPS DKI Jakarta (2024) persentase pengguna internet di DKI Jakarta sebesar 87,84%. BPS DKI Jakarta (2024) juga menyatakan dari persentase tersebut sebanyak 80,56% pengguna menggunakan media sosial ketika mengakses internet. Masyarakat yang menjadi pengguna transportasi Transjakarta dan aktif mengakses internet juga merupakan pengguna aktif media sosial. Hal ini dikarenakan informasi dan konten mengenai Transjakarta aktif diperbarui di media sosial (Alghoniyyu *et al.* 2025). Transjakarta juga memiliki akun media sosial X bernama @PT\_Transjakarta yang aktif hingga saat ini. Pengguna media sosial X melakukan interaksi dengan akun media sosial X Transjakarta. Interaksi antar pengguna ini membangun opini publik di media sosial X. Hal ini sesuai dengan penelitian oleh Qadri (2020) menjelaskan bahwa media sosial dapat dijadikan

sebagai penghubung antara komunikator politik dengan publik sehingga media sosial memberikan pengaruh dan dapat berperan membentuk opini publik.

Pengguna moda transjakarta yang banyak serta diiringi dengan pengguna media sosial yang tumbuh signifikan di Indonesia membuat transjakarta menjadi isu yang telah banyak dibahas oleh penelitian-penelitian pendahulu. Hal ini didukung dengan akun media sosial transjakarta yang aktif sehingga memunculkan berbagai sentimen mengenai transjakarta di media sosial X. Oleh karena itu, sentimen yang berkembang mengenai transjakarta di media sosial X merupakan hal yang bisa diteliti. Liu (2012) mengatakan di atas tahun 2000-an terjadi peningkatan penelitian mengenai *linguistic and natural language processing* (NLP). Alasan pertama adalah luasnya cakupan aplikasi mengenai bidang tersebut, hampir di setiap lini industri. Alasan kedua yang dikemukakan karena terdapat banyak sekali data opini yang tersedia dikarenakan media sosial di Internet sehingga penelitian mengenai *natural language processing* (NLP) menjadi semakin berkembang. Salah satu metode berbasis NLP yang dapat digunakan untuk mengetahui sentimen masyarakat terhadap topik di media sosial adalah analisis sentimen.

Dang *et al.* (2020) menjelaskan bahwa ada tiga pendekatan dalam melakukan analisis sentimen, yaitu *lexicon-based techniques*, *machine learning based techniques*, dan *hybrid approaches*. Beberapa penelitian pendahulu melakukan penelitian analisis sentimen menggunakan metode *machine learning*. Zhang dan Zheng (2016) melakukan penelitian yang menggunakan *machine leaning* untuk melakukan penelitian berbasis NLP dengan melakukan perbandingan analisis sentimen teks dengan menggunakan *support vector machine* (SVM) dan *extreme learning machine* (ELM). Samuels dan Mcgonical (2007) melakukan penelitian analisis sentimen di media sosial menggunakan metode *machine learning*. Metode yang digunakan adalah *support vector machine* dan *naive bayes*.

Penelitian analisis sentimen menggunakan metode *machine learning* terhadap media sosial X juga dilakukan sebelumnya. Han *et al.* (2020) melakukan penelitian menggunakan *support vector machine* dengan dataset dari media sosial X. Cindo *et al.* (2020) juga melakukan penelitian menggunakan *maximum entropy* dan *support vector machine* data dari media sosial X. Muzaki dan Witanti (2021) melakukan penelitian menggunakan *naive bayes clasifier* dengan topik Pilkada 2020 ketika pandemi covid-19. Novantirani *et al.* (2015) melakukan penelitian analisis sentimen pada twitter untuk penggunaan transportasi umum darat dalam kota menggunakan *support vector machine* (SVM). Hasil penelitian Novantirani *et al.* (2015) menunjukkan bahwa dengan menggunakan metode SVM diperoleh nilai akurasi sebesar 78,12%. Penelitian lain yang dilakukan oleh Chairunnisa *et al.* (2022) tentang analisis sentimen pengguna twitter terhadap vaksinasi Covid-19 di Indonesia menunjukkan bahwa analisis sentimen mampu mengklasifikasi sentimen media sosial menggunakan algoritma SVM sehingga diperoleh hasil akurasi sebesar 90%. Penelitian lainnya menggunakan algoritma *naive bayes* yang dilakukan oleh Muzaki dan Witanti (2021) mengenai pilkada serentak ketika covid-19 di media sosial X menghasilkan akurasi yang baik sebesar 92,2%. Penelitian lainnya menggunakan algoritma *maximum entropy* yang dilakukan oleh Cindo *et al.* (2020) melakukan analisis sentimen di media sosial X menghasilkan akurasi sebesar 92,6%. Penelitian lain yang lebih relevan dengan penelitian ini dilakukan oleh Go *et al.* (2009) melakukan analisis sentimen terhadap media sosial X dengan membandingkan tiga algoritma yaitu *maximum entropy*, *naive bayes* dan *support*

*vector machine*. Hasil yang didapatkan secara berurutan sebesar 80.5%, 81.3% dan 82.2%. Berbagai penelitian sebelumnya menunjukkan ketiga algoritma tersebut mendapatkan hasil yang cukup baik.

Meski demikian, masih belum banyak dilakukan penelitian analisis sentimen menggunakan algoritma *machine learning* dengan menggunakan data Transjakarta dari media sosial X. Nurlaela dan William (2023) melakukan penelitian analisis sentimen terhadap transjakarta dari twitter menggunakan *software WEKA* yang menerapkan algoritma *naive bayes* dengan jumlah data setelah praproses sebanyak 4.027 *tweets*. Penelitian ini berfokus kepada klasifikasi sentimen yang dibagi menjadi 17 variabel dengan sentimen negatif berjumlah 1.886, sentimen netral berjumlah 2.006 dan sentimen positif berjumlah 180. Aplikasi WEKA yang digunakan memberikan akurasi sebesar 86.68%. Penelitian lain yang dilakukan Iwandini *et al.* (2023) menggunakan *naive bayes* dan *k-nearest neighbour* sebagai metode dalam melakukan analisis sentimen terhadap pengguna transjakarta di media sosial twitter dengan jumlah data sebanyak 4.000 *tweets*. Hasil penelitian ini memberikan akurasi sebesar 61,1% untuk metode *naive bayes* dan 75,7% untuk metode *k-nearest neighbour*.

Meskipun terdapat beberapa penelitian terkait yang telah melakukan analisis terhadap data Transjakarta, namun data yang digunakan relatif terbatas dan belum menggunakan data dari tahun 2023. Berdasarkan hal tersebut, penelitian ini akan menguji kinerja tiga algoritma *machine learning* yaitu *maximum entropy*, *naive bayes*, dan *support vector machine* untuk menganalisis data sentimen masyarakat yang menggunakan Transjakarta dengan data dari media sosial X sebanyak 238.734 *tweets*. Data ini didapatkan dari permohonan *scrapping* data media sosial X kepada perusahaan bernama Ivosights dengan kata kunci "transjakarta". Hasil dari penelitian ini akan memberikan perbandingan dari ketiga algoritma yang digunakan.

## 1.2 Rumusan Masalah

Rumusan Masalah penelitian ini adalah sebagai berikut:

1. Bagaimana menganalisis sentimen publik menggunakan data media sosial X terhadap Transjakarta dengan data yang cukup besar?
2. Algoritma apa yang paling optimal digunakan untuk melakukan klasifikasi sentimen masyarakat terkait penggunaan moda transjakarta dengan jumlah data yang cukup besar?

## 1.3 Tujuan

Tujuan penelitian ini melakukan analisis sentimen publik terhadap transjakarta berbasis data media sosial X dengan menggunakan algoritma *maximum entropy*, *naive bayes* dan *support vector machine* kemudian membandingkan ketiga hasilnya.

## 1.4 Manfaat

Manfaat penelitian ini untuk mengetahui pendapat masyarakat terhadap moda transportasi Transjakarta. Manfaat lainnya adalah mendapatkan pilihan algoritma yang terbaik dari tiga algoritma yang digunakan dalam melakukan analisis sentimen



publik terhadap Transjakarta. Hasil dari penelitian ini memberikan gambaran proses analisis sentimen ketika menganalisis sentimen Transjakarta dalam jumlah data yang cukup besar sebanyak 238.734 data tweet.

## 1.5 Ruang Lingkup

Ruang lingkup penelitian ini menggunakan data media sosial X yang diberikan oleh perusahaan Ivosight dengan kata kunci “Transjakarta” berjumlah 238.734 data dari tanggal 1 Januari 2023 - 14 Juni 2023.

## II TINJAUAN PUSTAKA

### 2.1 Transjakarta

Menurut situs Transjakarta.co.id, Transjakarta merupakan moda transportasi publik berbentuk bus yang berada di wilayah Jakarta dan sekitarnya. Transjakarta merupakan sistem transportasi Bus Rapid Transit (BRT) pertama di Asia Tenggara dan Selatan yang beroperasi pertama kali pada 1 Februari 2004. Kebijakan moda transportasi Transjakarta pertama kali ditetapkan dalam bentuk Badan Pengelola Transjakarta dalam Keputusan Gubernur No.110/2003. Optimalisasi implementasi kebijakan Transjakarta terus dilakukan sampai saat ini, mulai dari penambahan armada, penambahan rute, penyesuaian metode pembayaran, dan hal lainnya. Tanggal 27 Maret 2014 Transjakarta berubah status menjadi Badan Usaha Milik Daerah (BUMD) dan resmi berganti nama menjadi PT. Transportasi Jakarta.

Transjakarta merupakan moda transportasi umum yang memiliki jumlah penumpang paling tinggi dibanding moda transportasi umum lainnya. Data Badan Pusat Statistik (BPS) Provinsi DKI Jakarta menyebutkan jumlah penumpang TransJakarta mencapai 180 Juta orang sepanjang Januari - Juni 2024. Tingginya jumlah orang yang menggunakan moda transportasi Transjakarta dikarenakan beberapa alasan dari penelitian-penelitian yang sebelumnya dilakukan. Transjakarta dari segi harga cukup terjangkau untuk semua segmen masyarakat (Rahadiano *et al.* 2019), terintegrasi baik dengan layanan moda transportasi lain (mikrotrans) (Dharmawan 2022), serta kualitas produk dan pelayanan yang baik (Prabantari 2020).

