TUTORIAL PENGGUNAAN ANALISIS KOMENTAR E-COMMERCE MENGGUNAKAN ALGORITMA NAÏVE BAYES & KNN

Wahyu Kurnia Sari
Roni Andarsyah
Mohamad Nurkamal Fauzan

TUTORIAL PENGGUNAAN ANALISIS KOMENTAR E-COMMERCE MENGGUNAKAN ALGORITMA NAÏVE BAYES & KNN

Wahyu Kurnia Sari
Roni Andarsyah
Mohamad Nurkamal Fauzan

bukupedia

TUTORIAL PENGGUNAAN ANALISIS KOMENTAR E-COMMERCE MENGGUNAKAN ALGORITMA NAÏVE BAYES & KNN

Penulis :
Wahyu Kurnia Sari
Roni Andarsyah
Mohamad Nurkamal Fauzan

ISBN : -

Editor :
Rolly Maulana Awangga

Penyunting :
Rolly Maulana Awangga

Desain sampul dan Tata letak :
Wahyu Kurnia Sari
Copyright : https://storyset.com/illustration/new-message/bro

Penerbit :
Penerbit Buku Pedia

Redaksi :
Athena Residence Blok. E No. 1, Desa Ciwaruga,
Kec. Parongpong, Kab. Bandung Barat 40559
Tel. 628-775-2000-300
Email : penerbit@bukupedia.co.id

Distributor :
Informatics Research Center
Jl. Sariasih No. 54
Bandung 40151
Email : irc@poltekpos.ac.id

Cetakan Pertama, 2022

KATA PENGANTAR

Allhamdulillah, Puji syukur penulis panjatkan ke hadirat illahi robi atas terselesaikannya buku ini.

Digitalisasi di era tekonologi 4.0 membuat masyarakat harus menyesuaikan zaman dengan hadirnya perkembangan teknologi yang canggih membuat masyarakat Indonesia dapat melakukan komunikasi dengan pihak lain tanpa batasan waktu dan jarak (Hakim, A, 2018). Indonesia mengalami pertumbuhan tahunan dalam *e-commerce.* Pertumbuhan ini tidak terlepas dari besarnya jumlah pengguna internet di Indonesia (Gumilang, Z. A. N., dkk, 2018). *E-commerce* menerapakan system bisnisnya dengan strategi B2C *(Bussines to Consumer).* Internet memungkinkan pengguna untuk membuat ulasan secara online diberbagai jenis platfom. Selain berbagi pengalaman pribadi, pengguna juga menunjukkan emosi atau perasaan mereka dalam *user-generated content* (UGC), seperti ulasan pengguna (Wu and Li 2020). Selain itu, Indonesia juga merupakan negara besar yang memiliki jumlah penduduk yang banyak sehingga membuat perkembangan *e-commerce* menjadi semakin meningkat. Berdasarkan survey yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) pada tahun 2020, pengguna internet di Indonesia mencapai angka 196,7 juta orang yang sebelumnya pada tahun 2018 hanya 171,17 juta jiwa.

Peningkatan jumlah pengunjung dan pengguna baru yang akan mengunduh aplikasi *E-commerce* berkaitan dengan komentar terhadap aplikasi *E-commerce*. Guna meningkatkan kepercayaan, pengguna bisa melihat komentar terhadap aplikasi *E-commerce* juga mampu meningkatkan kualitas terhadap layanan (Faadilah, A, 2020). Pengguna memberikan pendapat mereka tentang aplikasi *E-commerce* disediakan *google play store.* Ulasan tersebut yang dikemas menjadi bentuk komentar. Komentar yang ada meliputi komentar produk positif dan negatif. Oleh karena itu, diperlukan suatu teknik pengolahan data dan analisis terhadap komentar (Ratnawati, F. 2018). Sentimen atau komentar tersebut dapat dianalisis dengan menggunakan text *mining.*

Buku ini sendiri merupakan wujud dari “kepingan-kepingan puzzle”, yang antara lain bersumber dari bahan catatan kuliah, makalah, dan serangkaian diskusi di antara kami (penulis) yang terkait dengan gencaran persaingan *e-commerce* yang ada di Indonesia, sehingga masyarakat bisa menilai analisis dari aplikasi *e-commerce* melalui text mining yang bersumber dari *reviewer* komentar di *google playstore e- commerce*.

Akhir kata, tak ada gading yang tak retak, dan karenanya tidak ada karya yang sempurna. Terlebih, kesempurnaan hanyalah milik Allah Swt semata, sang pemilik ilmu yang sesungguhnya. Harapan kami, semoga buku ini bisa memberikan manfaat sebesar-besarnya bagi para pembaca Generasi Milenial Tantangan Membangun Komitmen Kerja/Bisnis dan Adversity Quotient (AQ) sekalian. Tidak lupa, kami juga senantiasa membuka diri, menanti kritik dan saran yang membangun demi perbaikan buku ini di masa mendatang. Selamat membaca!

Bandung, Agustus 2022

Penulis

DAFTAR ISI

BAGIAN SATU

1.1 PRAKATA

Digitalisasi di era tekonologi 4.0 membuat masyarakat harus menyesuaikan zaman dengan hadirnya perkembangan teknologi yang canggih membuat masyarakat Indonesia dapat melakukan komunikasi dengan pihak lain tanpa batasan waktu dan jarak (Hakim, A, 2018). Indonesia mengalami pertumbuhan tahunan dalam *e-commerce.* Pertumbuhan ini tidak terlepas dari besarnya jumlah pengguna internet di Indonesia (Gumilang, Z. A. N., dkk, 2018). *E-commerce* menerapakan system bisnisnya dengan strategi B2C *(Bussines to Consumer).* Internet memungkinkan pengguna untuk membuat ulasan secara online diberbagai jenis platfom. Selain berbagi pengalaman pribadi, pengguna juga menunjukkan emosi atau perasaan mereka dalam *user-generated content* (UGC), seperti ulasan pengguna (Wu and Li 2020). Selain itu, Indonesia juga merupakan negara besar yang memiliki jumlah penduduk yang banyak sehingga membuat perkembangan *e-commerce* menjadi semakin meningkat. Berdasarkan survey yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) pada tahun 2020, pengguna internet di Indonesia mencapai angka 196,7 juta orang yang sebelumnya pada tahun 2018 hanya 171,17 juta jiwa.

Salah satu perusahaan yang paling berdampak pada kehadiran internet bagi masyarakat adalah *e-commerce* (Rohandi, Mochamad Malik Akbar,2017). Pada saat ini, layanan *e- commerce* sudah ditawarkan diberbagai macam bentuk aplikasi. Sudah banyak bisnis *e-commerce* yang hadir di Indonesia dan *E-commerce* merupakan salah satu diantaranya. Pada tahun 2021, berdasarkan peta *e-commerce* Indonesia,

Peningkatan jumlah pengunjung dan pengguna baru yang berkaitan dengan komentar terhadap aplikasi *e-commerce*. Guna meningkatkan kepercayaan, pengguna bisa melihat komentar terhadap e-commerce juga mampu meningkatkan kualitas terhadap layanan. Pengguna memberikan pendapat mereka tentang aplikasi *E-commerce* disediakan

google play store. Ulasan tersebut yang dikemas menjadi bentuk komentar. Komentar yang ada meliputi komentar produk positif dan negatif. Oleh karena itu, diperlukan suatu teknik pengolahan data dan analisis terhadap komentar (Ratnawati, F. 2018). Sentimen atau komentar tersebut dapat dianalisis dengan menggunakan text *mining.*

Text *mining* diartikan sebagai suatu teknik untuk pengambilan informasi dari sejumlah data dari sebuah topik tertentu yang memiliki kualitas tinggi agar dapat diperoleh data-data permasalahan dalam teks (Ratnawati, F,2018). Penelitian ini menggunakan algoritma yang terdapat dalam teknik klasifikasi yaitu *k-Nearest Neighbor (KNN*) dan *Naïve bayes*. KNN merupakan algoritma yang digunakan sebagai pengklasifikasian suatu objek. terhadap nilai k data latih dengan syarat jaraknya yang terdekat dengan objek, nilai k harus kecil dari jumlah *training set* dan nilai k tidak boleh genap dan lebih dari satu (Hidayat, Assad, 2019). *Naïve bayes* merupakan metoda klasifikasi yang berdasarkan pada teorema *Bayes*.

Berdasarkan uraian di atas, pada buku ini membahas dilakukan penerapan *text mining* terhadap ulasan *E-commerce* menggunakan Algoritma *K-Nearest Neighbor dan Naïve Bayes.* Hal ini dikarenakan e-*commerce* menggunakan sistematika penjualanya dengan B2C (*Bussines to consumer)* sehingga, pengguna dapat menunjukkan emosi atau perasaan mereka dalam user generated content (UGC), yang akan mempengaruhi kualitas kepercayaan consumen melalui UGC (*user generated content)* yaitu suatu rivew sebelum membeli sesuatu. Didalam ulasan *E-commerce* terdapat ulasan komentar yang belum diketahui hasil baik buruknya. Mengenai hal tersebut digunakan text mining dikarenakan text mining merupakan suatu teknik dalam pengambilan dari sejumlah data yang tidak terstruktur dari suatu topik tertentu, karena peningkatan jumlah pengguna aplikasi dapat dilihat dari baik buruknya ulasan pada komentar. Pada buku ini diangkat untuk meningkatkan kepercayaan bagi pengguna layanan *e-commerce* melalui text mining serta menggunakan algoritma KNN dan *Naivebayes* untuk

mencari hasil akurasi tertinggi serta menggunakan *wordcloud* untuk mengetahui kata yang sering muncul.

1.2 TUJUAN INTRUKSIONAL DAN CAPAIAN PEMBELAJARAN

Pembelajaran setelah diadakan studi pustaka serta studi kasus yang ada Memiliki tujuan mulia yaitu untuk memberikan pengetahuan serta ilmu setelah apa yang telah dilakukan oleh penulis untuk memberikan keterampilan, pengetahuan, informasi kepada pembaca. Selain itu Adapun tujuan penulis membuat buku. Tujuan instruksional memiliki dua macam, yaitu Tujuan Instruksional Umum (TIU) dan Tujuan Instruksional Khusus (TIK).

a) TIU

   TIU lebih fokus pada hasil yang harus dicapai oleh peserta didik. Buku ini mengulas mengenai text mining berdasarkan ulasan yang ada di komentar yang terdapat di *google play store e-commerce* melalui *user generated content.*

b) TIK

   TIK memiliki dua aspek yang tidak kalah penting, yaitu aspek perilaku peserta didik dan aspek pesan isi yang disampaikan. Pada perihal ini dalam penyusunan buku ini tidak terlepas dari masukan serta saran dari dosen untuk adi guru/dosen/pengajar secara tidak langsung berperan untuk mendidik perilaku agar berkelakuan baik dan dapat mengamalkan ilmu yang telah di dapatkan.

