

**SKRIPSI**

**PERBANDINGAN METODE NAÏVE BAYES, KNN DAN SVM  
PADA ANALISIS SENTIMEN JASA TRANSPORTASI ONLINE**

**Disusun dan diajukan oleh :**

**AYU ADHE PUTRI**

**D121181027**



**PROGRAM STUDI SARJANA TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK**

**UNIVERSITAS HASANUDDIN**

**GOWA**

**2023**

**LEMBAR PENGESAHAN SKRIPSI**  
**PERBANDINGAN METODE NAÏVE BAYES, KNN DAN SVM PADA**  
**ANALISIS SENTIMEN JASA TRANSPORTASI ONLINE**

**Disusun dan diajukan oleh**

**AYU ADHE PUTRI**

**D121181027**

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Program Studi Teknik Informatika Fakultas Teknik Universitas Hasanuddin pada tanggal 12 April 2023 dan dinyatakan telah memenuhi syarat kelulusan.

Menyetujui,

Pembimbing Utama,

Pembimbing Pendamping,

Dr. Ir. Ingrid Nurtanio, M.T.  
Nip. 19610813 198811 2 001

Dr. Amr/ Ahmad Ilham, S.T., MIT.  
Nip. 19731010 199802 1 001

Ketua Program Studi,



Prof. Dr. H. Indrabayu, S.T., M.Bus.Sys., IPM, ASEAN .Eng.  
Nip. 19750716 2002/2 1 004

## PERNYATAAN KEASLIAN

Yang bertanda tangan di bawah ini:

Nama : Ayu Adhe Putri

NIM : D121181027

Departemen : Teknik Informatika

Jenjang : S1

Menyatakan dengan ini karya tulisan saya berjudul:

“PERBANDINGAN METODE NAÏVE BAYES, KNN DAN SVM PADA  
ANALISIS SENTIMEN JASA TRANSPORTASI ONLINE”

Adalah karya tulisan saya sendiri dan bukan merupakan pengambilalihan tulisan orang lain bahwa skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri.

Apabila di kemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Makassar, 18 April 2023

Yang menyatakan,



Ayu Adhe Putri

## ABSTRAK

**AYU ADHE PUTRI.** *Perbandingan Metode Naïve Bayes, KNN dan SVM pada Analisis Sentimen Jasa Transportasi Online.* (dibimbing oleh Ingrid Nurtanio dan Amil Ahmad Ilham)

Transportasi online sudah menjadi bagian dari kehidupan masyarakat. Dengan banyaknya opini yang beragam serta banyaknya pilihan transportasi online membuat masyarakat kesulitan dalam memilih transportasi online yang unggul. Menilai dari rating aplikasi saja nyatanya tidak cukup karena rating yang diberikan tidak jarang bertolak belakang dengan komentar yang diberikan. Oleh karena itu diperlukan perhatian khusus pada ulasan pengguna dengan memanfaatkan *sentiment analysis*. Tujuan dari penelitian ini adalah untuk menghasilkan rekomendasi transportasi online yang diperuntukkan kepada masyarakat agar dapat dengan mudah memilih transportasi online dengan layanan terbaik berdasarkan komentar pengguna. Penelitian ini menggunakan tiga model klasifikasi yaitu Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM) dan K-Nearest Neighbor (KNN) kemudian membandingkannya untuk mendapatkan model dengan performa terbaik yang dapat digunakan untuk memprediksi sentimen pada setiap ulasan pengguna. Transportasi online yang dianalisa adalah Grab, Gojek dan Maxim karena aplikasi ini adalah yang paling banyak diunduh oleh masyarakat Indonesia. Dari hasil pengujian yang dilakukan SVM memiliki performa paling baik dibandingkan dengan Multinomial NB dan KNN dengan nilai akurasi sebesar 85%. Sedangkan akurasi yang didapatkan untuk model Multinomial NB sebesar 84% dan KNN sebesar 68%. Maka dari itu SVM adalah model yang digunakan untuk memprediksi sentimen pada ulasan untuk menghasilkan sebuah rekomendasi transportasi online. Dari hasil perhitungan Customer Satisfaction yang didapatkan, Maxim memiliki nilai kepuasan pelanggan tertinggi sebesar 76%, kemudian Grab 55% dan Gojek 43%.

Kata Kunci: Analisis Sentimen, Transportasi Online, Multinomial Naïve Bayes, Support Vector Machine, K-Nearest Neighbor.

## ABSTRACT

**AYU ADHE PUTRI.** *Comparison of Naïve Bayes, KNN and SVM on Sentiment Analysis of Online Transportation Service.* (supervised by Ingrid Nurtanio dan Amil Ahmad Ilham)

Online transportation has become a part of our lives. With a lot of opinions and arguments also many online transportation options, it makes it difficult for people to choose superior online transportation. Judging only by rating is in fact not enough because the rating that given by users often contradicts with the comments. Therefore special attention is needed on user reviews by utilizing sentiment analysis. The purpose of this study is to make an online transportation recommendation that is intended for the public to help the community to choose online transportation with the best service based on user reviews. This study uses three classification models which is Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) and then compares them to obtain the model with the best performance that can be used to classify the sentiments in each user review. The online transportation that is analyzed are Grab, Gojek and Maxim because these applications are the most downloaded by people in Indonesia. The results of this study shows that SVM has the best performance compared to Multinomial NB and KNN with an accuracy of 85%. While the accuracy using Multinomial NB is 84% and 68% for KNN. Therefore SVM is used to classify the sentiment on reviews to make an online transportation recommendation. The results of Customer Satisfaction calculation, Maxim is the online transportation with the highest customer satisfaction score 76% then Grab 55% and Gojek 43%.

**Keywords:** Sentiment Analysis, Online Transportation, Multinomial Naïve Bayes, Support Vector Machine, K-Nearest Neighbor.