### 2.2 Analisis Sentimen

Analisis sentimen merupakan suatu metode untuk menganalisis kumpulan sentimen atau opini publik mengenai isu tertentu. Secara definisi, analisis sentimen merupakan analisis yang menggunakan pemrosesan bahasa alami atau *natural language processing* (NLP), analisis teks serta teknik komputasi untuk mengotomatisir ekstraksi dan klasifikasi sentimen berdasarkan tinjauan sentimen (Hussein 2018). Analisis sentimen juga dapat diartikan sebagai teknik pengolahan bahasa alami yang digunakan untuk mengidentifikasi dan mengklasifikasikan sentimen dalam teks atau data lainnya. Tujuan dari analisis sentimen adalah untuk memahami opini, pandangan, atau perasaan pengguna terhadap suatu topik atau produk tertentu (Pang dan Lee 2008).

Analisis sentimen memiliki tiga pendekatan, yaitu yaitu *lexicon-based techniques*, *machine learning based* (Dang *et al* 2020). Dari tiga pendekatan tersebut, pengaplikasian analisis sentimen digunakan dalam berbagai hal, diantaranya pemasaran, politik, keamanan siber, layanan kesehatan, serta analisis sosial (Liu 2012). Beberapa pengaplikasian lain dari analisis sentimen adalah mendapatkan *customer insights*, manajemen produk atau merek serta pengambilan keputusan dan strategi pengembangan (Aftab *et al.* 2023). Media sosial menjadi salah satu topik yang digunakan dalam beberapa penelitian analisis sentimen. Go *et al.* (2009) dan Chairunnisa *et al.* (2022) melakukan penelitian analisis sentimen dengan data yang bersumber dari media sosial X. Penelitian analisis sentimen yang menggunakan topik Transjakarta dilakukan oleh Iwandini *et al.* (2023) menggunakan *naive bayes* dan *k-nearest neighbour* sebagai metode dalam

melakukan analisis sentimen terhadap pengguna transjakarta di media sosial twitter dengan jumlah data sebanyak 4.000 tweets. Hasil penelitian ini memberikan akurasi sebesar 61,1% untuk metode *naive bayes* dan 75,7% untuk metode *k-nearest neighbour*.

### 2.3 Maximum Entropy

Algoritma MaxEnt didasarkan pada prinsip entropi maksimum untuk mengevaluasi parameter untuk setiap fitur (Xie *et al* 2017). Konsep mengenai algoritma *maximum entropy* adalah pemilihan model yang paling memenuhi suatu batasan tertentu. Algoritma MaxEnt adalah sebuah model berbasis fitur, Dalam skenario dua kelas, hal ini sama dengan menggunakan regresi logistik untuk menemukan distribusi atas kelas-kelas tersebut (Go *et al* 2009). Metode tersebut bertujuan untuk memaksimalkan entropi dalam sistem dengan memprediksi distribusi kondisi dari label dalam setiap kelas (Cindo *et al* 2020).

Persamaan dari MaxEnt dapat didefinisikan berikut (Go *et al* 2009):

$$(c|d) = \frac{(\sum_i (\lambda_i f_i(c,d)))}{\sum_i' (\sum_j (\lambda_j f_j(c',d)))}$$

- $P_{ME}(c|d, \lambda)$  : Probabilitas kondisional yaitu probabilitas dari sebuah data  $d$  termasuk dalam memiliki kelas  $c$  dengan parameter/bobot  $\lambda$
- $\lambda_i$  adalah parameter/bobot yang dipelajari untuk setiap fungsi fitur
- $f_i(c,d)$  adalah fungsi fitur yang menentukan kecocokan antara kelas  $c$  dan data  $d$ .
- $\sum_{c'}$  adalah penjumlahan atas semua kemungkinan kelas  $c'$
- $c$  adalah kelas atau label target
- $d$  adalah data input, biasanya berbentuk vektor fitur

### 2.4 Naive Bayes

*Naive bayes* adalah algoritma yang menggunakan aturan bayes dengan asumsi kuat bahwa atributnya bersyarat independen jika diberikan kelas. *Naive bayes* menyediakan sebuah mekanisme untuk menggunakan informasi dari suatu data untuk mengestimasi “*posterior probability*” ( $P(y|x)$ ) dari setiap kelas  $y$ , jika diberikan objek  $x$  (Sammut dan Webb 2010). Naive bayes digunakan secara luas dalam *machine learning* karena efisiensinya serta kemampuannya untuk menggabungkan sumber dari fitur dalam jumlah besar (Manning dan Schutze 1999). *Naive bayes* adalah sebuah model sederhana yang bekerja secara baik untuk melakukan kategorisasi teks.

Persamaan naive bayes dapat dituliskan sebagai berikut (Go *et al* 2009):

$$(c|d) := \frac{(c|d)\prod_i (f_i(d))}{(c)}$$

- $(c|d)$  disebut juga probabilitas posterior, adalah probabilitas  $d$  termasuk dalam kelas  $c$  (klasifikasi menggunakan Naive Bayes).

- $P(c)$  disebut juga probabilitas prior, adalah probabilitas awal dari kelas  $c$ , yang merupakan probabilitas yang dihitung sebelum melihat data (berdasarkan frekuensinya)
- $P(f_i|c)$  adalah probabilitas kondisional, yaitu dari fitur  $f_i$  diberikan kelas  $c$ , yang menggambarkan seberapa sering fitur  $f_i$  muncul dalam kelas  $c$ .
- $(\cdot)$  adalah frekuensi fitur, yaitu jumlah kemunculan fitur  $f_i$  dalam data/dokumen  $d$ .
- $c$  adalah kelas atau kategori yang menjadi target prediksi
- $d$  adalah dokumen atau satu unit data yang dianalisis
- $f_i$  adalah fitur ke- $i$ , yaitu sebuah ciri yang diekstrak dari dokumen  $d$
- $P(d)$  adalah probabilitas dari data  $d$ , yang biasanya digunakan sebagai faktor normalisasi untuk memastikan bahwa probabilitas kelas yang dihasilkan adalah distribusi probabilitas yang valid.

## 2.5 Support Vector Machine

*Support vector machine* adalah algoritma *machine learning* yang digunakan untuk membangun model prediktif dengan membagi data ke dalam kelas-kelas tertentu dengan memanfaatkan sebuah *hyperplane* (bidang pemisah) yang memaksimalkan margin (jarak) antara kelas-kelas tersebut (Abe 2010). *Support vector machine* merupakan algoritma yang cukup populer digunakan untuk melakukan analisis sentimen. Konsep dasar dari *support vector machine* (SVM) adalah menyelesaikan *hyperplane* pemisah yang dapat membagi set data pelatihan yang benar dan memiliki interval geometris terbesar (Chairunnisa *et al.* 2022). *Hyperplane* yang dibangun digunakan untuk melakukan klasifikasi opini analisis sentimen.

Persamaan *support vector machine* didapatkan dari mengembalikan nilai dua kelas permasalahan klasifikasi menggunakan model linear sehingga membentuk persamaan:

$$(y) = \Phi(x) + b$$

- $x$  adalah vektor input, yaitu representasi numerik dari satu titik data
- $y(x)$  adalah output keputusan, yaitu skor yang menentukan input  $x$
- $w$  adalah vektor bobot yang menentukan orientasi bidang pemisah (*hyperplane*)
- $\Phi(x)$  adalah fungsi kernel, yang memetakan vektor  $x$  ke ruang dimensi yang lebih tinggi.
- $b$  adalah Bias, yaitu nilai yang menggeser posisi bidang pemisah.

## 2.6 Data Cleansing

*Data Cleansing* atau juga disebut *data cleaning* atau *data scrubbing* adalah sebuah praproses data untuk membersihkan data agar data lebih berkualitas sebelum diolah lebih lanjut. Data cleansing mencakup perbaikan data-data yang buruk, menyaring beberapa data yang keliru dari dataset dan juga mengurangi atribut data yang tidak perlu (Garcia *et al.* 2014). Pembersihan data dilakukan untuk memenuhi persyaratan pemodelan yaitu melakukan segmentasi dengan menghapus beberapa simbol dan tanda baca (Diya dan Yixi 2019). Pembersihan data ini dilakukan dengan menghapus simbol maupun angka (*data outlier*) yang

memperburuk kualitas data. Hal hal tersebut dapat berupa angka, link *url*, emotikon, *hashtag*, dan *mention*.

## 2.7 Casefolding

*Casefolding* merupakan salah satu tahapan praproses data. *casefolding* merupakan salah satu bentuk normalisasi bahasa yang bertujuan untuk menyeragamkan seluruh bentuk kata agar memudahkan program menandai kata. Cara kerja *casefolding* yaitu mengubah kata dari *Capital Each Word* atau *Upper* atau huruf besar menjadi *lowercase* atau huruf kecil semua (Luqyana *et al.* 2018).

## 2.8 Stopwords Removal

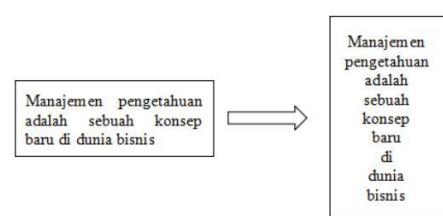
*Stopwords* adalah beberapa kata yang sangat umum dan kurang bermakna dalam suatu teks (Manning *et al.* 2009). Penghapusan ini bertujuan untuk mengurangi jumlah kata yang disimpan oleh sistem (Manning *et al.* 2009). Pada tahap ini, dilakukan penghapusan kata-kata yang sering muncul namun tidak penting atau tidak memberikan arti penting terhadap proses klasifikasi (Patel dan Shah 2013). Contoh kata-kata tersebut ialah kata depan seperti “yang”, “dan”, “di” dan lain-lain.