   Komponen membuat rumusan TIK yang lengkap meliputi tiga hal, yaitu *Terminal behavior, conditional of demonstration or rest dan standard of performance. Terminal behavior* sering juga disebut dengan tingkah laku akhir yang diharapkan. Misalnya, siswa bisa menjadi lebih paham.

   Berbeda dengan *conditional of demonstration or rest* atau yang disebut dengan kondisi demonstrasi.

BAGIAN 2

2.1 E-Commerce

E-Commerce atau perdagangan secara elektronik merupakan suatu kegiatan yang meliputi berbagai aktifitas seperti penyebaran, penjualan, pembelian, pemasaran produk yang berupa barang ataupun jasa yang memanfaatkan jaringan telekomunikasi. Atau dengan kata lain, *e- commerce* dapat diartikan sebagai transaksi jual beli secara elektronik (CloudHost, 2020). *E-business* mengarah pada definisi *e-commerce* yang lebih luas, tidak hanya ada proses jual beli, namun ada layanan pelanggan yang berkolaborasi dengan mitra bisnis lain. *E-Commerce* terbagi menjadi beberapa jenis, diantaranya *Consumen to Consumer (C2C), Business to Business (B2B), Consumen to Business (C2B), dan Business to Consumer (B2C)* (Gumilang, Z. A. N., dkk., 2018).

2.2 Machine Learning

Merupakan pembelajaran pendekatan algoritma untuk membuat prediksi dan keputusan berdasarkan pengalaman dan data. Model *machine learning* ini pada umumnya memiliki tiga kategori, diantaranya:

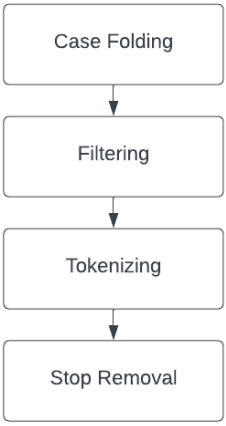
a. *Reinforcement learning*
b. *Supervised learning*
c. *Unsupervised learning*

Reinforcement learning merupakan algoritma *machine learning* yang dapat membuat *software* dan mesin bekerja secara otomatis untuk menentukan perilaku yang ideal dimana hal itu dapat memaksimalkan kinerja algoritmanya. *Supervised learning* adalah salah satu metode *machine learning* yang paling sering digunakan. Algoritma *supervised learning* merupakan *machine learning* yang harus diawasi karena bergantung pada kecocokan input dan output pada dataset yang diberikan. *Supervised learning* memiliki kelebihan dimana dapat mendefinisikan label dengan

lebih spesifik. *Unsupervised learning* merupakan *machine learning* yang menggunakan data tidak berlabel. Algoritma pada *unsupervised learing* ini mengidentifikasi dan mengelompokkan data berdasarkan fitur tertentu seperti kepadatan dan struktur datanya tanpa *training* data. Algoritma ini dapat memfilter data dan menemukan informasi dari data dengan mandiri. Adapun kelebihan *unsupervised learning* yaitu lebih sederhana dalam pengimplementasiannya. Model yang digunakan dalam penelitian ini termasuk dalam kategori *supervised learning* dimana algoritma ini terbagi menjadi dua jenis masalah, yaitu klasifikasi dan regresi. Klasifikasi (*classification*) menggunakan algoritma *supervised learning* untuk menetapkan *train* data kedalam kategori tertentu secara akurat. Klasifikasi ini mengelompokkan data kedalam kelas-kelas. Regresi (*regression*) menggunakan algoritma *supervised learning* untuk memahami hubungan antara variabel terikat dengan variabel bebas. Regresi ini data *training* menghasilkan nilai *output* tunggal dimana nilai ini merupakan interpretasi probabilitas yang mempertimbangkan kekuatan korelasi antar variabel *input*nya.

2.3 Text Preprocessing

Text preprocessing merupakan proses mengubah bentuk data yang sebelumnya tidak terstruktur kedalam bentuk data yang terstruktur. Berikut merupakan tahapan dari proses *preprocessing*:



Gambar 2. 2 Alur *Text Preprocessing*

a. *Case folding* merupakan tahapan mengubah semua huruf campuran seperti huruf besar (*uppercase*) maupun huruf kecil (*lowercase*) menjadi *lowercase* semua.
b. *Filtering* tahapan membersihkan data dari tanda baca, simbol, maupun elemen yang tidak dibutuhkan
c. *Tokenizing* merupakan tahapan memisahkan teks menjadi kata-kata
d. *Stopword removal* merupakan tahapan menghapus kata sambung

2.4 K-Fold Cross Validation

K-Fold Cross Validation (K-FCV) atau lebih sering disebut *cross validation* (validasi silang). Pada *cross validation*, dataset dibagi sebanyak k-lipatan. Pada setiap lipatan akan dipakai satu kali sebagai data uji dan lipatan sisanya dipakai sebagai data latih. Dengan menggunakan *cross validation* kita akan dapat hasil evaluasi yang lebih akurat karena model dievaluasi dengan seluruh data (Wayahdi et al., 2020).

Cross validation bersifat statistik. Algoritma ini terutama digunakan untuk memilih model untuk memperkirakan kesalahan uji model prediktif dengan lebih baik. Prinsip *cross validation* adalah membagi observasi sampel menjadi beberapa kelompok melalui pendekatan umum karena, mudah untuk dipahami dan mengarah pada penilaian yang kurang positif terhadap kemampuan model dibandingkan metode lain. Teknik *K-Fold Cross Validation*, semua kumpulan data pelatihan dipertimbangkan untuk pelatihan dan validasi. Dalam kumpulan data pelatihan, semua entri digunakan untuk validasi. Tahapan-tahapan algoritma K-FCV (Tamilarasi & Rani, 2020) :

a. Data pelatihan dibagi menjadi k subset yang sama seperti f1, f2, f3, …, fk. Semua subset disebut dengan lipatan *(fold).*
b. Dimulai dari i=0 sampai i=k.
c. f sebagai set validasi dan semua set k-1 lipatan yang tersisa ada di set pelatihan *cross validation.*
d. Menggunakan *cross validation*, latih model ML dan hitung akurasinya.
e. Evaluasi akurasi menggunakan semua k kasus *cross validation*

Cross validation merupakan prosedur untuk memperkirakan kinerja generalisasi. *Cross validation* memiliki banyak jenis seperti *Leave K-Out Cross Validation* (LKOCV), *Generalized Cross Validation* (GCV), dan R- Fold Cross Validation (RFCV), dalam hal perbedaan cara pemisahan data dan kompleksitas komputasi (Mu et al., 2018).

2.5 Python

Python adalah suatu bahasa pemrograman open source yang memakai contoh skrip (scripting languange) berorientasi objek . Python bersifat generik dengan juru bahasa dan dapat digunakan di domain aplikasi luas dan merupakan bahasa pemrograman tingkat tinggi yang fleksibel, sederhana, dan dinamis. Bahasa pemrograman ini dioptimalkan untuk kualitas perangkat lunak, produktivitas pengembang, portabilitas program, dan integritas komponen. Python telah digunakan untuk mengembangkan berbagai jenis software seperti *internet scripting, systems programming, user interfaces, product costumization, numberic programming*. Python saat ini menempati urutan ke-4 atau ke-5 bahasa pemrograman yang paling banyak digunakan di dunia. Bahasa pemrograman ini memiliki beberapa fitur yang dimanfaatkan oleh para pengembang perangkat lunak seperti Multi Paradigm Design, Open Source, Library Support, Portability, Extendable, dan Scalability (Tri Maya,dkk ,2019).

2.6 K-Nearest Neighbor

Metode K-NN merupakan salah satu metode yang populer dalam melakukan pengkategorian teks. Algoritma K-NN merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut (Liantoni, 2018). K-NN bertujuan untuk mengklasifikasikan objek baru berdasarkan atribut dan *training* sample (al Kautsar Aidilof & Kurniawan, n.d.). Dalam tahapnya, proses klasifikasi terhadap objek dilakukan berdasarkan data pembelajaran yang terdekat. Penentuan prediksi label kelas pada data uji ditentukan dengan nilai k yang menyatakan jumlah tetangga terdekat. K tetangga terdekat yang terpilih kemudian dilakukan voting kelas dari k tetangga terdekat tersebut. Label kelas hasil prediksi pada data uji tersebut bergantung pada kelas dengan jumlah suara tetangga terbanyak. K-NN memiliki beberapa karakteristik (Mardhyath et al., 2020), diantaranya :

a. K-NN termasuk algoritma yang menggunakan seluruh data latih untuk melakukan proses klasifikasi dimana hal ini

mengakibatkan proses prediksi yang sangat lama untuk data dalam jumlah sangat besar.

b. K-NN tidak membedakan fitur-fitur dengan suatu bobot.
c. K-NN masuk dalam kategori *lazy learning* yang menyimpan sebagian atau semua data dan hampir tidak ada proses pelatihan.
d. Penetuan nilai K yang paling sesuai merupakan hal yang rumit.
e. Dalam prinsip K-NN memilih tetangga terdekat, parameter jarak juga penting untuk dipertimbangkan sesuai dengan kasus datanya.