## DAFTAR ISI

<b>ABSTRAK</b> .....	ii
<b>ABSTRACT</b> .....	iii
<b>DAFTAR ISI</b> .....	iv
<b>DAFTAR GAMBAR</b> .....	vi
<b>DAFTAR TABEL</b> .....	vii
<b>DAFTAR SINGKATAN DAN ARTI SIMBOL</b> .....	viii
<b>DAFTAR LAMPIRAN</b> .....	ix
<b>KATA PENGANTAR</b> .....	x
<b>BAB I PENDAHULUAN</b> .....	1
1.1 Latar Belakang .....	1
1.2 Rumusan Masalah .....	2
1.3 Tujuan Penelitian .....	3
1.4 Manfaat Penelitian .....	3
1.5 Batasan Masalah .....	3
1.6 Sistematika Penulisan .....	3
<b>BAB II TINJAUAN PUSTAKA</b> .....	5
2.1 Transportasi Online .....	5
2.2 Gojek .....	5
2.3 Grab .....	6
2.4 Maxim .....	7
2.5 Web Scraping .....	7
2.6 Text Mining .....	7
2.7 Analisis Sentimen .....	8
2.8 Bag of Words .....	8
2.9 Naïve Bayes Classifier .....	9
2.10 Support Vector Machine .....	10
2.11 K-Nearest Neighbor .....	13
2.12 Confusion Matrix .....	14
2.13 Customer Satisfaction (CSAT) .....	15
<b>BAB III METODOLOGI PENELITIAN</b> .....	16
3.1 Tahapan Penelitian .....	16
3.2 Rancangan Sistem .....	17

3.3	Pengumpulan Data.....	19
3.4	Preprocessing Data.....	19
3.5	Perhitungan Bag of Words.....	22
3.6	Model Klasifikasi dengan Naïve Bayes.....	23
3.7	Model Klasifikasi dengan Support Vector Machine.....	26
3.8	Model Klasifikasi dengan K-Nearest Neighbor.....	31
<b>BAB IV HASIL DAN PEMBAHASAN.....</b>		<b>34</b>
4.1	Pengumpulan Data.....	34
4.2	Preprocessing Data.....	35
4.3	Analisis Sentimen Multinomial Naïve Bayes.....	36
4.4	Analisis Sentimen Support Vector Machine.....	37
4.5	Analisis Sentimen K-Nearest Neighbor.....	38
4.6	Perbandingan Kinerja Model.....	41
4.7	Rekomendasi.....	41
	4.7.1 Analisis Sentimen Gojek.....	42
	4.7.2 Analisis Sentimen Grab.....	42
	4.7.3 Analisis Sentimen Maxim.....	44
	4.7.4 Customer Satisfaction berdasarkan Analisis Sentimen.....	45
	4.7.5 Customer Satisfaction berdasarkan Rating.....	46
	4.7.6 Rekomendasi Transportasi Online.....	47
<b>BAB V KESIMPULAN DAN SARAN.....</b>		<b>48</b>
5.1	Kesimpulan.....	48
5.2	Saran.....	48
<b>DAFTAR PUSTAKA.....</b>		<b>49</b>
<b>LAMPIRAN.....</b>		<b>51</b>

## DAFTAR GAMBAR

Gambar 1 Ilustrasi <i>hyperplane</i> , <i>margin</i> dan <i>support vectors</i> .....	10
Gambar 2 Ilustrasi <i>margin</i> .....	11
Gambar 3 Data <i>non-linear</i> dalam 2D (a) Data <i>non-linear</i> dipisahkan oleh plane dalam 3D (b) .....	11
Gambar 4 Tahapan Penelitian .....	16
Gambar 5 Rancangan Sistem .....	18
Gambar 6 Tahap alur proses Multinomial Naïve Bayes .....	24
Gambar 7 Tahap alur proses Support Vector Machine.....	26
Gambar 8 Tahap alur proses K-Nearest Neighbor.....	31
Gambar 9 Sampel hasil scraping ulasan.....	34
Gambar 10 True Label vs Predicted Label Multinomial Naïve Bayes .....	36
Gambar 11 True Label vs Predicted Label Vector Machine.....	37
Gambar 12 Grafik Accuracy vs K-Value.....	39
Gambar 13 Best Value of K.....	39
Gambar 14 True Label vs Predicted Label K-Nearest Neighbor.....	40
Gambar 15 Sentimen ulasan Gojek.....	42
Gambar 16 Grafik sentimen ulasan Gojek .....	42
Gambar 17 Sentimen ulasan Grab.....	43
Gambar 18 Grafik sentimen ulasan Grab.....	43
Gambar 19 Sentimen ulasan Maxim .....	44
Gambar 20 Grafik sentimen ulasan Maxim .....	45
Gambar 21 Skor CSAT Gojek (AS).....	46
Gambar 22 Skor CSAT Grab (AS) .....	46
Gambar 23 Skor CSAT Maxim (AS).....	46
Gambar 24 Skor CSAT Gojek (Rating).....	47
Gambar 25 Skor CSAT Grab (Rating).....	47
Gambar 26 Skor CSAT Maxim (Rating) .....	47

## DAFTAR TABEL

Tabel 1 Sampel data .....	20
Tabel 2 Sampel data setelah tahapan case folding .....	20
Tabel 3 Sampel data setelah tahapan remove punctuation, number dan whitespace .....	20
Tabel 4 Sampel data setelah tahapan tokenizing .....	21
Tabel 5 Sampel data setelah tahapan stopword removal .....	21
Tabel 6 Sampel data setelah tahapan stemming .....	22
Tabel 7 Sampel data .....	22
Tabel 8 Sampel data hasil BoW .....	22
Tabel 9 Sampel data dengan sentimen .....	25
Tabel 10 Perhitungan $x, xi$ .....	28
Tabel 11 Perhitungan $\ x, xi\ $ .....	28
Tabel 12 Perhitungan RBF .....	28
Tabel 13 Nilai label ( $y$ ) .....	29
Tabel 14 Nilai $y$ tiap pernyataan .....	29
Tabel 15 Nilai Euclidean Distance .....	32
Tabel 16 Urutan Berdasarkan Euclidean Distance .....	33
Tabel 17 Tetangga Terdekat ( $K=3$ ) .....	33
Tabel 18 Voting .....	33
Tabel 19 Rincian dataset .....	35
Tabel 20 Sampel hasil <i>preprocessing</i> ulasan .....	35
Tabel 21 Evaluasi Model Multinomial Naïve Bayes .....	37
Tabel 22 Evaluasi Model Support Vector Machine .....	38
Tabel 23 Evaluasi Model K-Nearest Neighbor .....	40
Tabel 24 Perbandingan Kinerja Model .....	41
Tabel 25 Jumlah sentimen ulasan Gojek .....	43
Tabel 26 Jumlah sentimen ulasan Grab .....	44
Tabel 27 Jumlah sentimen ulasan Maxim .....	45