## 2.9 Normalisasi

Normalisasi adalah sebuah tahapan praproses data untuk menormalkan kaidah bahasa menjadi bentuk semula. Penormalan ini dilakukan sesuai dengan standardisasi yang berlaku yaitu sesuai dengan kamus bahasa. Normalisasi teks juga dapat diartikan sebagai konversi kata menjadi bentuk standar dan lebih sesuai (Jufrasky dan Martin 2025). Contoh penerapan normalisasi adalah mengubah singkatan kata menjadi bentuk umumnya, atau penyesuaian huruf kecil dan kapital.

## 2.10 Tokenisasi

Tokenisasi adalah menghilangkan tanda baca dan memisahkan antara kata satu dengan kata lainnya. Saat melakukan tokenisasi dilakukan proses pemisahan suatu rangkaian karakter berdasarkan karakter spasi, dan mungkin pada waktu yang bersamaan dilakukan juga proses penghapusan karakter tertentu, seperti tanda baca (Amin 2013). Token seringkali disebut sebagai istilah (*term*) atau kata. Sebuah token merupakan suatu urutan karakter dari dokumen tertentu yang dikelompokkan dan berguna untuk diproses. Kata-kata “computer”, “computing”, dan “compute” semua berasal dari *term* yang sama yaitu “comput”, tanpa pengetahuan sebelumnya dari morfologi bahasa Inggris. Pada Gambar 1 diilustrasikan proses tokenisasi.



Gambar 1 Ilustrasi tokenisasi

## 2.11 *Stemming*

*Stemming* merupakan metode memotong bagian akhir kata sehingga menghasilkan bentuk yang tepat, sering kali dengan menghapus imbuhan yang mengubah arti atau jenis kata. proses pencarian kata dasar atau akar kata dari suatu kata yang berimbuhan atau variasi lainnya (Manning *et al.* 2009). Kata tersebut disederhanakan menjadi bentuk asalnya untuk menggambarkan makna yang lebih umum. Pencarian akar sebuah kata dapat memperkecil hasil indeks tanpa harus menghilangkan makna. Proses *stemming* dapat dilakukan dengan dua pendekatan yaitu dengan menggunakan kamus dan menggunakan aturan-aturan imbuhan (Utomo 2013).

## 2.12 *Term Frequency-Inverse Document Frequency (TF-IDF)*

Metode TF-IDF (*Term Frequency – Inverse Document Frequency*) merupakan metode untuk memberikan bobot hubungan suatu kata (*term*) terhadap dokumen. *Term Frequency* (TF) menentukan seberapa penting sebuah kata dilihat dari seberapa sering kata tersebut muncul dalam sebuah dokumen. Sedangkan *Inverse Document Frequency* (IDF) sebuah kata dianggap penting dalam dokumen apabila kata tersebut tidak terlalu sering muncul di dokumen lain. *Term Frequency – Inverse Document Frequency* sejauh ini dikenal sebagai metode ekstraksi fitur terbaik untuk analisis teks (Nguyen 2018). Cara kerja metode ini adalah dengan menggabungkan frekuensi kemunculan kata dalam sebuah dokumen tertentu dengan invers frekuensi dokumen yang mengandung kata tersebut untuk perhitungan bobot tiap kata. Frekuensi kemunculan kata di dalam sebuah dokumen menunjukkan seberapa penting kata tersebut terhadap dokumen itu dan frekuensi dokumen yang memiliki kata tersebut akan menunjukkan seberapa sering kata tersebut sehingga bobot hubungan antara sebuah kata dan sebuah dokumen akan tinggi jika di dalam dokumen frekuensi kata tersebut tinggi dan frekuensi keseluruhan dokumen yang mengandung kata tersebut yang rendah pada kumpulan dokumen (Nurjannah 2013).

Formula dari TF-IDF adalah (Rajaraman dan Ullman 2011):

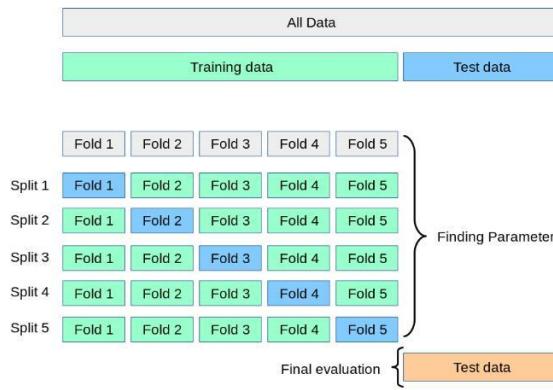
$$TF.IDF = \frac{f_i}{\max} \times \log_2 \left( \frac{N}{n_i} \right)$$

- $f_i$  adalah frekuensi kata i dalam dokumen j
- $\max$  adalah jumlah kemunculan maksimal kata dari dokumen (biasanya bernilai 1)
- N adalah jumlah data dalam dokumen
- $n_i$  jumlah kata i yang muncul dalam dokumen

## 2.13 *K-Fold Cross Validation*

*K-fold cross validation* adalah sebuah teknik untuk membagi dataset menjadi  $k$  sampel secara acak. Dari  $k$  sub-sampel, salah satu sampel digunakan untuk memvalidasi hasil, dan sisanya yaitu  $k-1$  sampel, digunakan untuk membangun model klasifikasi atau digunakan sebagai set data pelatihan. *K-fold cross validation* biasanya digunakan dengan nilai  $k=5$  atau  $k=10$ . Teknik *k-fold cross validation*

dapat mencapai hasil prediksi terbaik dengan mengoptimalkan pemilihan sampel dari set data yang tersedia untuk pelatihan dan validasi (Pachouly *et al.* 2022).



Gambar 2 Ilustrasi *k-fold cross validation*

### 2.14 Random Under Sampling

*Random undersampling* adalah metode dasar dan paling sederhana untuk melakukan *resampling* pada dataset yang tidak seimbang. *Random undersampling* mengambil sampel dari kelas mayoritas kemudian dieliminasi secara acak dari kelas tersebut untuk menyeimbangkan distribusi kelas untuk proses pembelajaran (Ali *et al.* 2019).

### 2.15 Confusion Matrix

*Confusion matrix* adalah matriks yang digunakan untuk melakukan evaluasi proses model klasifikasi berupa jumlah data uji yang benar dan salah. Matriks ini dapat mengetahui kualitas kinerja model klasifikasi (Normawati dan Prayogi 2021). Adanya *confusion matrix* untuk mengetahui sejauh mana *machine learning* bekerja sesuai dengan yang diinginkan. *Confusion matrix* berisi berbagai performa yang dapat diukur seperti akurasi, presisi, *recall* dan *F1-score* untuk mengetahui seberapa baik kinerja dari pemodelan yang telah dilakukan sebelumnya (Saputra 2022). Hasil *confusion matrix* dapat diketahui melalui tabel tingkat akurasi dari perhitungan klasifikasi berdasarkan jumlah data dan target kelasnya yang terdapat pada Tabel 1 (Tan *et al.* 2006).

Tabel 1 *Confusion matrix*

Kelas Aktual	Kelas Prediksi	
	Positif (P)	Negatif (N)
Positif (P)	True Positive (TP)	False Negative (FN)
Negatif (N)	False Positive (FP)	True Negative (TN)

Tabel di atas merupakan tabel *confusion matrix* dengan keterangan sebagai berikut:

- 1) *True Positive* (TP) = jumlah data nilai aktual positif dan nilai prediksi positif
- 2) *True Negative* (TN) = jumlah data nilai aktual negatif dan nilai prediksi negatif
- 3) *False Positive* (FP) = jumlah data nilai aktual positif dan nilai prediksi negatif
- 4) *False Negative* (FN) = jumlah data nilai aktual negatif dan nilai prediksi positif

### III METODE

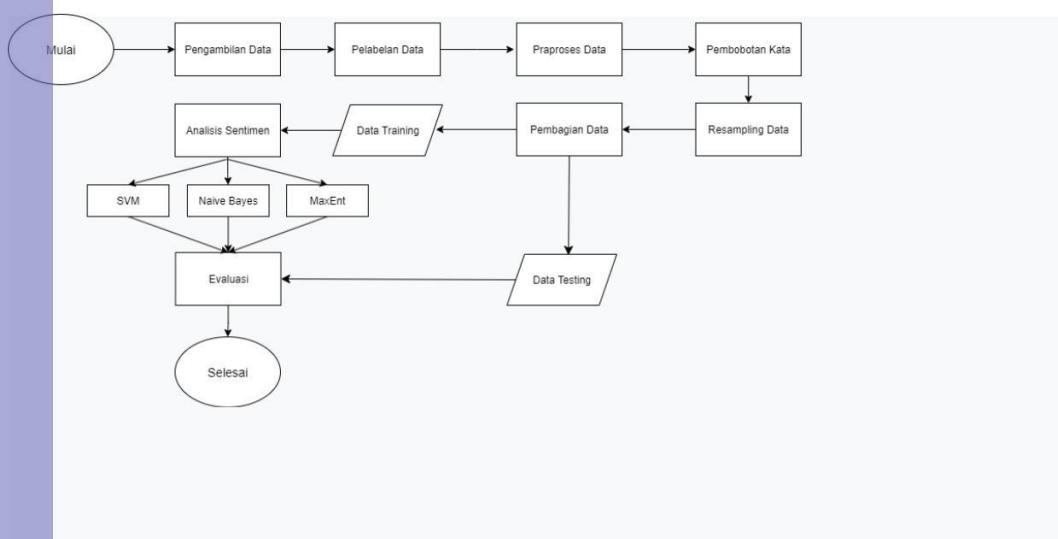
#### 3.1 Lingkungan Pengembangan

Spesifikasi perangkat lunak dan perangkat keras yang akan digunakan untuk penelitian ini adalah sebagai berikut:

1. Perangkat keras dengan spesifikasi:
  - Processor AMD Ryzen 5 7520U with Radeon Graphics (8 CPUs), ~2.8GHz.
  - RAM 16 GB.
  - SSD 512 GB.
2. Perangkat lunak dengan spesifikasi:
  - Windows 11 Home Single Language 64-bit
  - Python 3.11.0
  - Microsoft Visual Studio Code
  - Microsoft Excel 2021
  - Google Colab

#### 3.2 Tahapan Penelitian

Tahapan penelitian sesuai dengan Gambar 3 terdiri dari pengambilan data, praproses data, pembobotan kata, pembagian data menggunakan *k-fold cross validation* sebanyak  $k=10$  menjadi *data training* dan *data testing*, kemudian dilakukan *resampling data* pada *data training*. Tahapan berikutnya adalah dilakukan klasifikasi menggunakan tiga model algoritma *machine learning* yaitu *maximum entropy*, *naive bayes* dan *support vector machine* serta tahapan terakhir dilakukan evaluasi model menggunakan *confusion matrix* untuk mendapat tabel hasil klasifikasi dan menghitung akurasi model.