Adapun langkah-langkah dari algoritma *K-Nearest Neighbor* (Pratama, 2018), antara lain:

a. Menentukan parameter K = jumlah banyaknya tetangga terdekat
b. Menghitung jarak antara data baru dan semua data yang ada di data *training*
c. Penghitungan jarak antara data baru dan semua data yang ada menggunakan rumus *Euclidean Distance* :

$$d_{ij}(x_2, x_1) = \|x_2 - x_1\| = \sqrt{\sum_{k=1}^{m} \left|x_{ik} - x_{jk}\right|^2}$$

$$D(P_1, P_2) = \sqrt{(x_1 - x_2) + (y_1 - y_2)}$$

Keterangan :

$d_{ij=}jarak\ eucliedean\ objek\ data\ ke\ i\ dan\ objek\ data\ ke.j$

$m = banyaknya\ \ parameter\ yang\ digunakan$

$x_{2j} = objek\ data\ ke - i\ pada\ peubah\ ke - k$

$x_{1j} = objek\ data\ ke - j\ pada\ peubah\ ke - k$

a. Tentukan tetangga yang terdekat berdasarkan jarak minimum ke K dan urutkan jarak tersebut
b. Menentukan kategori dari tetangga terdekat

c. Terapkan kategori mayoritas yang sederhana dari tetangga yang terdekat sebagai nilai prediksi dari data yang baru.

2.7 Naïve Bayes

Metode K-NN merupakan salah satu metode yang populer dalam melakukan pengkategorian teks. Algoritma K-NN merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut (Liantoni, 2018). K-NN bertujuan untuk mengklasifikasikan objek baru berdasarkan atribut dan *training* sample (al Kautsar Aidilof & Kurniawan, n.d.). Dalam tahapnya, proses klasifikasi terhadap objek dilakukan berdasarkan data pembelajaran yang terdekat. Penentuan prediksi label kelas pada data uji ditentukan dengan nilai k yang menyatakan jumlah tetangga terdekat. K tetangga terdekat yang terpilih kemudian dilakukan voting kelas dari k tetangga terdekat tersebut. Label kelas hasil prediksi pada data uji tersebut bergantung pada kelas dengan jumlah suara tetangga terbanyak. K-NN memiliki beberapa karakteristik (Mardhyath et al., 2020), diantaranya :

a. K-NN termasuk algoritma yang menggunakan seluruh data latih untuk melakukan proses klasifikasi dimana hal ini mengakibatkan proses prediksi yang sangat lama untuk data dalam jumlah sangat besar.
b. K-NN tidak membedakan fitur-fitur dengan suatu bobot.
c. K-NN masuk dalam kategori *lazy learning* yang menyimpan sebagian atau semua data dan hampir tidak ada proses pelatihan.
d. Penetuan nilai K yang paling sesuai merupakan hal yang rumit.
e. Dalam prinsip K-NN memilih tetangga terdekat, parameter jarak juga penting untuk dipertimbangkan sesuai dengan kasus datanya.

Adapun langkah-langkah dari algoritma *K-Nearest Neighbor* (Pratama, 2018), antara lain:

a. Menentukan parameter K = jumlah banyaknya tetangga terdekat

b. Menghitung jarak antara data baru dan semua data yang ada di data *training*
c. Penghitungan jarak antara data baru dan semua data yang ada menggunakan rumus *Euclidean Distance* :
d. $d_{ij}(x_2, x_1) = \|x_2 - x_1\| = \sqrt{\sum_{k=1}^{m} \quad |x_{ik} - x_{jk}|^2}$
e. $D(P_1, P_2) = \sqrt{(x_1 - x_2) + (y_1 - y_2)}$

Keterangan :

1. $d_{ij} = jarak\ eucliedean\ objek\ data\ ke - i\ dan\ objek\ data\ ke - j$
2. $m = banyaknya\ peubah\ atau\ parameter\ yang\ digunakan$
3. $x_{2j} = objek\ data\ ke - i\ pada\ peubah\ ke - k$
4. $x_{1j} = objek\ data\ ke - j\ pada\ peubah\ ke - k$

f. Tentukan tetangga yang terdekat berdasarkan jarak minimum ke K dan urutkan jarak tersebut
g. Menentukan kategori dari tetangga terdekat
h. Terapkan kategori mayoritas yang sederhana dari tetangga yang terdekat sebagai nilai prediksi dari data yang baru.

2.8 TF-IDF

Term Frequency-Invers Document Frequency (TF-IDF) merupakan perhitungan bobot *term* pada sebuah dokumen berdasarkan seringnya kata tersebut muncul dimana bobot tersebut mengindikasikan pentingnya sebuah *term* terhadap dokumen, semakin banyak *term* tersebut muncul pada dokumen maka semakin tinggi nilai *term* tersebut (Yudiarta et al., 2018). Adapun rumus dari TF-IDF yaitu sebagai berikut

$$w_{i,j} = tf_{i,j} \times log\ log\ \frac{N}{df_i}$$

Keterangan:

$w_{i,j}$: bobot TF-IDF

$tf_{i,j}$: jumlah kemunculan *term i* pada dokumen
df_i : jumlah frekuensi dokumen tiap kata
N : jumlah total dokumen

Hasil dari pembobotan kata menggunakan TF-IDF ini adalah perkalian dari nilai TF dan IDF yang akan menghasilkan bobot lebih kecil apabila kata tersebut sering muncul pada setiap dokumen dalam koleksi, begitupun sebaliknya bobot TF-IDF akan lebih besar apabila kata tersebut jarang muncul pada setiap dokumen dalam koleksi (Andre Septian et al., 2019).

2.9 Confusion Matrix

Confusion matrix merupakan tabel yang menyatakan klasifikasi jumlah data uji yang benar dan jumlah data uji yang salah.

Tabel 2. 1 *confusion matrix*

		Prediksi	
		1	0
Fakta	1	TP	FN
	0	FP	TN

Keterangan:

TP (*True Positive*) = jumlah daya dari kelas 1 yang benar diklasifikasikan sebagai kelas 1

TN (*True Negative*) = jumlah data dari kelas 0 yang benar diklasifikasikan sebagai kelas 0

FP (*False Positive*) = jumlah data dari kelas 0 yang salah diklasifikasikan sebagai kelas 1

FN (*False Negative*) = jumlah data dari kelas 1 yang salah diklasifikasikan sebagai kelas 0 Adapun rumus dari *confusion matrix* yang digunakan untuk menghitung *accurrancy, precision, recall* :

$$Accurancy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision\ (negatif) = \frac{TN}{TN + FN}$$

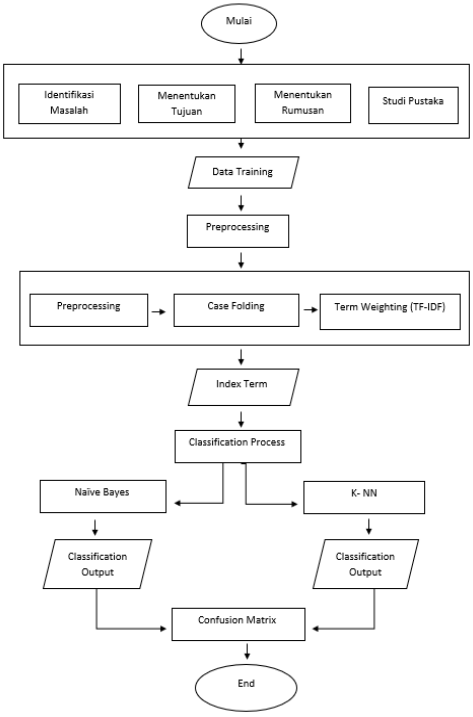
$$Precision\ (positif) = \frac{TP}{TP + FP}$$

$$Recall\ (negatif) = \frac{TN}{TN + FP}$$

$$recision\ (positif) = \frac{TP}{TP + FN}$$

2.10 Diagram Alur Metodologi Penelitian

Pada pembahasan buku ini dalam melakukan *text processing* dilakukan perancangan supaya menghasilkan data yang terstruktur.



Gambar 2.4 Diagram Alur Metodologi

BAGIAN TIGA

3.1 Analisis

Dalam tahapan ini hal pertama yang dilakukan adalah dengan menganalisa system. Analisa merupakan sebuah proses mempelajari suatu system dengan cara menguraikan system tersebut kedalam elemen yang membentuknya. Dengan menganalisa akan memudahkan tahapan selanjutnya untuk mengidentifikasi dan mengevaluasi permasalahan yang terjadi serta kebutuhan yang diperlukan sehingga diselesaikan dengan baik.

3.2 Tutorial Perhitungan Text Mining Pada Data

Pengumpulan data merupakan proses memperoleh data yang digunakan pada penelitian. Teknik pengumpulan data dan informasi yang digunakan dalam penelitian ini dengan melakukan *scrapping* pada ulasan aplikasi *E-Commerce* di *Google Play* dengan menggunakan teknik *web scraping di Jupiter notebook,* dengan menghasilkan data *scrapping* sebesar 4000 data. Berikut ini adalah tahapan dalam melakukan text mining.

Tabel 3. 1 Dataset

No.	userName	score	at	content
1	Siti Nur Lailatul	5	8/15/2022 7:37	mantap
2	Yin X	1	8/15/2022 7:47	Makin hari makin ampas ni aplikasi
3	AG .18	5	8/15/2022 7:47	bagus
4	heru triantoko	5	8/15/2022 7:48	saya pelanggan tokopedia selalu amanah
5	Ali Payo	5	8/15/2022 7:58	mantap
6	arief marret	5	8/15/2022 8:02	top
7	Dimas Ramadhani	5	8/15/2022 8:06	ok
9	Mistianah VCO	5	8/15/2022 8:10	meskipun ada pembatalan pesanan, bisa ttp blnja dg mengganti produk lainnya, trimksh tokopedia, menjaga kualitas pelayanannya, s
10	Aulia Hanafiah	1	8/15/2022 8:19	Yang kamu suka dibatalin otomatis pesanan gratis kamu itu krn ganti2 akun. 1 akun 1 hp 1 gratisan. Klo pengen nambah pk hp sodara
...	.....	..	.....	.....
3991	Fajar Ruhunussa	1	8/28/2022 4:15	Jadi males belanja disini lemotnya makin parah
3992	Wira Dharma	3	8/28/2022 4:17	Kriteria belanja bebas ongkir tidak jelas
3993	Dewi Eka rahayu	5	8/28/2022 4:18	oke
3994	Zaenal Albara	5	8/28/2022 4:20	ok
3995	Mutia Autoservice	5	8/28/2022 4:26	Sy pelanggan lama....Ga ada keraguan untuk TOPED LANCAR JAYA
3996	Seto Ludiro	1	8/28/2022 4:27	"jika pembeli sudah pernah verifikasi data diri sebelum membeli alkohol atau rokok elektrik sebelumnya, Tokopedia tidak akan memi
3997	Faisal pslx	1	8/28/2022 4:29	Toko pedia AMIG. Knapa pengiriman dipersulit bayar cargo bayar ongkir kurir lagi untuk bisa C.O.D mana Rp.100.000 lagi.trus bayar
3998	Lilik Andi	5	8/28/2022 4:32	Tokopedia masih yang terbaik
3999	Lucita fathan	1	8/28/2022 4:33	Good bye tokoped now, skrg beda jdi males belanja. skrg mending belanja di toko sebelah.
4000	Malawi Supendi	5	8/28/2022 4:38	makin topmarkotop
4001	Made Hendrawan	1	8/28/2022 4:40	Aduhh kok gak bisa update tokpednya