**DAFTAR SINGKATAN DAN ARTI SIMBOL**

---

Lambang/Singkatan	Arti dan Keterangan
MNB	<i>Multinomial Naïve Bayes</i>
SVM	<i>Support Vector Machine</i>
KNN	<i>K-Nearest Neighbor</i>
AS	<i>Analisis Sentimen</i>
KP	<i>Kepuasan Pelanggan</i>
NBC	<i>Naïve Bayes Classifier</i>
BoW	<i>Bag of Words</i>
OvA	<i>One vs All</i>
RBF	<i>Radial Basis Function</i>
K	Jumlah tetangga terdekat
CSAT	<i>Costumer Satisfaction Score</i>

## DAFTAR LAMPIRAN

<b>Lampiran 1.</b> Preprocessing Data .....	52
<b>Lampiran 2.</b> Class Multinomial Naïve Bayes .....	53
<b>Lampiran 3.</b> Class Support Vector Machine .....	55
<b>Lampiran 4.</b> Class K-Nearest Neighbor .....	58

## KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Allah SWT karena berkat rahmat dan karunia-Nya sehingga skripsi yang berjudul **“Perbandingan Metode Naïve Bayes, KNN dan SVM pada Analisis Sentimen Jasa Transportasi Online”** ini dapat terselesaikan dengan baik sebagai salah satu syarat dalam menyelesaikan jenjang Strata-1 pada Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin.

Penulis menyadari bahwa dalam proses penulisan skripsi ini banyak mengalami kendala dan kesulitan. Namun berkat bantuan, bimbingan dan dukungan dari berbagai pihak serta berkat Allah SWT sehingga kendala-kendala tersebut dapat diatasi. Oleh karena itu, dengan penuh kerendahan hati, penulis mengucapkan terimakasih kepada:

1. Allah SWT atas berkat dan rahmat-Nya sehingga penulis dapat menyelesaikan tugas akhir ini.
2. Kedua orang tua penulis, Bapak Muhammad Zainal dan Ibu Agnes Suriani yang senantiasa mendoakan, memfasilitasi dan memberikan dukungan dalam menyelesaikan perkuliahan.
3. Ibu Dr. Ir. Ingrid Nurtanio, M.T. selaku pembimbing I dan Bapak Dr. Amil Ahmad Ilham, S.T., M.IT. selaku pembimbing II, yang menyediakan waktu, tenaga dan pikiran dalam memberikan masukan dan arahan sampai penulisan skripsi ini selesai.
4. Bapak Prof. Dr. Ir. Indrabayu, S.T., M.Bus.Sys., IPM, ASEAN .Eng. dan Ibu Elly Warni. ST., M.T. selaku dosen penguji sidang skripsi yang telah memberikan kritik dan saran
5. Dosen dan Staff Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin yang telah banyak membantu penulis selama masa perkuliahan.
6. Teman-teman Prodi Teknik Informatika Angkatan 2018 atas dukungan, bantuan dan semangat yang telah diberikan.

7. Rekan-rekan dan semua pihak yang tidak dapat disebutkan namanya satu persatu yang telah membantu dalam penulisan skripsi ini.

Akhir kata, penulis berharap semoga Allah SWT membalas segala kebaikan dari semua pihak yang telah membantu penulis dalam penyusunan skripsi. Semoga skripsi ini dapat bermanfaat serta menambah wawasan ilmu untuk pembaca dan juga bagi penulis sendiri.

Makassar, Maret 2023



Penulis.

Ayu Adhe Putri

# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang

Pengaruh globalisasi dan perkembangan teknologi yang cukup pesat memberikan perubahan kepada pola hidup masyarakat. Banyak bisnis mulai bermunculan dengan memanfaatkan perkembangan teknologi komunikasi tersebut, salah satunya kemunculan penyedia layanan jasa transportasi online. Transportasi online adalah sebuah transportasi yang dimodifikasi dari ojek konvensional yang umumnya berada di suatu pangkalan. Lalu kemudian, transportasi online tersebut dikembangkan dengan beragam pelayanan, seperti pemesanan makanan, belanja, antar jemput barang, dan sebagainya. Menawarkan kepraktisan, kenyamanan dan biaya yang terjangkau membuat masyarakat semakin mengandalkan transportasi online.

Grab, Gojek dan Maxim adalah tiga teratas yang paling banyak digunakan oleh masyarakat Indonesia. Masyarakat perlu selektif dalam memilih layanan jasa transportasi yang dapat menjamin keamanan, kenyamanan serta keselamatan penggunanya. Begitupun dengan perusahaan, sangat penting untuk memperhatikan kualitas pelayanannya, memiliki ciri khas, kemampuan dan kekuatan untuk bersaing dalam menguasai pasar. Dengan banyaknya opini masyarakat yang beragam tentang penggunaan jasa transportasi online tersebut sehingga opini yang ada harus diselaraskan dan diberikan pandangan yang jelas dengan memanfaatkan teknologi. Setiap tanggapan yang diajukan mengandung sentimen yang mengekspresikan perasaan terhadap layanan tersebut. Oleh karena itu diperlukan suatu metode untuk menentukan respon dan tingkat kepuasan pelanggan terhadap layanan yang diberikan. Sehingga masyarakat dapat dengan mudah memilih transportasi online yang akan digunakan. Dan juga dari hasil informasi yang didapatkan bisa membantu perusahaan untuk meminjam kembali kekurangan layanan mereka.