Gambar 3 Tahapan penelitian

#### 3.3 Pengambilan Data

Data diambil dengan melakukan permohonan permintaan data kepada perusahaan Ivosight perihal tweet dengan kata kunci “Transjakarta”. Data yang

diamambil dari kurun waktu 1 Januari 2023 - 14 Juni 2023. Jumlah data sebanyak 238.734 data tweet. Atribut yang ada di data tersebut terdapat pada Tabel 2 berikut.

Tabel 2 Atribut data tweet

**Atribut Data Tweet**

Service
Date
Time
Datetime
Follower
Engagement
Comment
Likes
Shares
Views
user_id
Username
Name
user created at
user total post
Gender
Sentiment
Image
Message
url
in reply to

secara signifikan (Haddi *et al.* 2013). Pada penelitian ini dilakukan beberapa perlakuan praproses data yaitu *cleaning* (*filtering*), *casefolding*, tokenisasi, *stopwords removal*, normalisasi bahasa dan *stemming*.

### 3.6 Data Cleaning

Proses *cleaning* adalah proses penghapusan berbagai macam karakter atau data teks yang tidak diperlukan dalam “message” seperti url, mention (@), hashtag (#), emoji, angka atau tanda baca yang tidak relevan. Hasil dari “message” yang sudah melewati proses *data cleansing* tidak berisi karakter-karakter yang tidak diperlukan dalam menganalisis sentimen.

### 3.7 Casefolding

*Casefolding* adalah proses untuk melakukan standarisasi semua kata dalam sebuah data menjadi huruf kecil (non-kapital). Hal ini dilakukan agar memastikan kata yang kapital dengan yang tidak diperlakukan dengan sama. Hal ini supaya semua data bersifat seragam dari data satu ke data lainnya.

### 3.8 Normalisasi

Normalisasi bahasa adalah mengubah kata-kata yang tidak baku atau singkatan yang sering digunakan menjadi kata baku yang standar sesuai kamus. Peneliti menggunakan kamus normalisasi *colloquial indonesian lexicon* yang berasal dari Github nasalsabila. Dalam kampus tersebut terdapat kata slang dan kata formal.

### 3.9 Tokenisasi

Tokenisasi adalah tahapan untuk memecah teks dalam satu kalimat menjadi kata individual (token). Hal ini dilakukan sebelum tahapan *stopwords removal* agar proses tersebut dapat dilakukan.

### 3.10 Stopwords Removal

Setelah teks dibagi menjadi kata individual, dilakukan tahapan *stopwords removal*. *Stopwords removal* adalah tahapan untuk menghapus kata-kata yang tidak memiliki makna, seperti kata hubung atau kata depan. Contohnya seperti “yang”, “di”, “dan”. *Stopwords* yang dihapus berasal dari *library nltk.corpus* yang diimpor dari *stopwords*. Dari library tersebut peneliti menambah lagi beberapa kata tambahan yang termasuk dalam stopwords. Daftar kata-kata stopwords ini terdapat pada lampiran 1

### 3.11 Stemming

Terakhir *stemming* adalah tahapan kata imbuhan menjadi kata dasar. Hal ini dilakukan untuk mengurangi variasi kata yang sebenarnya memiliki arti yang sama. Tahapan *stemming* pada penelitian ini menggunakan *library Sastrawi* dan mengimpor dari *StemmerFactory*.

### 3.12 Pembobotan Kata

Metode TF-IDF (*Term Frequency – Inverse Document Frequency*) merupakan teknik pemodelan data yang berfungsi untuk menghitung bobot untuk setiap kata yang mengindikasikan pentingnya kata tersebut untuk dokumen tertentu dalam dokumen. TF (*Term Frequency*) mendefinisikan kemunculan sebuah kata  $w$  dalam sebuah dokumen  $d$ . IDF (*Inverse Document Frequency*) mengukur kelangkaan dari sebuah kata  $w$  dalam keseluruhan dokumen (Danilo *et al.* 2021). *Term Frequency – Inverse Document Frequency* sejauh ini dikenal sebagai metode ekstraksi fitur terbaik untuk analisis teks (Nguyen 2018).

Jika disesuaikan dengan penelitian yang dilakukan, maka formula dari TF-IDF akan menjadi seperti berikut:

$$TF.IDF(w, d) = TF(w, d) \times \log\left(\frac{N}{DF(w)}\right)$$

- $TF(w, d)$  adalah frekuensi kata  $w$  dalam dokumen  $d$
- $DF(w)$  adalah jumlah dokumen (data) yang mengandung kata  $w$
- $N$  adalah jumlah data

### 3.13 Resampling Data

*Random under sampling* (RUS) digunakan sebagai metode yang digunakan untuk mengatasi data yang tidak seimbang. *Random under sampling* dipilih karena metode ini mengurangi sampel mayoritas sehingga mempercepat waktu pelatihan model.

### 3.14 Pembagian Data

Data akan dibagi menjadi dua jenis data yaitu data latih dan data uji. Pembagian ini menggunakan *k-fold cross validation* sebagai pembaginya dengan  $k = 10$ . Dengan  $k = 10$  maka setiap iterasi  $k$  akan membagi data dengan proporsi data latih banding data uji sebesar 90%:10%.

### 3.15 Analisis Sentimen

Analisis sentimen yang mengklasifikasi dokumen menjadi kelas positif, netral dan negatif ini akan menunjukkan sentimen masyarakat terhadap Transportasi Transjakarta. Tahapan analisis sentimen dimulai dengan metode *WordCloud* untuk mengetahui kata yang dominan pada setiap kelompok sentimen. Hasil ini kemudian dianalisis untuk mengetahui sentimen masyarakat terhadap Transportasi Transjakarta. Analisis sentimen selanjutnya menggunakan metode machine learning dengan tiga algoritma. Pemodelan klasifikasi data tweet akan menggunakan tiga algoritma *machine learning* yaitu *maximum entropy* (MaxEnt), *naive bayes* dan *support vector machine* (SVM).

Jika kita sesuaikan variabel dari setiap algoritma machine learning yang digunakan, maka pada *maximum entropy* penjelasan setiap variabel akan menjadi seperti berikut:

$$\pi_{ij} = \frac{\sum_{k=1}^n \pi_{ijk}}{\sum_{k=1}^n \sum_{l=1}^m \pi_{ilk}}$$

- $P_{ME}(c|d, \lambda)$  adalah probabilitas dari sebuah data  $d$  (tweet) memiliki kelas  $c$  (sentimen) dengan parameter/bobot  $\lambda$ .
- $\lambda_i$  adalah parameter/bobot yang dipelajari untuk setiap fitur. Pada kasus ini menandakan seberapa kuat sebuah fitur menjadi indikator untuk suatu sentimen.
- $f_i(c,d)$  adalah fungsi fitur yang bernilai 1 jika memiliki ciri tertentu dan 0 jika tidak. Contohnya:
  - $f_1(\text{'positif'}, \text{tweet})$  akan bernilai 1 jika kata "nyaman" ada di dalam "message".
  - $f_2(\text{'negatif'}, \text{tweet})$  akan bernilai 1 jika kata "jelek" ada di dalam "message".
- $\sum_c$  adalah penjumlahan atas semua kelas sentimen yang memungkinkan (positif, netral, negatif) untuk menormalisasi hasil agar total probabilitas menjadi 1.
- $c$  adalah kelas sentimen yang ingin diprediksi dari kolom "sentiment", yaitu 'positif', 'netral' dan 'negatif'.
- $d$  adalah data tweet yang dianalisis, direpresentasikan oleh vektor TF-IDF dari kolom "message".

Kemudian pada *naive bayes* variabelnya akan menjadi seperti berikut:

$$( | ) \propto ( | ) \prod = ( | ) ( | )$$

- $(c|d)$  adalah probabilitas data  $d$  (tweet) termasuk dalam kelas  $c$  (sentiment).
- $P(c)$  adalah probabilitas awal dari kelas  $c$  (sentimen) dan dihitung dari frekuensi masing-masing sentimen dataset training.
- $P(f_i|c)$  adalah probabilitas sebuah kata  $f_i$  muncul dalam kumpulan data "message" yang memiliki sentimen (kelas  $c$ ).
- $( | )$  adalah jumlah kemunculan kata  $f_i$  dalam "message"  $d$ .
- $c$  adalah kelas target dari kolom "sentiment", yaitu 'positif', 'netral' dan 'negatif'.
- $d$  adalah teks dari satu baris kolom "message".
- $f_i$  adalah fitur ke- $i$ , yaitu setiap kata (token) unik dalam "message".
- $P(d)$  diabaikan karena praktiknya nilainya sama untuk semua kelas sehingga diindikasikan dengan simbol  $\propto$  yang berarti "sebanding dengan".

Kemudian pada *support vector machine* variabelnya akan menjadi seperti berikut:

$$( | ) = .\Phi( | ) +$$

- $x$  adalah data input berupa vektor TF-IDF yang dihasilkan dari teks pada kolom "message".
- $y(x)$  adalah skor sentimen sebuah tweet.
- $w$  adalah vektor bobot yang dipelajari oleh model. Vektor ini menentukan kemiringan dari *hyperplane*

- $\Phi(x)$  adalah fungsi kernel yang memetakan vektor  $x$  ke ruang dimensi yang lebih tinggi agar data sentimen positif, netral, dan negatif dapat dipisahkan oleh *hyperplane*
- $b$  adalah Bias, yaitu nilai yang menggeser posisi bidang pemisah. Hal ini agar bidang pemisah sesuai diantara kelompok-kelompok sentimen yang berbeda.