Adapun tahapan - tahapan yang dilakukan oleh peneliti antara lain seperti melakukan penginstalan pada *Jupiter notebook:*

1. Terdapat beberapa *library* yang digunakan diantaranya *librarypandas,numpy,matplotlib,seaborn,warnings,cv*

*re,string.*Tahapan pertama yang dilakukan yaitu meng*import pandas as pd* yaitu berfungsi sebagai kebutuhan analisis, manipulasi serta pembersihan data dengan pendukung CSV untuk dijadikan obyek *python* dengan *rows & columns* (dataframe). Selanjutnya pada *import numpy as np* yang berfungsi untuk menyimpan data dalam bentuk *array* dengan mengumpulkan variable yang memiliki tipe data yang sama.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sbn
import warnings
import cv2
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score
import re
from sklearn.feature_extraction.text import TfidfVectorizer,CountVectorizer
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords
from scipy.sparse import hstack
from sklearn.multiclass import OneVsRestClassifier
from sklearn.linear_model import SGDClassifier
import string
from tqdm import trange
```

Gambar 3.1 *Library* Proses *Crawling*

Tabel 3.2 Keterangan

No.	Nama	Keterangan
1.	*Pandas as pd*	Kebutuhan analisis, manipulasi dan pembersihan data, dengan pendukung CSV untuk dijadikan objek *python* dengan *rows & columns* (dataframe)
2.	*Numpy as np*	Menyimpan data dalam bentuk *array*, yaitu kumpulan variable yang memiliki tipe data yang sama
3.	*Matplotlib*	Visualisasi data membuat plot serta grafik

4.	*Seaborn as sbn*	*Library* yang dibangun diatas *matplotlib*
5.	*String*	Tipe data untuk teks yang terdiri dari gabungan huruf,angka dan berbagai karakter
6.	*Multinominal NB*	Permodelan yang dilakukan pada perhitungan di *naïve bayes*
7.	*Pandas as pd*	Permodelan untuk mendapatkan hasil akurasi tertinggi
8.	*Numpy as np*	Urutan karakter yang membentuk pola pencarian yang digunakan untuk memeriksa apakah *string* berisi pola pencarian yang ditentukan
9.	*Tfidf Vectorizer*	Algoritma yang dapat digunakan untuk menganalisa hubungan antara sebuah frase/kalimat dengan sekumpulan dokumen
10.	*Word_tokenize*	Operasi yang memisahkan teks menjadi potongan-potongan berupa token, dapat berupa potongan huruf, kata, atau kalimat, sebelum dianalisis lebih lanjut
11.	*Stemmer*	Pemotongan imbuhan (awalan, akhiran, sisipan, kombinasi) yang dijalankan dengan judul aslinya
12.	*Stopword*	*Stop words* adalah kata umum yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna.

13.	*Stack*	Salah satu struktur data yang digunakan untuk menyimpan sekumpulan objek ataupun variabel.

2. Selanjutnya dilakukan pemuatan ulang data .csv dengan total 4000 rows x 6 columns. Pada proses ini dilakukan pemrosesan data .csv yang akan diolah dengan beberapa fungsi antara lain adalah *username, content, score, at, replyContent,RepliedAt. Data hasil dari scrapping ditampilkan*

	userName	score	at	content
0	Siti Nur Lailatul	5	2022-08-15 07:37:13	mantap
1	Yin X	1	2022-08-15 07:47:47	Makin hari makin ampas ni aplikasi
2	AG .18	5	2022-08-15 07:47:52	bagus
3	heru triantoko	5	2022-08-15 07:48:09	saya pelanggan tokopedia selalu amanah
4	Ali Payo	5	2022-08-15 07:58:41	mantap
...	...	...	...	...
3995	Faisal pslx	1	2022-08-28 04:29:42	Toko pedia AMJG. Knapa pengiriman dipersulit b...
3996	Lilik Andi	5	2022-08-28 04:32:22	Tokopedia masih yang terbaik
3997	Lucita fathan	1	2022-08-28 04:33:48	Good bye tokoped now, skrg beda jdi males bela...
3998	Malawi Supendi	5	2022-08-28 04:38:52	makin topmarkotop
3999	Made Hendrawan	1	2022-08-28 04:40:32	Aduhh kok gak bisa update tokpednya

4000 rows × 4 columns

Gambar 3.2 Pemuatan dataset .csv

3. Pemberian label penilaian untuk membedakan antanya *content negative* dan *content positif.* Dengan pemrosesan final dari label_penilaian diubah menjadi *score*.

index	userName	score	at	content	sentiment
0	Siti Nur Lailatul	5	2022-08-15 07:37:13	mantap	Positif
1	Yin X	1	2022-08-15 07:47:47	Makin hari makin ampas ni aplikasi	Negatif
2	AG .18	5	2022-08-15 07:47:52	bagus	Positif
3	heru triantoko	5	2022-08-15 07:48:09	saya pelanggan tokopedia selalu amanah	Positif
4	Ali Payo	5	2022-08-15 07:58:41	mantap	Positif

Show 25 per page

Gambar 3.3 Pelabelan data sentiment

4. Selanjutnya dilakukan pendownload *library* nltk yaitu library yang digunakan untuk membantu dalam bekerja dengan teks. Library ini memudahkan untuk memproses teks seperti melakukan *classification,*

tokenization, stemming, tagging, parsing, dan *semantic reasoning* yang selanjutnya dilakukan pendownloadan untuk dilakukan proses *stemming* antaralain yaitu *punkt,stopword,wordnet*

```
nltk.download('punkt')
nltk.download('stopword')
nltk.download('wordnet') #stemming
```

Gambar 3.5 Proses download untuk *stemming*

5. Pada tahapan selanjutnya adalah data menampilkan data *content*

```
data.content
```

```
0                                                    mantap
1                          Makin hari makin ampas ni aplikasi
2                                                      bagus
3                     saya pelanggan tokopedia selalu amanah
4                                                    mantap
                                ...
3995    Toko pedia AMJG. Knapa pengiriman dipersulit b...
3996                           Tokopedia masih yang terbaik
3997    Good bye tokoped now, skrg beda jdi males bela...
3998                                       makin topmarkotop
3999                    Aduhh kok gak bisa update tokpednya
Name: content, Length: 4000, dtype: object
```

Gambar 3.6 Data *Content*

6. Data *Cleaning* merupakan suatu prosedur untuk memastikan kebenaran, konsistensi dan kegunaan suatu data yang ada dalam dataset. Caranya dengan mendeteksi adanya *error* atau *corrupt* pada data kemudian memperbaiki atau menghapus data apabila diperlukan. Pada proses *cleaning* data ada beberapa tahapan seperti *remove* url,merubah teks menjadi huruf kecil, *remove mention, remove hastag, remove next character* serta *remove punctuation.* Data yang berkualitas buruk akan memberikan hasil dan algoritma yang tidak bisa menjamin kebenaranya meski proses analisisnya benar. Berikut ini beberapa alasan mengapa data *cleaning* harus dilakukan diantaranya

menghilangkan kesalahan yang muncul saat beberpa data sources dikumpulkan dalam satu dataset. Meningkatkan efisiensi kerja karena proses ini memudahkan dalam pengolah data untuk menemukan apa yang dibutuhkan dari data. Selanjutnya tingkat *error* yang rendah juga akan mendatangkan kepuasan pelanggan dan mengurangi beban kerja .

```
#cleaning
#remove url
data['content'] = data['content'].str.replace('https\S','', case=False)

#merubah teks menjadi huruf kecil
data['content'] = data['content'].str.lower() #case folding

#remove mention
data['content'] = data['content'].str.replace('@\S+','',case=False)

#remove hastag
data['content'] = data['content'].str.replace('#\S+','',case=False)


#remove next character
data['content'] = data['content'].str.replace("\'\W+",'',case=False)

#remove punctuation
data['content'] = data['content'].str.replace('[^\w\s]','',case=False)
```

Gambar 3.6 *Cleaning* Data

7. *Tokenizing* merupakan metode untuk melakukan pemisahan kata dalam suatu kalimat dengan tujuan untuk proses analisis teks lebih lanjut. Sebelum melakukan *tokenizing* biasanya melakukan proses *case folding* yang didalamnya mencakup proses menghapus angka dan tanda baca yang tidak perlu, dan *whitespace*. Pada *tokenizing* proses pengumpulan diskripsi yang semula dari kalimat menjadi data testing.Selain itu juga ada word segmentation adalah suatu permasalahan terkait pembagian string bahasa tertulis menjadi kata-kata komponennya.

```
#tokenizing merupakan proses pengumpulan penguraian diskripsi yang semula berupa kalimat mejadi kata
#testing
from nltk.tokenize import word_tokenize

x = data.iloc[0]
print(nltk.word_tokenize(x['content']))
```

```
['mantap']
```

Gambar 3.7 *Tokenizing*

8. Pada proses ini dilakukan pengidentifikasian pada token. Token juga berasal dari jumlah total dari kata

yang terdapat didalam sebuah kalimat terlepas seberapa sering kata tersebut diulang.