Menurut Iqbal, solusi untuk pengambilan informasi yang berbentuk teks adalah *text mining* (Muttaqina & Kharisudin, 2021). Salah satu proses pengambilan

informasi *text mining* adalah analisis sentimen. Analisis sentimen digunakan untuk melihat kecenderungan suatu sentimen atau pendapat, apakah pendapat tersebut cenderung beropini positif atau negatif. Metode ini dapat digunakan untuk menghasilkan sebuah rekomendasi transportasi online yang digemari masyarakat. Sumber data terbaik untuk mendapatkan dataset dalam penelitian analisis sentimen adalah Google Playstore, dikarenakan data yang diperoleh lebih bersih dan tidak mengandung unsur iklan ataupun promosi.

Pada penelitian yang dilakukan oleh Immanuel (DjajaPutra et al., 2020) menganalisis opini masyarakat terhadap layanan Grab dan Gojek dengan implementasi *text mining* menggunakan algoritma *Naïve Bayes Classifier* (NBC) digunakan untuk mengklasifikasikan data tweet ke dalam kelas sentimen positif dan sentimen negatif. Penelitian yang dilakukan oleh Savira (Rohwinasakti et al., 2021) membuat sistem *sentiment analysis* dengan mengklasifikasikan komentar Instagram menggunakan metode *K-Nearest Neighbor* (KNN) dengan akurasi yang cukup tinggi sebesar 94.44%. Pada penelitian yang dilakukan Iqbal dalam pengklasifikasian ulasan aplikasi Gojek menggunakan metode *Support Vector Machine* (SVM) dan KNN, dari hasil yang didapatkan disimpulkan bahwa metode SVM melakukan klasifikasi lebih baik dibandingkan metode KNN (Muttaqin & Kharisudin, 2021). Oleh karena itu, pada penelitian ini mengembangkan sistem yang serupa dengan membandingkan ketiga metode yaitu NBC, KNN dan SVM. Hasil dari penelitian berupa informasi rekomendasi transportasi online. Rekomendasi ini dapat digunakan oleh masyarakat agar dapat dengan mudah memilih dan menentukan transportasi online yang unggul untuk digunakan.

## 1.2 Rumusan Masalah

1. Bagaimana menggunakan algoritma klasifikasi Naïve Bayes, KNN dan SVM pada analisis sentimen?
2. Bagaimana menghasilkan rekomendasi transportasi online berdasarkan sentimennya?

### **1.3 Tujuan Penelitian**

1. Menentukan metode klasifikasi dengan performa terbaik pada analisis sentimen.
2. Memberikan rekomendasi transportasi online berdasarkan ulasan dari para pengguna.

### **1.4 Manfaat Penelitian**

Penelitian ini diharapkan dapat membantu:

1. Masyarakat, informasi rekomendasi transportasi online yang dihasilkan dapat dijadikan bahan pertimbangan oleh masyarakat dalam memilih jasa transportasi online yang unggul untuk digunakan
2. Bagi institusi pendidikan, penelitian ini diharapkan dapat menjadi referensi bagi akademis maupun peneliti berikutnya dalam melakukan penelitian mengenai analisis sentiment

### **1.5 Batasan Masalah**

Batasan masalah dalam penelitian ini dapat dijabarkan menjadi beberapa poin sebagai berikut:

1. Dataset terdiri dari kumpulan review dari pengguna aplikasi Grab, Gojek dan Maxim yang diperoleh dari scraping data pada Google Play Store
2. Dataset yang digunakan menggunakan teks Bahasa Indonesia
3. Metode klasifikasi yang digunakan adalah Naïve Bayes, KNN dan SVM

### **1.6 Sistematika Penulisan**

Untuk memberikan gambaran singkat mengenai isi tulisan secara keseluruhan, maka tugas akhir ini diuraikan kedalam beberapa tahapan penulisan, yaitu:

## BAB I PENDAHULUAN

Bab ini membahas mengenai latar belakang, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah serta sistematika penulisan.

## BAB II TINJAUAN PUSTAKA

Bab ini berisikan sejumlah kajian pustaka yang berhubungan dengan topik penelitian dan teori-teori yang menunjang percobaan penelitian yang dilakukan.

## BAB III METODOLOGI PENELITIAN

Bab ini memberikan gambaran mengenai alur dan rancangan sistem serta berisikan tentang metode-metode yang akan digunakan pada penelitian.

## BAB IV HASIL DAN PEMBAHASAN

Bab ini memaparkan hasil penelitian serta pembahasan atau evaluasi dari penelitian yang dilakukan secara keseluruhan.

## BAB V PENUTUP

Bab ini memberikan kesimpulan dari hasil penelitian serta saran untuk pengembangan sistem lebih lanjut.

## **BAB II**

### **TINJAUAN PUSTAKA**

#### **2.1 Transportasi Online**

Transportasi online adalah pelayanan jasa transportasi yang berbasis internet dalam setiap kegiatan transaksinya, mulai dari pemesanan, pemantauan jalur, pembayaran dan penilaian terhadap pelayanan jasa itu sendiri. Di Indonesia, bisnis transportasi online sebenarnya sudah ada sejak tahun 2010, lalu seiring dengan perkembangan teknologi komunikasi, bisnis transportasi online berbasis *smartphone* baru mengalami perkembangan pada tahun 2015. Pada tahun 2018 pengguna rata-rata harian layanan transportasi online 8juta/hari di wilayah Asia Tenggara dan negara Indonesia merupakan negara terbesar dibanding dengan negara Asia Tenggara lainnya (Rohwinasakti et al., 2021). Masyarakat yang sudah jengah dengan transportasi konvensional merasa terbantu dan dimudahkan dengan adanya transportasi berbasis aplikasi online. Seiring dengan perkembangan teknologi dan kreativitas manusia, transportasi online kemudian dikembangkan dengan beragam pelayanan seperti antar jemput barang, pemesanan makanan, belanja di mini market dan sebagainya.