### 3.16 Evaluasi

Algoritma yang dikembangkan dievaluasi menggunakan *confusion matrix*. Evaluasi digunakan untuk menghitung matriks *accuracy*, *precision*, *recall* dan *F1-score* berdasarkan hasil *confusion matrix*. Untuk menghitung *accuracy*, *precision*, *recall* dan *F1-score* kita membutuhkan nilai *confusion matrix*. Karena kelas sentimen pada penitian ini berjumlah tiga yaitu ‘positif’, ‘netral’, dan ‘negatif’ maka model *confusion matrix*-nya akan menjadi seperti Tabel 3 berikut.

Tabel 3 Model *confusion matrix*

		Prediksi			
		Aktual	-1	0	1
Aktual	-1	TN	TN	FP	
	0	TN	TN	FP	
	1	FN	FN	TP	

Maka untuk menghitung *accuracy*, *precision*, *recall* dan *F1-score* dibutuhkan rumus berikut :

- ACC (*Accuracy*) =  $(TP + TN) / (P + N) = (TP + TN) / (TP+TN+FP+FN)$
- Precision/PPV (*Positive predictive value*) =  $TP / (TP+FP)$
- Recall/TPR (*True positive rate*) =  $TP/P = TP/(TP+FN)$
- *F1-score* =  $2(PPV)(TPR) / PPV + TPR$

## IV HASIL DAN PEMBAHASAN

### 4.1 Pengambilan Data

Data diambil dengan melakukan permohonan permintaan data kepada perusahaan Ivosight mengenai tweet dengan kata kunci “Transjakarta”. Data yang diambil dari kurun waktu 1 Januari 2023 - 14 Juni 2023. Jumlah data sebanyak 238.734 data tweet. Atribut yang digunakan untuk penelitian dari Tabel 2 adalah ”datetime”, ”username”, ”sentiment”, ”message”. Klasifikasi awal berupa pelabelan sentimen sudah dilakukan oleh perusahaan Ivosight dengan metode *counting* berdasarkan dari bank sentimen data yang dimiliki oleh perusahaan tersebut. Proses tersebut menghasilkan jumlah data yang memiliki sentimen positif sebanyak 6.225, sentimen negatif sebanyak 6.879 dan sentimen netral sebanyak 225.630.

### 4.2 Pelabelan Data

Setelah data didapatkan, data-data duplikat dihapus agar data yang digunakan lebih valid. Penghapusan data ini dilakukan dengan data yang berasal dari retweet tanpa kalimat tambahan. Setelah data duplikat tersebut dihapus, data berkurang menjadi 155.957 data tweet. Proporsi klasifikasi dari data tersebut adalah 5.360 data positif, 3.915 data negatif, dan 143.487 data netral.

Setelah dilakukan pengurangan data yang duplikat, dilakukan pelabelan sentimen secara manual. Pelabelan ini dilakukan oleh tim peneliti secara manual agar mendapatkan perspektif sentimen dari manusia secara langsung. Setelah dilakukan pelabelan secara manual, proporsi klasifikasi dari data tersebut adalah 6.139 data positif, 6.331 data negatif, dan 143.487 data netral.

### 4.3 Praproses Data

Praproses data dilakukan dengan tahapan *data cleansing*, *casefolding*, normalisasi, tokenisasi, *stopwords removal* dan *stemming* secara berurutan. *Data cleansing* adalah tahapan menghapus berbagai karakter yang tidak relevan dengan kalimat. Merujuk pada kalimat diatas perubahan dari tweet awal setelah dilakukan *data cleansing* tertera pada Tabel 4 berikut.

**Tabel 4 Tahapan *data cleansing* para praproses data**

<b><i>Tweet awal</i></b>	<b><i>Setelah <i>data cleansing</i></i></b>
@PT_Transjakarta @RaniSihotang8 Kak, yg menuju cengkareng/kalideres Dari arah kuningan yg masih buka jalur mana aja ya sampe jam berapa?	Kak yg menuju cengkarengkalideres Dari arah kuningan yg masih buka jalur mana aja ya sampe jam berapa

Pada tahapan tersebut, karakter *add* (@), *underscore* (\_) dan tanda tanya(?) dihapus untuk membersihkan data. Dihapus juga kata yang beriringan dengan karakter @ karena dianggap data yang tidak relevan. Hasil dari tahapan *data*



*cleansing* terdapat pada kolom setelah *cleansing*. Tahapan berikutnya adalah *casefolding*, yaitu merubah seluruh huruf kapital menjadi huruf kecil. Contoh dari tahapan *casefolding* ada pada Tabel 5 berikut.

Tabel 5 Tahapan *casefolding* para praproses data

Setelah <i>data cleansing</i>	Setelah <i>casefolding</i>
Kak yg menuju cengkareng/kalideres Dari arah kuningan yg masih buka jalur mana aja ya sampe jam berapa	kak yg menuju cengkarengkalideres dari arah kuningan yg masih buka jalur mana aja ya sampe jam berapa

Pada tahapan *casefolding* seluruh kata yang masih mengandung huruf kapital dirubah menjadi huruf kecil. Hal ini bertujuan agar seluruh teks seragam dan memiliki perlakuan yang sama. Tahapan berikutnya adalah normalisasi yaitu mengubah kata-kata yang tidak baku menjadi kata-kata baku. Contoh dari tahapan normalisasi ada pada Tabel 6 berikut.

Tabel 6 Tahapan normalisasi para praproses data

Setelah <i>casefolding</i>	Setelah <i>normalisasi</i>
kak yg menuju cengkarengkalideres dari arah kuningan yg masih buka jalur mana aja ya sampe jam berapa	kak yang menuju cengkarengkalideres dari arah kuningan yang masih buka jalur mana aja ya sampai jam berapa

Pada tahapan ini kata "yang" yang ditulis tidak baku dengan tulisan "yg" dirubah ke bentuk baku yaitu "yang". Kata lain yang berubah adalah kata "sampe" menjadi "sampai". Perubahan ini didasarkan pada kamus yang berasal dari Github milik nasalsabila berjudul *colloquial indonesian lexicon*. Kamus yang berbentuk dokumen ini diinput dalam program dalam format .csv. Tahapan berikutnya melakukan tokenisasi, yaitu membagi kalimat menjadi teks-teks individual (*token*). Contoh tahapan tokenisasi ada pada Tabel 7 berikut.

Tabel 7 Tahapan tokenisasi para praproses data

Setelah <i>normalisasi</i>	Setelah <i>tokenisasi</i>
kak yang menuju cengkarengkalideres dari arah kuningan yg masih buka jalur mana aja ya sampai jam berapa	[kak], [yang], [menuju], [cengkarengkalideres], [dari], [arah], [kuningan], [yang], [masih], [buka], [jalur], [mana], [aja], [ya], [sampai], [jam], [berapa]

Setelah kalimat dibagi menjadi teks-teks individual, tahapan selanjutnya adalah *stopwords removal*. *Stopwords removal* adalah penghapusan kata-kata stopwords, yaitu kata-kata yang dianggap tidak memiliki makna. Daftar kata-kata

yang dihapus dapat dilihat pada lampiran 1. Contoh tahapan *stopwords removal* dapat dilihat pada Tabel 8 berikut.

Tabel 8 Tahapan *stopwords removal* para praproses data

Setelah tokenisasi	Setelah <i>stopword removal</i>
[kak], [yang], [menuju], [cengkarengkalideres], [dari], [arah], [kuningan], [yang], [masih], [buka], [jalur], [mana], [aja], [ya], [sampai], [jam], [berapa]	[kak], [cengkarengkalideres], [arah], [kuningan], [buka], [jalur], [jam]

Pada kolom *stopword removal* beberapa kata yang dianggap *stopwords* telah hilang dan menghasilkan kalimat yang masih berbentuk *token* yang hanya berisi kalimat-kalimat yang memiliki substansi. Tahapan berikutnya yaitu *stemming*, yaitu merubah kata menjadi bentuk dasarnya. Namun pada contoh ini semua kata sudah dalam bentuk dasar sehingga tidak ada yang perlu dirubah. Setelah berbagai tahapan praproses data, perubahan pada proses ini dapat dilihat secara signifikan pada Tabel 9 berikut.

Tabel 9 Contoh tahapan praproses data

Tweet awal	Setelah cleansing sampai normalisasi	Setelah stemming
@PT_Transjakarta @RaniSihotang8 Kak, yg menuju cengkareng/kalideres Dari arah kuningan yg masih buka jalur mana aja ya sampe jam berapa?	kak yang menuju cengkarengkalideres dari arah kuningan yang masih buka jalur mana saja ya sampai jam berapa	[kak, cengkarengkalideres, arah, kuningan, buk...]

#### 4.4 Pembobotan Kata

Pembobotan kata atau juga biasa disebut ekstraksi fitur dilakukan dengan menggunakan metode TF-IDF. Pembobotan kata dilakukan sebelum pembagian data menggunakan metode *k-fold cross validation*. Perhitungan bobot suatu kata didasarkan pada rumus TF-IDF. Bobot tersebut dihitung pada setiap data dalam satu dokumen dikali dengan logaritma dari total data dalam dokumen dibagi berapa kali kata tersebut muncul dalam dokumen. Dari perhitungan tersebut akan didapatkan bobot dari setiap kata dalam teks. Bobot ini digunakan untuk proses selanjutnya yaitu analisis sentimen menggunakan *machine learning*.