```
def identify_tokens(row) :
  text = row['content']
  tokens =nltk.word_tokenize(text)
  token_words = [w for w in tokens if w.isalpha()]
  return token_words

data['content'] = data.apply(identify_tokens, axis = 1)
data.content
```

```
0                                              [mantap]
1              [makin, hari, makin, ampas, ni, aplikasi]
2                                               [bagus]
3           [saya, pelanggan, tokopedia, selalu, amanah]
4                                              [mantap]
                                ...
3995    [toko, pedia, amjg, knapa, pengiriman, dipersu...
3996                    [tokopedia, masih, yang, terbaik]
3997    [good, bye, tokoped, now, skrg, beda, jdi, mal...
3998                                 [makin, topmarkotop]
3999          [aduhh, kok, gak, bisa, update, tokpednya]
Name: content, Length: 4000, dtype: object
```

Gambar 3.8 Identifikasi token

9. Berikut ini merupakan hasil dari pengidentifikasian token pada 4000 data yang digunakan yang terbagi menjadi content, label_penilaian serta *score.*

	content	label_Penilaian	score
0	[mantap]	Positif	1
1	[makin, hari, makin, ampas, ni, aplikasi]	Negatif	0
2	[bagus]	Positif	1
3	[saya, pelanggan, tokopedia, selalu, amanah]	Positif	1
4	[mantap]	Positif	1
...	...	...	...
3995	[toko, pedia, amjg, knapa, pengiriman, dipersu...	Negatif	0
3996	[tokopedia, masih, yang, terbaik]	Positif	1
3997	[good, bye, tokoped, now, skrg, beda, jdi, mal...	Negatif	0
3998	[makin, topmarkotop]	Positif	1
3999	[aduhh, kok, gak, bisa, update, tokpednya]	Negatif	0

4000 rows × 3 columns

Gambar 3.9 Identifikasi token

10. *Stemming* adalah proses pemetaan dan penguraian bentuk dari suatu kata menjadi bentuk kata dasarnya, pemotongan imbuhan yang akan dijalankan untuk membentuk kata dasar yang akan diproses pada text mining.

```
#stemming (pembentukan kata dasar)
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer
stemming = PorterStemmer
```

```
def stem_list(row) :
  text = row ['content']
  stem = [steaming.stem(word) for word in text]
  return(stem)

  data['content'] = data.apply(stem_list, axis =1)
  data.reviewContent
```

Gambar 3.10 Stemming

11. Tahapan *stopword* merupakan proses penyaringan kata-kata sebelum dan sesudah pemrosesan, pada *stopword* merupakan kata yang diabaikan dalam pemrosesan dan biasanya disimpan di dalam *stop lists*. *Stop list* ini berisi daftar kata umum yang mempunyai fungsi tapi tidak mempunyai arti. Karakteristik utama dalam pemilihan *stopwords* biasanya adalah kata yang mempunyai frekuensi kemunculan yang tinggi misalnya kata penghubung

```
from nltk.corpus import stopwords
import nltk
nltk.download('stopwords')
#from nltk.tokenize import word_tokenize
stops = set (stopwords.words('indonesian'))
```

Gambar 3.11 Proses *Stopword*

12. Pengambilan kolom label dalam *variable y_train*

```
# mengambil hanya kolom label dalam variabel y_train
y_train = []
for index, row in data.iterrows():
    y_train.append(row["score"])
print("Jumlah score1: ", len(y_train))
```

```
Jumlah score1:  4000
```

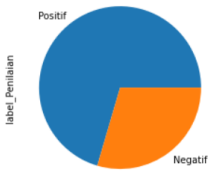
Gambar 3.12 Pelabelan Kolom

13. Pada Pemrosesan pada Pie plot digunakan untuk menyajikan data dalam bentuk persentase, yang mana setiap potongan pie berisikan data dengan ukuran

tertentu. Secara visual *pie plot* menarik untuk penyajian data serta memudahkan pembaca dalam memahami data tersebut

```
data.label_Penilaian.value_counts().plot.pie()
data.label_Penilaian.value_counts()
```

```
Positif    2818
Negatif    1182
Name: label_Penilaian, dtype: int64
```



Gambar 3.13 Plot Pie

14. *Penyimpanan hasil pelabeling akan disimpan kedalam csv.* Berikut ini adalah hasil data dari proses pelabelan

		label_Penilaian	content
1	0	Positif	['mantap']
2	1	Negatif	kin', 'ampas', 'ni', 'aplikasi']
3	2	Positif	['bagus']
4	3	Positif	opedia', 'selalu', 'amanah']
5	4	Positif	['mantap']
6	5	Positif	['top']
7	6	Positif	['ok']
8	7	Positif	anannya', 'sukses', 'terus']
9	8	Negatif	, 'buat', 'pelanggan', 'baru']
10	9	Positif	ko', 'nya', 'semua', 'penipu']
11	10	Negatif	h', 'dr', 'x', 'jam', 'ga', 'juga']
12	11	Positif	mbed', 'premium', 'cotton']
13	12	Positif	['sanjay', 'baik']
14	13	Positif	['mantap']
15	14	Positif	['bagus']
16	15	Negatif	ivasi', 'konsumen', 'dijaga']
17	16	Negatif	naan', 'data', 'heran', 'deh']
18	17	Negatif	, 'harus', 'ada', 'iklan', 'nya']
19	18	Positif	'klo', 'mau', 'beli', 'barang']
20	19	Positif	iya', 'mudah', 'komplainya']
21	20	Negatif	'utk', 'jasa', 'ekspedisinya']
22	21	Negatif	, 'malah', 'di', 'bikin', 'rumit']
23	22	Negatif	['tidak', 'kompoten']

Gambar 3.14 Penyimpanan csv

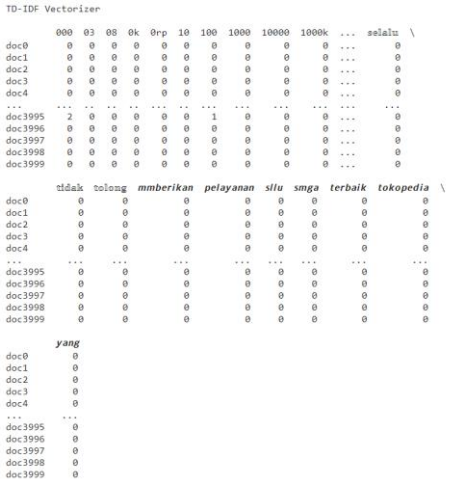
15. Perhitungan *Vector,* menghitung frekuensi kata dalam dokumen. *Count Vectorizer* dapat mengubah fitur teks menjadi sebuah representasi *vector*. TF – IDF, atau

pembobotan kata merupakan skema yang digunakan untuk menghitung bobot setiap kata yang paling umum digunakan.

```
#menghitung vector
from sklearn.feature_extraction.text import CountVectorizer
import pandas as pd
```

Gambar 3.15 Perhitungan *Vector*

16. *TF IDF* merupakan sebuah metode algoritma yang berguna untuk menghitung bobot setiap kata yang umum digunakan. Pada proses ini akan menghitung bobot setiap kata yang umum digunakan. Metode ini digunakan untuk mengetahui beberapa sering suatu kata yang muncul dalam dokumen.



TD-IDF Vectorizer

	000	03	08	0k	0rp	10	100	1000	10000	1000k	...	selaltz
doc0	0	0	0	0	0	0	0	0	0	0	...	0
doc1	0	0	0	0	0	0	0	0	0	0	...	0
doc2	0	0	0	0	0	0	0	0	0	0	...	0
doc3	0	0	0	0	0	0	0	0	0	0	...	0
doc4	0	0	0	0	0	0	0	0	0	0	...	0
...	...	..	..	..	...	..	...	...	...	...	...	...
doc3995	2	0	0	0	0	0	1	0	0	0	...	0
doc3996	0	0	0	0	0	0	0	0	0	0	...	0
doc3997	0	0	0	0	0	0	0	0	0	0	...	0
doc3998	0	0	0	0	0	0	0	0	0	0	...	0
doc3999	0	0	0	0	0	0	0	0	0	0	...	0

	tidak	tolong	mmberikan	pelayanan	sllu	smga	terbaik	tokopedia
doc0	0	0	0	0	0	0	0	0
doc1	0	0	0	0	0	0	0	0
doc2	0	0	0	0	0	0	0	0
doc3	0	0	0	0	0	0	0	0
doc4	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...
doc3995	0	0	0	0	0	0	0	0
doc3996	0	0	0	0	0	0	0	0
doc3997	0	0	0	0	0	0	0	0
doc3998	0	0	0	0	0	0	0	0
doc3999	0	0	0	0	0	0	0	0

	yang
doc0	0
doc1	0
doc2	0
doc3	0
doc4	0
...	...
doc3995	0
doc3996	0
doc3997	0
doc3998	0
doc3999	0

Gambar 3.16 Proses TFidf

17. Split dataset dalam pre *prosessing* untuk mendapatkan dataset yang proposional

```
# spliting data
from sklearn.model_selection import train_test_split
x_train, x_test,y_train, y_test = train_test_split(
data01['content'],data01['label_Penilaian'], test_size=0.2, random_state=7)
```

Gambar 3.17 Proses splitting data

18. Pemrosesan menggunakan *confusion_matrix* merupakan salah satu cabang dari disiplin ilmu kecerdasan buatan (*artificial intelligence*) yang membahas bagaimana sistem dibangun berdasarkan pada data. Pada proses ini menggunakan algortima KNN. Pada penelitian ini, perhitungan jarak data pada KNN dengan perhitungan *confusion matrix* yang digunakan untuk mengukur kinerja dari model klasifikasi di *machine learning. Confusion matrix* adalah salah satu *tools* analitik prediktif yang menampilkan dan membandingkan nilai aktual atau nilai sebenarnya dengan nilai hasil prediksi model yang dapat digunakan untuk menghasilkan metrik evaluasi seperti *Accuracy* (akurasi), *Precision*, *Recall*, dan *F1-Score* atau *F-Measure*. Nilai akurasi didapatkan dari jumlah data bernilai positif yang diprediksi positif dan data bernilai negatif yang diprediksi negatif dibagi dengan jumlah seluruh data di dalam dataset. *Precision* adalah peluang kasus yang diprediksi positif yang pada kenyataannya termasuk kasus kategori positif. Recall.*Recall* adalah peluang kasus dengan kategori positif yang dengan tepat diprediksi positif. Nilai *F1-Score* atau dikenal juga dengan nama *F-Measure* didapatkandari hasil *Precision* dan *Recall* antara kategori hasil prediksi dengan kategori sebenarnya. Pada hasil penelitian ini mendapatkan nilai *precision negative* sebesar *0.84, recall 0.19, f1-score 0.30 support 220* serta pada *precision* positif mendapatkan hasil sebesar *0.75, recall 0.99, f1-score 0.86* dan *support 580*