#### **2.2 Gojek**

Di Indonesia, Gojek adalah pencetus pertama yang menghadirkan aplikasi ojek online sebelum pesaing bermunculan. Gojek merupakan perusahaan decacorn karya anak bangsa yang didirikan pada tahun 2010 di Jakarta oleh Nadiem Makarim (Muttaqin & Kharisudin, 2021). Ide mendirikan Gojek muncul dari pengalaman pribadi Nadiem menggunakan transportasi ojek yang hampir setiap hari ke tempat kerjanya untuk menembus kemacetan di Jakarta. Nadiem melihat ternyata sebagian besar waktu yang dihabiskan oleh pengemudi ojek hanyalah sekedar mangkal menunggu penumpang. Dari pengalamannya tersebut, Nadiem melihat adanya peluang untuk membuat sebuah layanan yang dapat menghubungkan penumpang dengan pengemudi ojek.

Pada tanggal 5 Oktober 2009, Gojek resmi berdiri dengan 20 orang pengemudi. Pada saat itu, Gojek masih mengandalkan call center untuk menghubungkan penumpang dengan pengemudi ojek. Pada pertengahan 2014, berkat popularitas Uber kala itu, Nadiem Makarim mulai mendapatkan tawaran investasi. Pada tanggal 7 Januari 2015, Gojek akhirnya meluncurkan aplikasi berbasis Android dan iOS untuk menggantikan sistem pemesanan menggunakan call center. Gojek memiliki tujuan mendorong perubahan agar sektor informal seperti ojek agar yang tadinya bekerja serabutan dengan pendapatan yang tidak menentu bisa beroperasi secara profesional dengan pendapatan yang lebih baik.

### **2.3 Grab**

Grab atau yang sebelumnya dikenal sebagai GrabTaxi adalah sebuah perusahaan asal Malaysia yang bermarkas di Singapura. Anthony Tan bersama rekannya Tan Hooi merupakan nama dibalik keberhasilan layanan jasa transportasi online Grab. Grab didirikan pada tahun 2011. Saat ini Grab telah beroperasi di sejumlah negara di Asia Tenggara seperti Malaysia, Singapura, Thailand, Vietnam, Indonesia dan Filipina. Indonesia merupakan pasar terbesar Grab berdasarkan jumlah perjalanan yang diselesaikan. Layanan Grab ditujukan untuk memberikan alternatif berkendara bagi para pengemudi dan penumpang yang menekankan pada kecepatan, keselamatan dan kepastian. Saat ini, Grab sendiri mempunyai sejumlah layanan mulai dari Grab Express, Grab Car, Grab Bike, Grab Food, Grab Taxi dan layanan jasa lainnya.

Beberapa tahun terakhir, Grab sangat berambisi untuk memberikan layanan jasa di semua wilayah Asia. Negara Indonesia sendiri merupakan salah satu tujuan utama dari perusahaan transportasi ini. Grab berhasil masuk ke wilayah Indonesia pada pertengahan tahun 2014 lalu. Walaupun kedatangannya membawa pro dan kontra, tetapi kenyataannya Grab kini telah mampu diterima dengan baik oleh warga Indonesia. Tak hanya itu, para pelaku usaha dalam negeri pun saat ini telah banyak melakukan Kerjasama dengan Grab. Dengan adanya Grab diharapkan dapat membuat perkembangan ekonomi masyarakat menjadi semakin maju.

## 2.4 Maxim

Maxim merupakan perusahaan transportasi online asal Rusia, tepatnya di Chardnisk, pegunungan Ural yang sudah beroperasi sejak tahun 2003. Pada tahun 2014, Maxim membuka cabang diluar Rusia, yakni di Ukraina, Kazakhstan, Georgia, Bulgaria, Belarusia, Azerbaijan hingga Italia. Hingga sejak tahun 2018, Maxim telah beroperasi di Indonesia. Selama 4 tahun beroperasi, layanan Maxim dapat dinikmati di lebih dari 70 kota di Indonesia. Kantor pusat Maxim di Indonesia berada di Tebet, Jakarta Selatan.

Perusahaan teknologi internasional ini menawarkan layanan transportasi online (motor dan mobil) serta menawarkan layanan tambahan seperti pesan-antar makanan dan barang, kargo, jasa pembersih dan laundry. Tidak seperti pesaingnya yang berwarna hijau, Maxim identik dengan warna kuning mencolok pada helm dan jaketnya. Selain itu, tarif yang mereka tetapkan pun terbilang lebih rendah. Maxim hanya mematok Rp.3.000 per kilometer. Tarif ini jauh lebih rendah dari Peraturan Menteri Perhubungan (Permenhub) No 348 sejak 1 Mei 2019, yang menetapkan tarif minimal Rp.7.000 hingga Rp.10.000 per kilometer. Seperti yang tertera pada halaman websitenya Misi Maxim yakni meningkatkan interaksi secara terus menerus diantara para pengguna dan membantu banyak orang untuk melakukan perjalanan ke tujuan masing-masing.

## 2.5 Web Scraping

*Web scraping* dapat didefinisikan sebagai proses pengambilan data atau ekstraksi dari sebuah website lalu kemudian disimpan dalam sebuah format tertentu. Pada penelitian ini *web scraping* sangat efisien untuk mengumpulkan informasi data komentar pengguna dari beberapa aplikasi transportasi online karena prosesnya dalam mengumpulkan data terbilang cepat apalagi untuk jumlah data yang besar serta mampu mengumpulkan data secara detail.

## 2.6 Text Mining

*Text mining* adalah proses pengolahan data yang berbentuk teks. *Text mining* dapat menghasilkan informasi atau intisari melalui pemrosesan, pengelompokan

dan analisis data-data tidak terstruktur dalam jumlah besar (Prabowo & Wiguna, 2021). *Text mining* sebenarnya merupakan subtype dari *data mining*, bedanya pada *teks mining* berfokus pada teks, sumber datanya tidak terstruktur (*unstructured*), sebaliknya pada *data mining*, bentuk datanya lebih terstruktur (*structured data*). Permasalahan yang dihadapi pada *text mining* sama dengan permasalahan yang terdapat pada *data mining*, yaitu jumlah data yang besar, dimensi yang tinggi, data dan struktur yang berubah serta data noise.