#### 4.5 Resampling Data

Data yang akan dilatih berjumlah 155.957 data tweet dengan proporsi 6.139 data positif, 6.331 data negatif, dan 143.487 data netral. Dengan kondisi data yang tidak seimbang diperlukan proses penyeimbangan data. Untuk menyeimbangkan data yang digunakan sebagai data latih, dilakukan *resampling* pada data latih dengan metode *random under sampling* (RUS). Metode RUS digunakan karena

mampu mengurangi waktu yang digunakan untuk menjalankan program. Setelah dilakukan proses RUS data yang digunakan menjadi 18.417 data. Data tersebut terdiri dari 6.139 data positif, 6.139 data netral dan 6.139 data negatif.

## 4.6 Pembagian Data

Data dibagi menjadi data latih dan data uji menggunakan metode *k-fold cross validation* iterasi  $k = 10$  dengan proporsi pembagian data latih dan uji sebesar 90%:10%. Pilihan  $k = 10$  digunakan dalam beberapa penelitian karena mampu menunjukkan model yang digunakan tidak bergantung pada ukuran sampel pelatihan sehingga bisa mengurangi bias estimasi dibanding menggunakan  $k$  dengan ukuran lebih kecil (Kohavi 1995, Kim 2009). Data latih akan digunakan untuk mempelajari model dan data uji digunakan untuk mengevaluasi akurasi model.

#### 4.7 Analisis Sentimen

Metode *wordcloud* untuk melihat intensitas kata yang sering muncul pada data yang diteliti. Kata dominan yang muncul pada setiap kelas kemudian dianalisis untuk mendapatkan sentimen masyarakat mengenai transportasi Transjakarta. Hasil *wordcloud* dilakukan sesuai dengan label sentimen tweet yaitu sentimen positif dan negatif. Hasil dari *wordcloud* dapat dilihat pada Gambar 4 berikut.



Gambar 4 *Wordcloud* sentimen kelas positif (kiri) dan sentimen kelas negatif (kanan)

Pada *wordcloud* sentimen kelas positif kata yang menonjol dan relevan adalah “bagus”, “baik”, “cepat”, “keren”, “nyaman”, “murah”. Hal ini menunjukkan pengalaman positif dan kepuasan masyarakat ketika menggunakan moda transportasi Transjakarta. Transjakarta dipandang memberikan fasilitas yang baik, memberikan kenyamanan dengan harga yang terjangkau ke masyarakat. Pada *wordcloud* sentimen kelas negatif kata yang menonjol dan relevan adalah “bodoh”, “lama”, “penuh”, “rusak”, “jelek”, “macet”. Hal ini menunjukkan pengalaman negatif dan ketidaknyamanan masyarakat ketika menggunakan moda transportasi Transjakarta. Meski ada sentimen positif mengenai Transjakarta, masyarakat juga beranggapan ada hal-hal negatif pada moda transportasi Transjakarta.

Kesimpulan yang dapat diambil adalah pelayanan moda transportasi Transjakarta sudah memberikan pelayanan yang baik karena respon masyarakat yang positif. Moda transportasi dinilai sudah memberikan kenyamanan oleh masyarakat. Harga yang terjangkau juga menjadi poin yang sangat meningkatkan sentimen positif masyarakat mengenai Transjakarta. Akan tetapi, moda transportasi Transjakarta masih perlu meningkatkan pelayanannya karena ada sentimen negatif

di masyarakat. Seperti optimalisasi rute atau penambahan armada untuk kasus sentimen “lama” atau “penuh”. Juga dilakukan perbaikan fasilitas karena ada sentimen yang mengatakan fasilitas Transjakarta yang “rusak” dan “jelek”.

Setelah dilakukan pemetaan kata menggunakan *wordcloud*, tahapan berikutnya melakukan analisis sentimen dengan *machine learning* menggunakan tiga algoritma. Tiga model algoritma klasifikasi *machine learning* tersebut yaitu *naive bayes*, *maximum entropy* serta *support vector machine*. Model *naive bayes* menggunakan *package scikit-learn* bernama *multinomialNB*. Model *maximum entropy* menggunakan *package scikit-learn* bernama *LogisticRegression*. Model *support vector machine* menggunakan *package scikit-learn* bernama SVC (*Support Vector Classification*).

Algoritma yang digunakan pada penelitian ini adalah *maximum entropy* atau juga disebut *logistic regression* diterapkan model klasifikasi yang memaksimalkan entropi dari probabilitas dengan mempertimbangkan berbagai fitur. Konsepnya adalah menentukan nilai pada setiap kata dalam satu kalimat. Nilai yang paling tinggi dijadikan prediksi untuk menentukan sentimen. Algoritma ini akan mengekstrak fitur-fitur dalam hal ini setiap kata pada data tweet dan mempelajari bobot (yang pada rumus dilambangkan sebagai  $w_i$ ) untuk setiap fitur. Pemilihan kelas pada *maximum entropy* juga didasarkan dari prediksi kelas dengan nilai probabilitas tertinggi. Hal ini karena algoritma *maximum entropy* termasuk ke dalam jenis algoritma probabilitas (*probabilistic classifier*) (Dang *et al.* 2020).

Algoritma *maximum entropy* mampu melakukan klasifikasi tiga kelas (positif, negatif, dan netral) menggunakan fungsi *softmax* ketika menangani kondisi kelas yang lebih dari dua. Fungsi ini sudah termasuk dalam *package scikit-learn* bernama *LogisticRegression* yang digunakan. Fungsi *softmax* adalah mengubah skor mentah atau dalam hal ini bobot kata menjadi probabilitas dan bertujuan mencari kelas yang paling sesuai dari rentang probabilitas tersebut. Penerapan akhir latihan model dengan algoritma *maximum entropy* adalah sentimen yang berasal dari prediksi kelas dengan nilai yang paling besar.

Algoritma yang digunakan berikutnya dalam penelitian ini adalah algoritma *naive bayes*. Konsep *naive bayes* adalah menebak sentimen dari seberapa sering kata dalam tweet muncul pada masing-masing kelas. Algoritma *naïve bayes* melakukan perhitungan probabilitas dengan aturan bayes, *package scikit-learn multinomialNB* mampu melakukan hal tersebut. Aturan bayes adalah sebuah teori probabilitas yang menghubungkan probabilitas bersyarat dengan suatu kejadian dengan informasi yang baru.

Pada proses pelatihan menggunakan algoritma *naive bayes*, proses klasifikasi sentimen dimulai dari penghitung *prior probability* pada setiap kelas yang berarti pada penelitian ini adalah positif, negatif, dan netral berdasarkan banyaknya kemunculan data dalam data latih. Proses berikutnya adalah *likelihood* atau kemungkinan muncul suatu probabilitas pada masing-masing kelas dihitung. Fitur di sini merupakan kata, yang berarti probabilitas kemunculan ini dihitung berdasarkan frekuensi kata yang muncul dalam kelas. Proses prediksi kelas ditentukan dengan nilai *prior probability* yang paling tinggi dari setiap kelas yang dihitung. Jika nilai dari suatu kelas bernilai paling tinggi, maka sentimen dari data teks adalah kelas tersebut.

Algoritma terakhir yang digunakan adalah *support vector machine*. *Support vector machine* adalah algoritma yang bekerja dengan menemukan *hyperplane*

(bidang pemisah) paling optimal untuk memisahkan kelas terhadap fitur dengan margin terbesar. *Support vector machine* berbeda dengan dua algoritma sebelumnya. Algoritma *maximum entropy* dan *naive bayes* adalah *probabilistic classifier* sementara *support vector machine* adalah *linear classifier*. *Linear classifier* adalah algoritma klasifikasi yang memprediksi hasil berdasarkan dari pemisahan linear antara kelas menggunakan fitur. *Support vector machine* yang termasuk *linear classifier* menggunakan fungsi kernel linear untuk mentransformasikan data ke ruang dimensi yang lebih tinggi. *Support vector machine* menggunakan bobot kata dari proses TF-IDF dalam menentukan *hyperplane*. Parameter  $c$  pada fungsi support vector machine adalah variabel yang “mengizinkan” seberapa besar toleransi terhadap kesalahan klasifikasi.

Tahapan awal dari algoritma *support vector machine* adalah mengubah teks menjadi vektor fitur. Hal ini sudah dilakukan diawal dengan tahapan penelitian pembobotan kata menggunakan TF-IDF. Setelah itu algoritma akan mencari *hyperplane* pemisah dengan margin maksimal antar kelas dalam ruang fitur. Jika kelas yang diberikan bersifat biner (misal positif dan negatif) maka hasil prediksi sesuai dengan output fungsi *support vector machine*. Jika outputnya lebih besar atau sama dengan 0 maka prediksi kelas bernilai positif dan jika lebih kecil dari 0 maka hasil prediksi kelas bernilai negatif. Namun dalam menangani kasus multi-kelas (kelas lebih dari 2) maka ada dua pendekatan khusus yang bisa dilakukan, yaitu *one-vs-rest* (OvR) dan *one-vs-one* (OvO). *One-vs-rest* adalah pendekatan yang melatih model dengan membangun satu model biner per kelas dengan setiap model memisahkan satu kelas dari semua kelas lainnya. Contoh kasus klasifikasi dengan tiga kelas (positif, negatif, netral) maka kelas akan membangun tiga model yaitu positif-vs-negatif + netral, negatif-vs-positif + netral, dan terakhir netral-vs-positif + negatif. Hasil prediksi berdasarkan dari nilai fungsi tertinggi dari ketiga model tersebut. Sementara untuk pendekatan *one-vs-one* adalah pendekatan yang melatih model dengan setiap pasangan kelas. Contoh kasus dengan kelas yang sama, maka kelas akan membangun tiga model yaitu positif-vs-negatif, positif-vs-netral, dan negatif-vs-netral. Hasil prediksi ditentukan dengan menghitung semua model dan memilih berdasarkan kelas yang paling banyak muncul.

Pada penelitian ini pemodelan algoritma *support vector machine* menggunakan *package scikit-learn support vector classification*. *Library* ini mampu melakukan proses analisis sentimen menggunakan algoritma *support vector machine*. Ketika bertemu dengan kasus klasifikasi dengan multi-kelas, maka secara *default* pendekatan yang dilakukan oleh *library* ini adalah menggunakan *one-vs-one*.