```
model5 = KNeighborsClassifier(n_neighbors=5)
model5.fit(features_train,y_train)
prediction_knn5 = model5.predict(features_test)
print(accuracy_score(y_test,prediction_knn5))
print(classification_report(y_test,prediction_knn5))
```

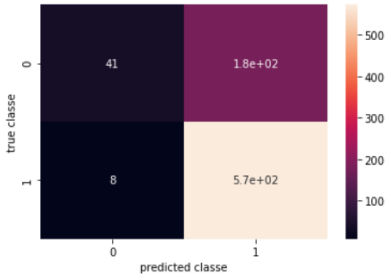
```
0.76625
              precision    recall  f1-score   support

     Negatif       0.84      0.19      0.30       220
     Positif       0.76      0.99      0.86       580

    accuracy                           0.77       800
   macro avg       0.80      0.59      0.58       800
weighted avg       0.78      0.77      0.71       800
```

Gambar 3.18 Algortima KNN

19. *Seaborn* merupakan salah satu pustaka visual *python* yang berlandaskan pada *matplotlib*. Seaborn menyediakan antar-muka tingkat tinggi untuk menangani permasalahan terkait visualisasi data secara statistik agar tampak lebih menarik. *Seaborn* merupakan salah satu pustaka visual *Python* yang berlandaskan pada *matplotlib.* Seaborn menyediakan antar-muka tingkat tinggi untuk menangani permasalahan terkait visualisasi data secara statistik agar tampak lebih menarik.



Gambar 3.19 *Seaborn*

20. Hasil akhir dari perhitungan text mining adalah

```
print(accuracy_score(y_test,prediction_knn9))
```

```
0.76
```

Gambar 3.20 Nilai Akurasi KNN

21. Setelah mengetahui hasil dari perhitungan KNN, dilanjutkan dengan pemrosesan *naïve bayes* dengan pengklasifikasi data berdasarkan probabilitas. Pada *naïve bayes* ini menetapkan label kelas ke i*nstance*/catatan menggunakan probabilitas bersyarat.

```
#Import Library
import pandas as pd
import string
import numpy as np
import nltk
import string
import re
import collections
from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory

import joblib
import pickle
from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
from sklearn.multiclass import OneVsRestClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.model_selection import train_test_split
from sklearn.utils.multiclass import unique_labels
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import LabelBinarizer,OrdinalEncoder,OneHotEncoder
from sklearn.metrics import accuracy_score
```

Gambar 3.21 Naïve Bayes

Tabel 3.3 Keterangan

No.	Nama	Keterangan
1.	*Pandas as pd*	Kebutuhan analisis, manipulasi dan pembersihan data, dengan pendukung CSV untuk dijadikan objek *python* dengan *rows & columns*
2.	*Numpy as np*	Menyimpan data dalam bentuk *array*, yaitu kumpulan variable yang memiliki tipe data yang sama

3.	*Matplotlib*	Visualisasi data membuat plot serta grafik
4.	*Seaborn as sbn*	*Library* yang dibangun diatas *matplotlib*
5.	*String*	Tipe data untuk teks yang terdiri dari gabungan huruf,angka dan berbagai karakter
6.	*Multinominal NB*	Permodelan yang dilakukan pada perhitungan di *naïve bayes*
7.	*Pandas as pd*	Permodelan untuk mendapatkan hasil akurasi tertinggi
8.	*Numpy as np*	Urutan karakter yang membentuk pola pencarian yang digunakan untuk memeriksa apakah *string* berisi pola pencarian yang ditentukan
9.	*Tfidf Vectorizer*	Algoritma yang dapat digunakan untuk menganalisa hubungan antara sebuah frase/kalimat dengan sekumpulan dokumen
10.	*Word_tokenize*	Operasi yang memisahkan teks menjadi potongan-potongan berupa token, dapat berupa potongan huruf, kata, atau kalimat, sebelum dianalisis lebih lanjut
11.	*Stemmer*	Pemotongan imbuhan (awalan, akhiran, sisipan, kombinasi) yang dijalankan dengan judul aslinya
12.	*Stopword*	*Stop words* adalah kata umum yang biasanya muncul dalam

		jumlah besar dan dianggap tidak memiliki makna.
13.	*Stack*	Salah satu struktur data yang digunakan untuk menyimpan sekumpulan objek ataupun variabel.

22. Melakukan pelabelan ke setiap content, dengan sistematika apabila positif maka akan bernilai 1, dan sebaliknya apabila negative akan bernilai 0

```
label_positive = 1
label_negative = 0
```

Gambar 3.22 *Pelabelan*

23. Selanjutnya adalah melihat kembali data yang akan diproses menggunakan algoritma *Naïve Bayes.*

	Unnamed: 0	label_Penilaian	content
0	0	Positif	['mantap']
1	1	Negatif	['makin', 'hari', 'makin', 'ampas', 'ni', 'apl...
2	2	Positif	['bagus']
3	3	Positif	['saya', 'pelanggan', 'tokopedia', 'selalu', '...
4	4	Positif	['mantap']
5	5	Positif	['top']
6	6	Positif	['ok']
7	7	Positif	['meskipun', 'ada', 'pembatalan', 'pesanan', '...
8	8	Negatif	['yang', 'kamu', 'suka', 'dibatalin', 'otomati...
9	9	Positif	['toko', 'nya', 'semua', 'penipu']

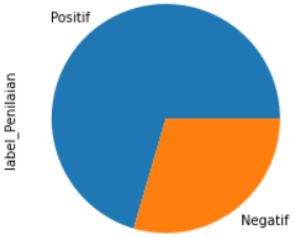
Gambar 3.23 Data Awal

24. Pemrosesan pada Pie plot digunakan untuk menyajikan data dalam bentuk persentase, yang mana setiap potongan pie berisikan data dengan ukuran tertentu. Secara visual *pie plot* menarik untuk penyajian data

serta memudahkan pembaca dalam memahami data tersebut

```
f.label_Penilaian.value_counts().plot.pie()
f.label_Penilaian.value_counts()
```

```
Positif    2818
Negatif    1182
Name: label_Penilaian, dtype: int64
```



Gambar 3.24 *Plot Pie*

25. Pada fungsi ini menampilkan data yang terdapat pada *content*

Gambar 3.25 Data *Content*

26. Pada tahapan *Tokenizing* melakukan proses yang paling awal dalam melakukan text mining. Dalam proses ini, input *stream* yang didapat dari file text akan dipecah-pecah menjadi bagian bagian yang lebih kecil. Sebagai contoh pemecahan kalimat menjadi kata-kata (*tokens*).

	Unnamed: 0	label_Penilaian	content	Cleaning	Tokenization
0	0	Positif	['mantap']	['mantap']	[, mantap,]
1	1	Negatif	['makin', 'hari', 'makin', 'ampas', 'ni', 'apl...	['makin' 'hari' 'makin' 'ampas' 'ni' 'aplikasi']	[, makin, hari, makin, ampas, ni, aplikasi,]
2	2	Positif	['bagus']	['bagus']	[, bagus,]
3	3	Positif	['saya', 'pelanggan', 'tokopedia', 'selalu', '...	['saya' 'pelanggan' 'tokopedia' 'selalu' 'aman...	[, saya, pelanggan, tokopedia, selalu, amanah,]
4	4	Positif	['mantap']	['mantap']	[, mantap,]
5	5	Positif	['top']	['top']	[, top,]
6	6	Positif	['ok']	['ok']	[, ok,]
7	7	Positif	['meskipun', 'ada', 'pembatalan', 'pesanan', '...	['meskipun' 'ada' 'pembatalan' 'pesanan' 'bisa...	[, meskipun, ada, pembatalan, pesanan, bisa, t...
8	8	Negatif	['yang', 'kamu', 'suka', 'dibatalin', 'otomati...	['yang' 'kamu' 'suka' 'dibatalin' 'otomatis' '...	[, yang, kamu, suka, dibatalin, otomatis, pesa...
9	9	Positif	['toko', 'nya', 'semua', 'penipu']	['toko' 'nya' 'semua' 'penipu']	[, toko, nya, semua, penipu,]

Gambar 3.26 Pelabelan

27. *Stop removal* didefinisikan sebagai sekumpulan kata yang tidak berhubungan *(irrelevant)* dengan subyek utama yang dimaksud, meskipun kata tersebut sering muncul didalam data yang digunakan. Kata-kata yang dimaksud biasanya adalah jenis kata sambung, imbuhan, dan lain sebagainya.

```
#Stop Removal
stopword = nltk.corpus.stopwords.words('indonesian')

def remove_stopwords(text):
    text = [word for word in text if word not in stopword]
    return text
f['Stop_Removal']= f['Tokenization'].apply(lambda x: remove_stopwords(x))
f.head(10)
```

	Unnamed: 0	label_Penilaian	content	Cleaning	Tokenization	Stop_Removal
0	0	Positif	['mantap']	['mantap']	[, mantap,]	[, mantap,]
1	1	Negatif	['makin', 'hari', 'makin', 'ampas', 'ni', 'apl...	['makin' 'hari' 'makin' 'ampas' 'ni' 'aplikasi']	[, makin, hari, makin, ampas, ni, aplikasi,]	[, ampas, ni, aplikasi,]
2	2	Positif	['bagus']	['bagus']	[, bagus,]	[, bagus,]
3	3	Positif	['saya', 'pelanggan', 'tokopedia', 'selalu', '...	['saya' 'pelanggan' 'tokopedia' 'selalu' 'aman...	[, saya, pelanggan, tokopedia, selalu, amanah,]	[, pelanggan, tokopedia, amanah,]
4	4	Positif	['mantap']	['mantap']	[, mantap,]	[, mantap,]
5	5	Positif	['top']	['top']	[, top,]	[, top,]
6	6	Positif	['ok']	['ok']	[, ok,]	[, ok,]
7	7	Positif	['meskipun', 'ada', 'pembatalan', 'pesanan', '...	['meskipun' 'ada' 'pembatalan' 'pesanan' 'bisa...	[, meskipun, ada, pembatalan, pesanan, bisa, t...	[, pembatalan, pesanan, ttp, blnja, dg, mengga...
8	8	Negatif	['yang', 'kamu', 'suka', 'dibatalin',	['yang' 'kamu' 'suka' 'dibatalin' 'otomatis'	[, yang, kamu, suka, dibatalin, otomatis,	[, suka, dibatalin, otomatis, pesanan,

Gambar 3.27 Stop removal

28. Pada penelitian ini, perhitungan jarak data pada Naïve Bayes dengan perhitungan *confusion matrix* yang digunakan untuk mengukur kinerja dari model klasifikasi di *machine learning. Confusion matrix* adalah salah satu *tools* analitik prediktif

yang menampilkan dan membandingkan nilai aktual atau nilai sebenarnya dengan nilai hasil prediksi model yang dapat digunakan. Nilai akurasi didapatkan dari jumlah data bernilai positif yang diprediksi positif dan data bernilai negatif yang diprediksi negatif dibagi dengan jumlah seluruh data di dalam dataset. *Precision* adalah peluang kasus yang diprediksi positif yang pada kenyataannya termasuk kasus kategori positif. Recall.*Recall* adalah peluang kasus dengan kategori positif yang dengan tepat diprediksi positif. Nilai *F1-Score* atau dikenal juga dengan nama *F-Measure* didapatkan *Precision* dan *Recall* antara kategori hasil prediksi dengan kategori sebenarnya. Pada hasil penelitian ini mendapatkan nilai *accuracy 88%.*