## 2.7 Analisis Sentimen

Salah satu analisis pada *text mining* adalah analisis sentimen atau sering juga disebut dengan *opinion mining*. Analisis sentimen digunakan untuk melihat kecenderungan suatu sentimen atau pendapat, apakah pendapat tersebut cenderung beropini positif atau negatif (Muttaqin & Kharisudin, 2021). Ada 2 cara yang umumnya digunakan untuk menentukan apakah suatu teks mengandung kalimat opini atau tidak. Cara pertama adalah dengan membaca isi teks secara keseluruhan, kemudian memberikan label secara manual. Cara ini memiliki presisi yang cukup tinggi tapi sangat sulit diterapkan untuk data yang berjumlah sangat besar karena memakan banyak waktu. Cara kedua adalah dengan menerapkan metode yang dapat mengekstrak kalimat opini secara otomatis, yaitu dengan analisis sentimen. Pada penelitian ini metode yang digunakan untuk analisis sentimen antara yaitu Naïve Bayes, Support Vector Machine (SVM) dan K-Nearest Neighbor.

## 2.8 Bag of Words

Bag of words (BoW) merupakan salah satu metode paling sederhana dalam mengubah data teks menjadi vector yang dapat dipahami oleh komputer. Metode ini hanya menghitung frekuensi kemunculan kata pada seluruh dokumen. Model Bow umumnya digunakan dalam metode klasifikasi dokumen, dimana frekuensi kemunculan setiap kata digunakan sebagai fitur untuk training klasifikasi. Nilai frekuensi ini disimpan dalam bentuk vektor, dimana setiap dimensi mewakili sebuah kata dalam korpus. Sehingga vektor ini adalah kumpulan kata-kata yang mewakili keseluruhan teks (Schoot Uiterkamp, 2019).

## 2.9 Naïve Bayes Classifier

Naïve Bayes adalah model klasifikasi menggunakan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes. Naïve Bayes memprediksi peluang dimasa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal dengan *Teorema Bayes*. Metode ini berpotensi baik dalam klasifikasi dalam hal presisi dan komputasi data (Prabowo & Wiguna, 2021). Naïve Bayes menerapkan teknik *supervised* klasifikasi objek di masa depan dengan menetapkan label kelas ke *instance*/catatan menggunakan probabilitas bersyarat. Probabilitas bersyarat adalah ukuran peluang suatu peristiwa yang terjadi berdasarkan peristiwa lain yang telah terjadi.

Perhitungan probabilitas dalam Naïve Bayes menggunakan pendekatan teorema Bayes menggunakan persamaan (1):

$$P(\mathbf{A} | \mathbf{B}) = \frac{P(\mathbf{B} | \mathbf{A}) P(\mathbf{A})}{P(\mathbf{B})} \quad (1)$$

Dimana:

$P(\mathbf{A} | \mathbf{B})$  / *Posterior* : Probabilitas A terjadi jika B terjadi

$P(\mathbf{B} | \mathbf{A})$  / *Likelihood* : Probabilitas B terjadi jika A terjadi

$P(\mathbf{A})$  / *Prior* : Peluang terjadinya A

$P(\mathbf{B})$  / *Evidence* : Peluang terjadinya B

Pada penelitian yang akan dilakukan menggunakan algoritma Multinomial Naïve Bayes. Pemilihan algoritma ini sesuai dengan jenis data yang ingin diolah yaitu data teks. Multinomial Naïve Bayes akan memilih kelas dengan nilai probabilitas tertinggi yang dirumuskan dengan persamaan

$$\hat{P}(t|c) = \frac{T_{ct}+1}{(\sum_{t \in V} T_{ct'})+|V|} \quad (2)$$

Dimana:

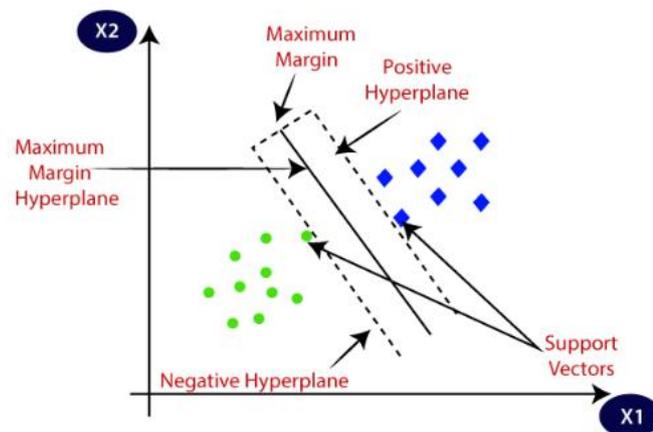
$T_{ct}$  = jumlah kemunculan  $t$  dalam dokumen data *training* kelas  $c$

$|V|$  = jumlah kata unik pada keseluruhan dokumen

Pada persamaan tersebut digunakan *add-one* dengan menambahkan 1 untuk menghindari hasil probabilitas sama dengan 0 dari hasil perkalian *conditional probability* yang dikarenakan tidak terdapat kata  $t$  dalam satu kelas.