#### 4.8 Evaluasi

Setelah dilakukan pemodelan, ketiga model dievaluasi menggunakan *confusion matrix* serta dihitung rata-ratanya dengan  $k = 10$ . Evaluasi tersebut menghasilkan tabel hasil klasifikasi serta tabel *confusion matrix* sesuai dengan Tabel 10-16.

Hak Cipta Dilindungi Undang-undang  
 1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

- Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
- Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Tabel 10 Rata-rata *confusion matrix maximum entropy*

Aktual	Prediksi		
	-1	0	1
-1	424	121	67
0	128	430	54
1	88	86	439

Tabel 11 Rata-rata hasil klasifikasi *maximum entropy*

Sentimen	precision	Recall	F1-score	support
-1	0.66	0.69	0.68	6139.0
0	0.67	0.70	0.69	6139.0
1	0.78	0.72	0.75	6139.0
Accuracy			0.70	
macro avg	0.70	0.70	0.70	1842.0
weigthed avg	0.70	0.70	0.70	1842.0

Pada Tabel 10 terdapat hasil dari rata-rata *confusion matrix* menggunakan algoritma *maximum entropy*. Dengan rumus menghitung *accuracy*, *precision*, *recall* dan *F1-score*, didapatkan *precision* sebesar 70%, *recall* 70%, *F1-score* 70% dan algoritma *maximum entropy* memberikan *accuracy* sebesar 70%. Hasil evaluasi ini dapat dilihat pada Tabel 11.

Tabel 12 Rata-rata *confusion matrix naive bayes*

Aktual	Prediksi		
	-1	0	1
-1	501	57	55
0	207	344	61
1	206	71	335

Tabel 13 Rata-rata hasil klasifikasi *naïve bayes*

Sentimen	precision	Recall	F1-score	support
-1	0.55	0.82	0.66	6139.0
0	0.73	0.56	0.63	6139.0
1	0.74	0.55	0.63	6139.0
Accuracy			0.64	
macro avg	0.67	0.64	0.64	1842.0
weigthed avg	0.67	0.64	0.64	1842.0



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
- b. Pengutipan tidak merugikan kepentingan yang wajar IPB University

Pada Tabel 12 terdapat hasil dari rata-rata *confusion matrix* menggunakan algoritma *naive bayes*. Dengan rumus menghitung *accuracy*, *precision*, *recall* dan *F1-score*, didapatkan *precision* sebesar 67%, *recall* 64%, *F1-score* 64% dan algoritma *naive bayes* memberikan *accuracy* sebesar 64%. Hasil evaluasi ini dapat dilihat pada Tabel 13.

Tabel 14 Rata-rata *confusion matrix support vector machine*

<b>Aktual</b>	<b>Prediksi</b>		
	<b>-1</b>	<b>0</b>	<b>1</b>
-1	451	106	56
0	148	424	41
1	99	79	434

Tabel 15 Rata-rata hasil klasifikasi *support vector machine*

<b>Sentimen</b>	<b>precision</b>	<b>Recall</b>	<b>F1-score</b>	<b>support</b>
-1	0.65	0.74	0.69	6139.0
0	0.70	0.69	0.69	6139.0
1	0.82	0.71	0.76	6139.0
Accuracy			0.71	
macro avg	0.72	0.71	0.71	1842.0
weighited avg	0.72	0.71	0.71	1842.0

Pada Tabel 14 terdapat hasil dari rata-rata *confusion matrix* menggunakan algoritma *support vector machine*. Dengan rumus menghitung *accuracy*, *precision*, *recall* dan *F1-score*, didapatkan *precision* sebesar 72%, *recall* 71%, *F1-score* 71% dan algoritma *support vector machine* memberikan *accuracy* sebesar 71%. Hasil evaluasi ini dapat dilihat pada Tabel 15.

Tabel 16 Perbandingan hasil klasifikasi tiga algoritma machine learning

<b>Model</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>	<b>Accuracy</b>
SVM	0.72	0.71	0.71	0.71*
MaxEnt	0.70	0.70	0.70	0.70**
Naive Bayes	0.67	0.64	0.64	0.64***

\*SVM = 0.711406519730048, \*\* MaxEnt = 0.7030452457918057, \*\*\* NB = 0.6417969627751523

Perbandingan hasil evaluasi dari ketiga algoritma yang digunakan dapat dilihat pada Tabel 16. Tabel tersebut diurutkan berdasarkan tingkatan akurasi yang paling tinggi hingga paling rendah secara berurutan dari atas ke bawah. Algoritma *support vector machine* menjadi algoritma yang mendapatkan hasil akurasi paling



tinggi yaitu sebesar 71% dengan *precision* sebesar 72%, *recall* sebesar 71% dan *F1-score* sebesar 71%.

Hasil evaluasi dari ketiga algoritma menunjukkan bahwa *support vector machine* menjadi algoritma yang memiliki performa paling tinggi dibandingkan *maximum entropy* dan *naive bayes*. Hal ini dikarenakan algoritma *support vector machine* memiliki keunggulan dibandingkan dua algoritma lainnya yang secara fundamental memiliki prinsip kerja berbeda dalam menangani data teks yang kompleks dan berdimensi tinggi.

Secara prinsip kerja, *naive bayes* memiliki karakteristik independensi kata yang berarti bahwa setiap kata bersifat independen satu sama lain jika kelasnya sudah diketahui. Pada bahasa alami sehari-hari hal ini justru sering kali tidak sesuai dengan makna sebenarnya. Misalkan pada latihan algoritma *naive bayes* menemukan frasa "pelayanan tidak bagus" algoritma ini akan memproses kata "tidak" dan "bagus" secara terpisah. Proses ini dapat menyebabkan menghilangnya konteks yang ingin didapatkan ketika melatih fitur. Karakteristik ini menjadi kekurangan utama algoritma *naive bayes* dalam memproses fitur yang memiliki dimensi yang tinggi.

Algoritma *maximum entropy* diimplementasikan pada penelitian ini dengan *logistic regression* yang tidak memiliki karakteristik independensi kata seperti *naive bayes*. Sehingga bisa dilihat pada hasil evaluasi akurasi algoritma *maximum entropy* lebih baik dibandingkan dengan *naive bayes*. Namun *maximum entropy* memiliki prinsip kerja yang berbeda dengan *support vector machine* yang membuat tujuan kedua algoritma berbeda. *Maximum entropy* bekerja dengan memaksimalkan probabilitas kemiripan (*likelihood*) dari keseluruhan dari *data training*. Setelah model dilatih model bisa digunakan untuk membuat prediksi pada data uji maupun data baru.

*Support vector machine* pada prinsip kerjanya memaksimalkan margin dengan mencari *hyperplane* pemisah antar vektor kelas. Fungsi kernel pada *support vector machine* mampu untuk menangkap interaksi yang kompleks antar kata. *Support vector machine* memiliki efektivitas dalam ruang berdimensi tinggi, cenderung menciptakan model yang lebih *robust* dan memiliki kemampuan generalisasi yang lebih baik pada data uji. Dengan prinsip kerja inilah *support vector machine* memberikan akurasi tertinggi dan memiliki keunggulan dibanding kedua algoritma lainnya

## V SIMPULAN DAN SARAN

### 5.1 Simpulan

Dalam penelitian ini peneliti telah membangun sebuah model analisis sentimen untuk mengetahui pendapat masyarakat mengenai moda transportasi Transjakarta. Pendapat tersebut berasal dari media sosial X yang diambil dalam kurun waktu enam bulan dari Januari-Juni 2023 dengan kata kunci “Transjakarta”. Dari data tweet yang didapatkan dalam kurun waktu tersebut terdapat berbagai macam sentimen, baik sentimen positif, netral maupun negatif. Kelas sentimen yang paling dominan adalah kelas netral, diikuti kelas negatif kemudian kelas positif. Menggunakan *WordCloud* didapatkan beberapa kata positif yang muncul dengan dominan di *dataset* tersebut. Kata-kata tersebut yaitu “bagus”, “baik”, “cepat”, “keren”, “nyaman”, “murah”. Hal tersebut menunjukkan sentimen positif masyarakat mengenai moda transportasi Transjakarta. Meskipun begitu terdapat sentimen negatif masyarakat mengenai moda transportasi Transjakarta. Hal itu ditunjukkan dari kata-kata negatif yang muncul dalam dataset tersebut. Kata-kata tersebut adalah “bodoh”, “lama”, “penuh”, “rusak”, “jelek”, “macet”.

Selain mengetahui pendapat masyarakat mengenai moda transportasi Transjakarta, penelitian ini juga melakukan analisis sentimen menggunakan metode *machine learning* dan membandingkan tiga model algoritma untuk mengetahui algoritma terbaik untuk penelitian ini. Algoritma *machine learning* tersebut yaitu *naive bayes*, *maximum entropy* dan *support vector machine*. Dilakukan pemodelan dengan ketiga algoritma tersebut dan dihitung nilai *confusion matrix* masing-masing algoritma untuk dapat dievaluasi. Peneliti mencari nilai *precision*, *recall*, *F1-score* dan *accuracy* untuk mendapatkan algoritma terbaik dari ketiga algoritma yang digunakan.

Setelah dilakukan evaluasi, algoritma *maximum entropy* mendapatkan nilai *precision* sebesar 70%, *recall* 70%, *F1-score* 70% dan memberikan *accuracy* sebesar 70%. Algoritma *naive bayes* mendapatkan nilai *precision* sebesar 67%, *recall* 64%, *F1-score* 64% dan memberikan *accuracy* sebesar 64%. Algoritma *support vector machine* mendapatkan nilai *precision* sebesar 72%, *recall* sebesar 71%, *F1-score* sebesar 71% dan *accuracy* sebesar 71%. Pada penelitian ini, dapat disimpulkan bahwa *support vector machine* merupakan algoritma yang paling baik untuk digunakan karena menghasilkan hasil yang terbaik dengan *precision* sebesar 72%, *recall* sebesar 71%, *F1-score* sebesar 71% dan akurasi sebesar 71%.