```
print(classification_report(Y_test,preds))
```

	precision	recall	f1-score	support
0	0.85	0.68	0.76	228
1	0.88	0.95	0.92	572
accuracy			0.88	800
macro avg	0.87	0.82	0.84	800
weighted avg	0.87	0.88	0.87	800

```
print(accuracy_score(Y_test,preds))
```

```
0.87625
```

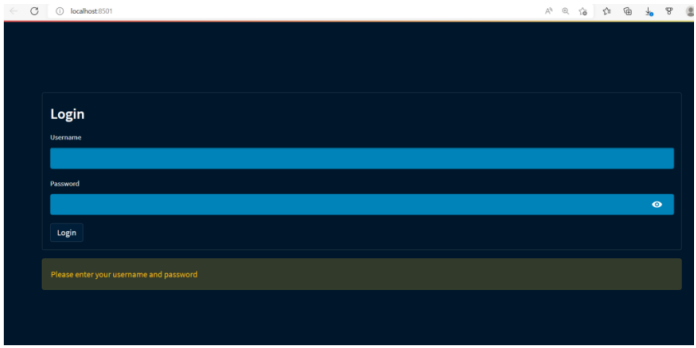
Gambar 3.28 Hasil *Accuracy*

29. Hasil perbandingan antara metode KNN dengan *Naïve Bayes.* Dari hasil yang telah dilakukan perhitungan menggunakan 4000 dataset dari reviwer dapat disimpulkan seperti dibawah ini, metode Naïve Bayes lebih unggul daripada metode KNN

Tabel 3. 4 Hasil Metodologi

No.	KNN				Naïve Bayes			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
Negatif	84 %	19%	30%	220	85%	68%	76%	228
Positif	76%	99%	86%	580	88%	95%	92%	572
Accuracy			77%	800			88%	800
Macro Avg	80%	59%	58%	800	87%	82%	84%	800
Weighted Avg	78%	77%	71%	800	87%	88%	87%	800

30. Implementasi Dashboard, *Dashboard login* pada perhitungan comparasi KNN dan *Naïve Bayes*

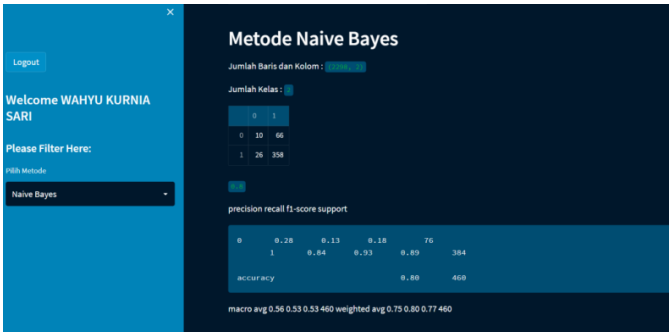


Gambar 3.31 Login *Dashboard*

31. Perhitungan Dataset pada *dashboard*

Gambar 3.32 Hasil *Dataset*

32. Perhitungan dengan model



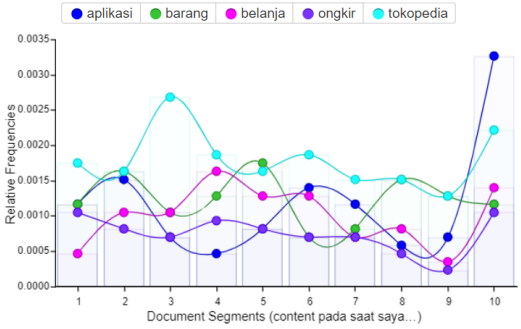
Gambar 3.33 Hasil Perhitungan Model

33. Visualisasi Data

Visualisasi tampilan berupa grafis atau visual dari informasi dan data.Dengan kata lain, data *visualization* mengubah kumpulan data menjadi hal lebih sederhana untuk ditampilkan. Dengan menggunakan elemen visual tersebut, pembaca akan lebih mudah memahami tren, *outliers,* dan pola dalam suatu data. data bertujuan untuk melihat jumlah frekuensi kata yang sering muncul pada setiap ulasan, visualisasi data dilakukan dengan memanfaatkan situs yaitu *https://voyant-tools.org/*. Hasil dari visualisasi data berdasarkan opinion. Berikut ini adalah visualisasi data.

Gambar 3.34 Visualisasi *Word Cloud*

Berikut ini adalah grafik terhadap 5 data tertinggi pada *process text mining* di *e- commerce* berdasarkan hasil *reviewer* pengunjung.



Gambar 3 .35 Grafik *Text Processing*

DAFTAR PUSTAKA

Rohandi, Mochamad Malik Akbar. 2017. “Effectiveness C2C e-commerce Media In Bandung (Case Study at Tokopedia.com and Bukalapak.com)”. Jurnal Managemen dan Bisnis (Performa) Universitas Islam Bandung

Hakim, A. (2018). Klasifikasi sentimen terhadap bukalapak dengan menggunakan metode naïve bayes classifier (Unpublished doctoral dissertation). Universitas Islam Negeri Sultan Syarif Kasim Riau.2021. “Peta E-Commerce Indonesia”. [diakses 4 Juli 2022]

Faadilah, A. (2020). Analisis sentimen pada ulasan aplikasi tokopedia di google play store menggunakan metode long short term memory (B.S. thesis). Fakul- tas Sains dan Teknologi Universitas Islam Negeri Syarif Hidayatullah.

Ratnawati, F. 2018. “Implementasi Algortima Naïve Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter”. Jurnal Invotek Polbeng, 3

Purnomo, W. G., dan Purnomo, P. P. (2018). Akurasi text mining menggunakan algoritma knearest neighbour pada data content berita sms. Format, 6(1), 1–13.

Pristiyani, R, I., Fauzi, M, A., dan Muflikhah, L. 2018. “Sentimen Analisis Peringkasan Review Film Menggunakan Metode Information Gain dan K-Nearest Neighbor”. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Computer (J-PTIIK) Universitas Brawijaya.

Mardi, Y. (2017). Data mining: Klasifikasi menggunakan algoritma c4. 5. Edik Informatika, 2(2), 213– 219.

Syaputri, Astia Weni. 2020. “Analisa Sentimen pada Ulasan Hotel Grand Elite di website Traveloka menggunakan Algoritma K-Nearest Neighbor”. Skripsi. Universitas Islam negeri Sultan Syarif Kasim Riau

Bhavani, A., dan B. Santhosh Kumar. "Sebuah Tinjauan Seni Negara dari Algoritma Klasifikasi Teks." *Konferensi Internasional ke-5 tentang Metodologi dan Komunikasi Komputasi (ICCMC) 2021* . IE, 2021.

Zheng, Ting, et al. "Compositionally Graded KNN-Based Multilayer Composite with Excellent Piezoelectric Temperature Stability." *Advanced Materials* 34.8 (2022): 2109175.

Purnomo, W. G., dan Purnomo, P. P. (2018). Akurasi text mining menggunakan algoritma knearest neighbour pada data content berita sms. Format, 6(1), 13

Pristiyani, R, I., Fauzi, M, A., dan Muflikhah, L. 2018. "Sentimen Analisis Peringkasan Review Film Menggunakan Metode Information Gain dan K-Nearest Neighbor". Jurnal Pengembangan Teknologi Informasi Dan Ilmu Computer (J-PTIIK) Universitas Brawijaya

Mardi, Y. (2017). Data mining: Klasifikasi menggunakan algoritma c4. 5. Edik Informatika, 2(2), 213– 219.

Rohandi, Mochamad Malik Akbar. 2017. "Effectiveness C2C e-commerce Media In Bandung (Case Study at Tokopedia.com and Bukalapak.com)". Jurnal Managemen dan Bisnis (Performa) Universitas Islam Bandung

Syaputri, Astia Weni. 2020. "Analisa Sentimen pada Ulasan Hotel Grand Elite di website Traveloka menggunakan Algoritma K-Nearest Neighbor". Skripsi. Universitas Islam negeri Sultan Syarif Kasim Riau

Bhavani, A., dan B. Santhosh Kumar. "Sebuah Tinjauan Seni Negara dari Algoritma Klasifikasi Teks." *Konferensi Internasional ke-5 tentang Metodologi dan Komunikasi Komputasi (ICCMC) 2021* . IE, 2021.

Zheng, Ting, et al. "Compositionally Graded KNN-Based Multilayer Composite with Excellent Piezoelectric Temperature Stability." *Advanced Materials* 34.8 (2022): 2109175.

Xiong, Lei, and Ye Yao. "Study on an adaptive thermal comfort model with K-nearest-neighbors (KNN) algorithm." *Building and Environment* 202 (2021): 108026.

Zhao, Dongdong, et al. "K-means clustering and kNN classification based on negative databases." *Applied Soft Computing* 110 (2021): 107732.

Shamrat, F. M. J. M., et al. "Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm." *Indonesian Journal of Electrical Engineering and*

Computer Science 23.1 (2021): 463-470.

Lin, Jinfeng, et al. "Significantly Photo-Thermochromic KNN-Based “Smart Window” for Sustainable Optical Data Storage and Anti-Counterfeiting." *Advanced Optical Materials* 9.17 (2021): 2100580.

Dogan, Alican, and Derya Birant. "Machine learning and data mining in manufacturing." *Expert Systems with Applications* 166 (2021): 114060.

Anandarajan, Murugan, Chelsey Hill, and Thomas Nolan. "Text preprocessing." *Practical Text Analytics*. Springer, Cham, 2019. 45-5 4

Trianto, Rahmawan Bagus, Andri Triyono, and Dhika Malita Puspita Arum. "Klasifikasi Rating Otomatis pada Dokumen Teks Ulasan Produk Elektronik Menggunakan Metode N-gram dan Naïve Bayes." *Jurnal Informatika Universitas Pamulang* 5.3 (2020): 295.

Hidayat, Assad, et al. "Implementasi Algoritma K-Nearest Neighbor dan Probabilistic Neural Network untuk Analisis Opini Masyarakat Terhadap Toko Online di Indonesia." *Seminar Nasional Teknologi Informasi Komunikasi dan Industri*. 2019.

GLOSARIUM

A

AUC: luas area di bawah curve ROC, atau integral dari fungsi ROC

Accuracy Score: Permodelan untuk mendapatkan hasil akurasi tertinggi

---

B

B2C: model penjualan dari bisnis langsung ke pelanggan

C

Case Folding: tahapan mengubah semua huruf campuran seperti huruf besar (*uppercase*) maupun huruf kecil (*lowercase*) menjadi *lowercase* semua.

Cross Validation: tahapan membagi training set dan testing set

Confusion Matrix: table yang menyatakan klasifikasi jumlah data yan benar dan jumlah data uji yang salah.

D

Development: Pengembangan, pembanganan, perkembangan, dan pertumbuhan. Oleh karena itu, definisi *development* merupakan setiap aktivitas yang terukur dan terstruktur untuk meningkatkan kemajuan organisasi atau perusahaan melalui inovasi mulai dari sisi produk hingga sumber daya manusia.