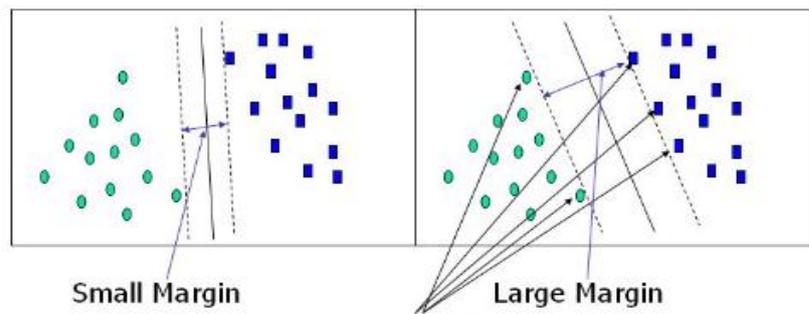
## 2.10 Support Vector Machine

Support Vector Machine (SVM) merupakan salah satu algoritma *machine learning* dengan pendekatan berbasis *supervised learning* yang digunakan untuk melakukan prediksi, baik dalam kasus regresi maupun klasifikasi (Muttaqin & Kharisudin, 2021). Dalam SVM ada tiga hal yang harus ditentukan dari data yang ada yaitu *support vectors*, *hyperplane* dan *margin*. Prinsip kerja SVM yaitu mencari *hyperplane* yang optimal dengan margin maksimum yang bertindak sebagai batas keputusan (*decision boundary*) untuk memisahkan dua buah kelas atau lebih pada *input space* (Goh & Lee, 2019). *Hyperplane* dapat berupa *line* atau garis pada dua dimensi dan dapat berupa *flat plane* pada *multiple plane*. Prinsip dasar SVM adalah *linear classifier*, dan selanjutnya dikembangkan agar dapat bekerja pada *problem non-linear* dengan memasukkan konsep *kernel trick* pada ruang kerja berdimensi tinggi (Pisner & Schnyer, 2019).



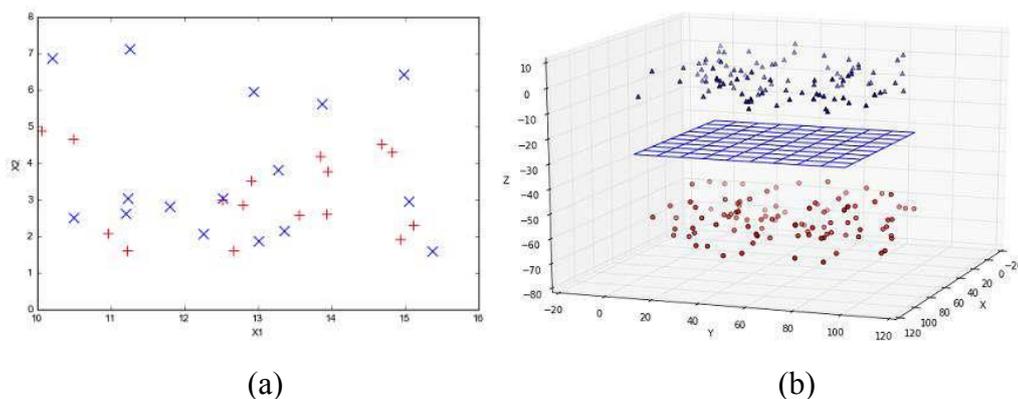
Gambar 1 Ilustrasi *hyperplane*, *margin* dan *support vectors* (Pisner & Schnyer, 2019)

*Support Vector* merupakan titik data atau vektor yang paling dekat dengan *hyperplane* dan yang mempengaruhi posisi *hyperplane*. Karena vektor-vektor ini mendukung *hyperplane*, maka disebut *support vector*.



Gambar 2 Ilustrasi margin (Pisner & Schnyer, 2019)

Margin adalah jarak *support vector* dari masing-masing kelas di sekitar *hyperplane*. Pada gambar 2.10.3 margin diilustrasikan dengan jarak antara 2 garis putus. Margin terbesar (*max margin*) dapat ditemukan dengan memaksimalkan nilai jarak antara *hyperplane* dan titik terdekatnya. Dapat dilihat pada gambar bagian kanan memiliki margin lebih besar daripada gambar bagian kiri. Secara intuitif, margin yang lebih besar akan menghasilkan performa klasifikasi yang lebih baik.



Gambar 3 Data *non-linear* dalam 2D (a) Data *non-linear* dipisahkan oleh plane dalam 3D (b) (Goh & Lee, 2019)

Terdapat beberapa kasus dimana kelas yang ada tidak bisa dipisahkan dengan memanfaatkan garis linear. Untuk mengatasi kondisi tersebut, SVM akan memproyeksikan data yang ada ke dimensi yang lebih tinggi dan menerapkan *decision boundary* dalam bentuk bidang datar yang disebut dengan *plane*.

Pencarian lokasi *hyperplane optimal* merupakan inti dari metode SVM. Diasumsikan bahwa terdapat data *learning* dengan *points*  $x_i$  ( $i = 1, 2, \dots, m$ ) dan memiliki dua kelas  $y_i = \pm 1$  yaitu kelas positif (+1) dan negatif (-1) sehingga akan diperoleh *decision function* berikut.

$$f(x) = \text{sign}(w \cdot x + b) \quad (3)$$

Dengan  $(\cdot)$  adalah *scalar* sehingga

$$w \cdot x = w^T x \quad (4)$$

*Margin maximum* dapat diperoleh dengan cara memaksimalkan nilai jarak antara *hyperplane* dan titik terdekatnya yaitu  $\frac{1}{\|w\|}$ . Hal ini dirumuskan sebagai QP (*Quadratic Programming Problem*) dengan mencari titik minimal seperti pada persamaan berikut.

$$\min \tau(w) = \frac{1}{2} \|w\|^2 \quad (5)$$

Sedangkan untuk subjek *constrain* atau kendala persamaannya adalah:

$$y_i(w \cdot x_i + b) \geq 1 \forall_i \quad (6)$$

Persamaan di atas merupakan permasalahan optimisasi kendala dimana diminimalkan fungsi objek persamaan (5) dengan kendala pada persamaan (6). Permasalahan tersebut dapat direduksi dengan menggunakan fungsi *Langrange* yang terdiri dari jumlahan fungsi objektif dan m kendala dikalikan dengan pengganda *Langrange* seperti berikut (Campbell & Yiming, 2011).

$$L(w, b) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^m \alpha_i (y_i(w \cdot x_i + b) - 1) \quad (7)$$

Dimana  $\alpha_i$  merupakan *Langrange Multipliers* dan nilai  $\alpha_i \geq 0$ .