*Support vector machine* menjadi algoritma yang memiliki akurasi paling tinggi dibanding dua algoritma lainnya karena memiliki prinsip kerja yang berbeda. Algoritma *maximum entropy* dan *naive bayes* adalah algoritma jenis *probabilistic classifier* sementara *support vector machine* adalah *linear classifier*. Prinsip kerja *support vector machine* yang memaksimalkan margin dengan membentuk *hyperplane* pemisah membuat *support vector machine* menunjukkan kinerja terbaik. Dengan fungsi kernel *support vector machine* mampu menjadi algoritma yang tidak kaku yang mampu memahami hubungan antar kata. Hal ini yang membuat *support vector machine* memiliki keunggulan dibanding kedua algoritma lainnya.



## 5.2 Saran

Penelitian ini dapat ditingkatkan karena data yang digunakan hanya terbatas dengan data dari awal tahun 2023 sehingga akan lebih baik jika penelitian ini dilanjutkan dengan penambahan data. Penelitian yang dilakukan saat ini juga dilakukan dengan tiga algoritma *machine learning* yang masih sangat bisa dikembangkan lebih jauh, melihat banyak sekali opsi algoritma *machine learning* untuk analisis sentimen lainnya. Hasil evaluasi yang didapatkan saat ini juga dapat ditingkatkan di penelitian lainnya dengan melakukan berbagai macam eksplorasi metode yang mungkin lebih efektif dan efisien. Komparasi antar kelas data sentimen juga menjadi hal yang perlu diperhatikan dalam penelitian kedepannya.

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
- b. Pengutipan tidak merugikan kepentingan yang wajar IPB University

## DAFTAR PUSTAKA

- Abe S. 2010. *Support Vector Machines for Pattern Classification*. New York: Springer.
- Aftab F, Bazai SU, Marjan S, Baloch L, Aslam S, Amphawan A, Neo T-K. 2023. A Comprehensive Survey on Sentiment Analysis Technique. *International Journal of Technology*. 14(6):1288-1298.
- Alghoniyyu F, Zyiafira F, Junita FM, Sa'adah AN. 2025. Pengaruh Konten Media Sosial Instagram @buswayfansclub Terhadap Pemenuhan Kebutuhan Informasi Seputar Transjakarta (Survei Followers Instagram @buswayfansclub). *Journal of Social and Economic Research*. 7(1):470-477.
- Ali H, Salleh MNM, Hussain K, Ahmad A, Ullah A, Muhammad A, Naseem R, Khan M. 2019. A Review on Data Preprocessing Methods for Class Imbalance Problem. *International Journal of Engineering & Technology*. 8(3):390-397.
- [BPS] Badan Pusat Statistik. 2023. Statistik Telekomunikasi Indonesia 2022. Jakarta: BPS.
- [BPS] Badan Pusat Statistik. 2024. Perkembangan Transportasi DKI Jakarta Juni 2024. Jakarta: BPS.
- [BPS] Badan Pusat Statistik. 2024. Statistik Kesejahteraan Rakyat Provinsi DKI Jakarta 2024. Jakarta: BPS.
- Bishop CM. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.
- Chairunnissa QA, Herdiyeni Y, Hardhienata MKD, Adisantoso J. 2022. Analisis Sentimen Pengguna Twitter Terhadap Program Vaksinasi Covid-19 di Indonesia Menggunakan algoritma Support Vector Machine. *Jurnal Ilmu Komputer Agri-Informatika*. 9(1):79-89.
- Cindo M, Rini DP, Ermatita. 2020. Sentiment Analysis on Twitter by Using Maximum Entropy and Support Vector Machine Method. *Sinergi*. 24(2):87-94.
- Dang NC, Moreno-Garcia MN, De la Prieta F. 2020. Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics*. 9(483):1-29.
- Danilo D, Helaoui R, Kumar V, Recupero DR, Riboni D. 2021. TF-IDF vs Word Embeddings for Morbidity Identification in Clinical Notes: An Initial Study. *ArXiv*. 1-12.
- Dharmawan H. 2022. Hubungan Antara Integrasi Layanan Paratransit Terhadap Jumlah Pengguna Bus Rapid Transit: Studi Kasus Mikrotrans Transjakarta. *Jurnal Transportasi Multimoda*. 20(2022):19-25.
- Diya W and Yixi Z. 2019. Using News to Predict Investor Sentiment: Based on SVM Model. *Procedia Computer Science*. 174(2020):191-199.
- Garcia S, Luengo J, Herrera F. 2014. *Data Preprocessing in Data Mining*. Cham: Springer
- Go A, Bhayani R, Huang L. 2009. Twitter Sentiment Classification using Distang Supervision. *Stanford Project Report CS224N*. 1(12).

- Haddi E, Liu X, Shi Y. 2013. The Role of Text Pre-proccesing in Sentiment Analysis. *Procedia Computer Science*. 17(2013):26-32.
- Han K-X, Chien W, Chiu C-C, Cheng Y-T. 2020. Application of Support Vector Machine (SVM) in the Sentiment Analysis of Twitter DataSet. *Appl. Sci.* 10(3):1125.
- Hussein D. 2016. A Survey on Sentiment Analysis Challenges. *Journal of King Saud University – Engineering Sciences*. 30(4):330-338.
- Jwandini I, Triayudi A, Soepriyono. 2023. Analisa Sentimen Pengguna Transportasi Jakarta Terhadap Transjakarta Menggunakan Metode Naive Bayes dan K-Nearest Neighbor. *Journal of Information System Reasearch*. 4(2):543-550.
- Jurafsky D, Martin JH. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. Ed ke-3. [diakses 2025 Agu 10]. <https://web.stanford.edu/~jurafsky/slp3>.
- Kim JH. 2009. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics and Data Analysis*. 53(2009):3735-3745.
- Kohavi R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Di dalam: Proceedings of Fourteenth International Joint Conference on Artificial Intelligence (IJCAI). Montreal, CA. hlm 1137-1143.
- Liu B. 2012. *Sentiment Analysis and Opinion Mining*. California:Morgan & Claypool.
- Luqyana WA, Cholissodin I, Perdana RS. 2018. Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Universitas Brawijaya*. 2(11): 4704-4713.
- Manning CD, Schutze H. 1999. *Foundations of Statistical Natural Language Processing*. London: MIT Press
- Manning CD, Raghavan P, Schütze H. 2009. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
- Muzaki A, Witanti A. 2021. Sentiment Analysis of The Community in The Twitter to The 2020 Election in Pandemic Covid-19 by Method Naive Bayes Classifier. *Jurnal Teknik Informatika(Jutif)*. 2(2):101-107.
- Nguyen H, Veluchamy A, Diop M, Iqbal R. 2018. Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches. *SMU Data Science Review*. 1(4):1-22.
- Novantirani A, Sabariah MK, Effendy V. 2015. Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine. *e-Proceeding of Engineering*. hlm 1177-1183.
- Nurjannah M, Hamdani, Astuti IF. 2013. Penerapan Algoritma Term Frequency-Inverse Document Frequency (TF-IDF) Untuk Text Mining. *Jurnal Informatika Mulawarman*. 8(3):110–113.
- Nurlaela S, William A. 2023. TransJakarta Service Evaluation in Controlling COVID-19 Transmission Using Twitter Sentiment Analysis. *Journal of Regional and City Planning*. 34(2):156-174.

- Pachouly J, Ahirrao S, Kotecha K, Selvachandran G, Abraham A. 2022. A Systematic Literature Review on Software Defect Prediction Using Artificial Intelligence: Datasets, Data Validation Methods, Approaches, and Tools. *Engineering Application of Artificial Intelligence*. 111:104773.
- Pang B, Lee L. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*. 2(1-2):1-135.
- Patel B, Shah D. 2013. Significance of Stopword Elimination in Meta Search Engine. Di dalam: International Conference on Intelligent System and Signal Processing. ISSP, 52-55.
- Prabantari BVK. 2020. Analisis Hubungan Kualitas Pelayanan Terhadap Tingkat Kepuasan Pelanggan Transportasi Transjakarta. *Jurnal Transaksi* 12(1):25-39.
- Pristanto A, Lauren OT, Aryani D, Sahara S. 2023. Analisis Kepuasan Pelanggan Terhadap Fasilitas yang Disediakan oleh Pihak Transjakarta. *MRI: Jurnal Manajemen Riset Inovasi*. 1(3):9-17.
- Qadri M. 2020. Pengaruh Media Sosial Dalam Membangun Opini Publik. *Qaumiyyah: Jurnal Hukum Tata Negara*. 1(1):49-63.
- Rahadianto NA, Maarif S, Yuliaty LN. 2019. Analysis of Intention to Use Transjakarta Bus. *Independent Journal of Management & Production (IJM&P)*. 10(1):301-324.
- Rajaraman A, Ullman JD. 2011. *Mining of Massive Datasets*. Cambridge:Cambridge University Press.
- Sahara S, Ferdiansyah A. 2023. Pengaruh Keamanan Halte Transjakarta Terhadap Kenyamanan Pelanggan Bus TransJakarta (Studi Penelitian Halte UNJ). *INNOVATIVE: Journal of Social Science Research*. 3(6):7806-7814.
- Samuels A, Mcgonical J. 2007. Sentiment Analysis on Social Media Content. *arXiv*.02338.
- Sammut C, Webb GI. 2010. *Encyclopedia of Machine Learning*. New York: Springer
- Tan PN, Steinbach M, Kumar V. 2006. *Introduction to Data Mining*. Boston: Pearson Addison Wesley
- Utomo M. 2013. Implementasi Stemmer Tala pada Aplikasi Berbasis Web. *Jurnal Teknologi Informasi DINAMIK*. 18(1):41-45.
- Xie X, Ge S, Hu F, Xie M, Jiang N. 2017. An improved algorithm for sentiment analysis based on maximum entropy. *Soft Comput*. 23(2019):599-611.
- Zhang X, Zheng X. 2016. Comparison of Text Sentiment Analysis based on Machine Learning. Di dalam: 15th International Symposium on Parallel and Distributed Computing (ISPDC), 230-233.