E

E-Commerce: kegiatan yang meliputi berbagai aktifitas seperti penyebaran, penjualan, pembelian, pemasaran produk yang berupa barang ataupun jasa yang memanfaatkan jaringan telekomunikasi.

F

Filtering: Tahapan membersihkan data dari tanda baca, simbol, maupun elemen yang tidak dibutuhkan

---

G

Google Play Store: toko online yang dikunjungi pengguna untuk menemukan aplikasi, game, film, acara TV, buku, dan konten lainnya.

H

Hash: Fungsi apa pun yang dapat digunakan untuk memetakan data dengan ukuran arbitrer ke nilai ukuran tetap.

I

***Indirect* Material**: Bahan Baku Tidak Langsung

Instance: Memiliki struktur data yang sama dengan instance lain, tetapi nilai yang disimpan dalam instance tersebut terpisah.

J

Jupiter Notebook: proyek dengan tujuan untuk mengembangkan perangkat lunak sumber terbuka, standar terbuka, dan layanan untuk komputasi interaktif di berbagai bahasa pemrograman

---

K

KNN: Metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut.

K-Fold Cross Validation: Validasi silang untuk pada dataset yang dibagi sebanyak k lipatan. Pada setiap lipatan akan dipakai satu kali sebagai data uji dan lipatan sisanya dipakai sebagai data latih

---

L

LDA: Generalisasi diskriminan linear Fisher, yaitu sebuah metode yang digunakan dalam ilmu statistika, pengenalan pola dan pembelajaran mesin untuk mencari kombinasi linear fitur yang menjadi ciri atau yang memisahkan dua atau beberapa objek atau peristiwa.

M

Machine Learning: Pembelajaran pendekatan algoritma untuk membuat prediksi dan keputusan berdasarkan pengalaman dan data

Market Palace: *platform* yang disediakan untuk para penjual berkumpul dan bisa menjual barang atau jasanya kepada pelanggan meski tanpa bertemu secara fisik.

Matplotlib: Visualisasi data membuat plot serta grafik

Multinominal NB: Permodelan yang dilakukan pada perhitungan di *naïve bayes*

N

Naïve Bayes: Sebuah metoda klasifikasi yang berakar pada teorema *Bayes*. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yg dikemukakan oleh ilmuwan Inggris *Thomas Bayes*, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai *Teorema Bayes*.

Numpy: Menyimpan data dalam bentuk *array*, yaitu kumpulan variable yang memiliki tipe data yang sama

---

O

Online: istilah saat kita sedang terhubung dengan internet atau dunia maya, baik itu terhubung dengan akun media sosial kita, email dan berbagai jenis akun lainnya yang kita pakai atau gunakan lewat internet.

P

Python: Bahasa pemrograman open source yang memakai contoh skrip (scripting languange) berorientasi objek . Python bersifat generik dengan juru bahasa dan dapat digunakan di domain aplikasi luas dan merupakan bahasa pemrograman tingkat tinggi yang fleksibel, sederhana, dan dinamis.

Pre-Processing Data: Menyeleksi data dan mengubahnya menjadi data yang lebih terstruktur dengan beberapa proses *tokenizing, filtering, stemming*

Pandas: Kebutuhan analisis, manipulasi dan pembersihan data, dengan pendukung CSV untuk dijadikan objek *python* dengan *rows & columns* (dataframe)

Q

QR: *Quick Ratio*

Queri: kemampuan untuk menampilkan data dari database untuk diolah lebih lanjut yang biasanya diambil dari tabel tabel dalam database.

R

Roc: alat ukur performance untuk classification problem dalam menentukan threshold dari suatu model.

Re: Urutan karakter yang membentuk pola pencarian yang digunakan untuk memeriksa apakah *string* berisi pola pencarian yang ditentukan

---

S

Scrapping : Teknik atau metode otomatisasi yang memungkinkan seseorang untuk mengekstrak data dari sebuah website, database, aplikasi enterprise, atau sistem legacy yang kemudian dapat menyimpannya ke dalam sebuah file dengan format tabular atau spreadsheet

Stopword Removal : Tahapan menghapus kata sambung

Stopword : *Stop words* adalah kata umum yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna.

Stemmer: Pemotongan imbuhan (awalan, akhiran, sisipan, kombinasi) yang dijalankan dengan judul aslinya

Stack: Salah satu struktur data yang digunakan untuk menyimpan sekumpulan objek ataupun variabel.

String: Tipe data untuk teks yang terdiri dari gabungan huruf,angka dan berbagai karakter

---

T

TF-IDF: perhitungan bobot *term* pada sebuah dokumen berdasarkan seringnya kata tersebut muncul dimana bobot tersebut mengindikasikan pentingnya sebuah *term* terhadap dokumen, semakin banyak *term* tersebut muncul pada dokumen maka semakin tinggi nilai *term*

Text Preprocessing: Proses mengubah bentuk data yang sebelumnya tidak terstruktur kedalam bentuk data yang terstruktur.

Taksonomi: kategorisasi benda atau konsep, serta prinsip-prinsip yang mendasari kategorisasi tersebut.

Text Mining: Suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan tools analisis yang merupakan komponen-komponen dalam data mining yang salah satunya adalah kateagorisasi.

U

User Generated Content: Penentu keputusan ketika akan membeli sebuah produk.

V

Variable Budget: Merencanakan anggaran secara sistematis dan menjelaskan secara lebih rinci tingkat perubahan biaya kegiatan perusahaan relatif terhadap biaya tidak langsung.

W

Word Cloud: ambar yang menunjukkan daftar kata-kata yang digunakan dalam sebuah teks, umumnya semakin banyak kata yang digunakan semakin besar ukuran kata tersebut dalam gambar.

Word Tokenize: Operasi yang memisahkan teks menjadi potongan-potongan berupa token, dapat berupa potongan huruf, kata, atau kalimat, sebelum dianalisis lebih lanjut

TENTANG PENULIS

Wahyu Kurnia Sari, lahir di Kabupaten Karanganyar, Jawa Tengah pada tanggal 21 Agustus. Pendidikan tingkat dasar hingga menengah dan atas ditempuh di Karanganyar. Melanjutkan untuk merantau ke kota Bandung untuk menempuh pendidikan Diploma di Universitas Logistik Bisnis Internasional, Bandung program studi D4 Teknik Informatika. Pada tahun ini 2022 allhamdulillah dapat menyelesaikan studi tepat waktu dengan menyandang gelar S.Tr.Kom. Dengan adanya buku ini semoga dapat bermanfaat dan berguna bagi pembaca, mohon maaf apabila masih banyak kekurangan dalam isi maupun penulisan. Semoga dengan buku ini dapat membuat semangat penulis untuk berkarya dan bagi pembaca dapat terinspirasi.

SINOPSIS

Google play store menyediakan fitur kepada pengguna untuk dapat memberikan ulasan terhadap aplikasi yang digunakan, salah satunya yaitu ulasan terhadap aplikasi *e-commerce*. *E-Commerce* salah satu aplikasi yang menggunakan model bisnis *marketplace* dan mall *online*. Peningkatan jumlah pengunjung dan pengguna baru yang akan mengunduh aplikasi *e-commerce* berkaitan dengan komentar terhadap aplikasi. Ulasan aplikasi dibagi menjadi ulasan *review* produk positif dan negatif. Guna meningkatkan kepercayaan, pengguna bisa melihat komentar terhadap aplikasi. Oleh karena itu, diperlukan suatu teknik pengolahan data dan analisis terhadap komentar. Komentar tersebut dapat dianalisis dengan menggunakan *text mining*. *Text mining* merupakan teknik dalam pengambilan informasi dari sejumlah data tak terstruktur dari sebuah topik tertentu yang memiliki kualitas tinggi serta dapat diperoleh data-data permasalahan dalam teks. Penulisan buku ini bertujuan untuk mengklasifikasikan ulasan barang yang dipesan di *e-commerce* menggunakan algoritma *K-Nearest Neighbor* dan *Naïve Bayes.* Algoritma *K-Nearest Neighbor* merupakan salah satu algoritma yang terdapat dalam teknik klasifikasi. Pada tahun 2021, berdasarkan peta *e-commerce* Indonesia, Selanjutnya hasil akurasi akan di*compare* dengan metode *Naïve bayes* untuk mendapatkan hasil yang lebih unggul.