Untuk memetakan dimensi awal (dimensi yang lebih rendah) ke dimensi baru (dimensi yang lebih tinggi) digunakan fungsi kernel. Macam-macam fungsi kernel adalah sebagai berikut (Aggarwal, 2018).

1. Kernel Linear

$$K(x, y) = x \cdot y \quad (8)$$

2. Kernel *Gaussian Radial Basic Function* (RBF)

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (9)$$

3. Kernel *Polynomial*

$$K(x_i, x_j) = ((x_i \cdot x_j) + c)^d \quad (10)$$

Dengan  $x_i$  dan  $x_j$  merupakan pasangan dua kata *training*. Parameter  $\sigma, c, d > 0$  merupakan konstanta.

## 2.11 K-Nearest Neighbor

Algoritma K-NN merupakan salah satu algoritma paling populer dalam machine learning karena prosesnya mudah dan sederhana. Algoritma klasifikasi ini bekerja dengan mengambil sejumlah K data terdekat (tetangganya) sebagai acuan untuk menentukan kelas dari data baru. Dalam menentukan nilai K dapat dilakukan dengan menggunakan metode *Elbow* untuk mendapatkan nilai K optimal. Perhitungan jarak dilakukan dengan konsep *Euclidean*. Jumlah kelas yang paling banyak dengan jarak terdekat akan menjadi kelas dimana data uji berada. Algoritma ini mengklasifikasikan data berdasarkan *similarity* atau kemiripan atau kedekatannya terhadap data lainnya (Cunningham & Delany, 2021).

Adapun langkah-langkah dari algoritma K-NN adalah:

1. Tentukan parameter K
2. Hitung jarak antara data uji dengan data latih dengan menggunakan *euclidean distance* seperti pada persamaan (11).

$$d(x_i, y_i) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (11)$$

Perhitungan *euclidean distance* adalah untuk menentukan jarak antara dua titik pada data latih dan data uji

Keterangan:

$x_i$  = data training

$y_i$  = data testing

$d(x_i, y_i)$  = jarak

$i$  = data ke- $i$

$n$  = jumlah data

3. Urutkan jarak yang terbentuk dengan cara *ascending*. Perhitungan tiap dokumen pada data latih dan data uji didapatkan dari perhitungan ekstraksi ciri tiap kata yang telah dinormalisasi.
4. Menggunakan parameter K untuk menentukan jumlah tetangga terdekat.
5. Memilih jumlah kelas terbanyak dari tetangga terdekat lalu di klasifikasikan.

## 2.12 Confusion Matrix

*Confusion matrix* adalah cara yang dapat digunakan untuk mengevaluasi kinerja model klasifikasi. Setiap baris pada confusion matrix mewakili kelas yang sebenarnya, sedangkan setiap kolom mewakili kelas yang diprediksi (Géron, 2019). Ada empat istilah yang merupakan representasi hasil proses klasifikasi pada confusion matrix yaitu *True Positive*, *True Negative*, *False Positive* dan *False Negative*.

True Positive (TP) : Prediksi kelas positif dan benar

True Negative (TN) : Prediksi kelas negatif dan benar

False Positive (FP) : Prediksi kelas positif dan salah

False Negative (FN) : Prediksi kelas negatif dan salah

- Precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (12)$$

Dari persamaan diatas dapat disimpulkan dengan, dari semua kelas yang diprediksi positif, berapa banyak yang benar-benar positif. Sehingga presisi merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif.

- Recall

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (13)$$

Dari persamaan diatas dapat disimpulkan dengan, dari semua kelas positif berapa banyak yang diprediksi dengan benar. *Recall* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif.

- Accuracy

$$\text{Accuration} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (14)$$

Dari semua kelas (positif dan negatif), berapa banyak yang telah diprediksi dengan benar. Dalam hal ini akurasi merupakan tingkat kedekatan antara nilai prediksi dengan nilai aktual.

- F-measure

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (15)$$

$$\text{F}\beta \text{ score} = (1 + \beta^2) \frac{\text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}} \quad (16)$$

Secara definisi, *F1-score* adalah *harmonic mean* dari *precision* dan *recall*. Jika *F1-score* memiliki skor yang baik mengindikasikan bahwa model klasifikasi kita punya *precision* dan *recall* yang baik. Sedangkan, *F-beta score* digunakan ketika presisi dan *recall* sama-sama penting akan tetapi salah satunya lebih penting. Ketika presisi dianggap lebih penting daripada *recall* maka nilai beta < 1. Sebaliknya jika *recall* dianggap lebih penting daripada presisi maka nilai beta > 1. Apabila keduanya sama-sama penting dan seimbang maka nilai beta = 1. Penggunaan jenis *F-measure* tergantung dengan tujuan klasifikasi dan jenis *error* apa yang lebih *problematic*. *Accuracy* bekerja kurang baik apabila dataset yang digunakan tidak seimbang, sedangkan *F-measure* masih dapat mengukur kinerja secara objektif walaupun kelas tidak seimbang.

### 2.13 Customer Satisfaction (CSAT)

*Customer satisfaction* atau kepuasan pelanggan didefinisikan dengan mengukur bagaimana suatu produk atau jasa memenuhi atau bahkan melebihi harapan klien (Dam & Dam, 2021). Kotler dan Keller mengatakan bahwa kepuasan pelanggan adalah persepsi pelanggan tentang kesenangan atau kekecewaan karena adanya perbandingan produk atau layanan yang diberikan dengan ekspektasi pelanggan (Kotler & Keller, 2016).

$$\text{Skor CSAT} = \frac{\text{total respon positif}}{\text{total respon}} \times 100 \% \quad (17)$$

Skor CSAT dapat diperoleh dengan menghitung jumlah dari semua respon positif yang dibagi dengan total respon kemudian dikalikan dengan 100%. Hasilnya memberikan persentase keseluruhan pelanggan yang puas pada produk atau layanan yang diberikan.