BAB II

TINJAUAN PUSTAKA

2.1 Data Mining

Data *Mining* adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran computer (machine learning) untuk menganalisis dan mengekstraksi pengetahuan *(knowledge)* secara otomatis. Data *Mining* merupakan proses iteratif dan interaktif untuk mengemukakan pola atau model baru sahih (sempurna), bermanfaat dan dapat dimengerti dalam suatu *database* yang sangat besar *(massive databse).*

Menurut Abdan, Data *Mining* berisi pencarian tren atau pola yang diinginkan dalam *database* besar untuk membantu pengambilan keputusan di waktu yang akan datang. Pola – pola ini dikenali oleh perangkat tertentu yang dapat memberikan suatu analisa data berguna dan berwawasan yang kemudian dapat dipelajari dengan lebih teliti, yang mungkin saja mengguankan perangkat pendukung keputusan yang lainnya (Hermawati, 2013).

2.2 Text Mining

Text mining dapat diartikan sebagai penemuan informasi yang baru dan tidak diketahui sebelumnya oleh komputer, dengan secara otomatis mengekstrak informasi dari sumber-sumber yang berbeda. Kunci dari proses ini

adalah menggabungkan informasi yang berhasil diekstraksi dari berbagai sumber (Hearst, 2003). Sedangkan menurut (Harlian, 2006) text mining memiliki definisi menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen.

Menurut Khotimah dalam penelitiannya , perbedaan antara *text mining* dengan data mining terletak pada sumber data yang digunakan. Dalam *text mining* pola-pola yang di ekstrak dari data tekstual yang tidak terstruktur bukan berasal dari suatu *database*. Beberapa kesamaanya adalah data yang digunakan merupakan data besar dan data berdimensi tinggi dengan struktur yang terus berubah. Dalam *data mining* data yang diolah adalah data yang terstruktur dari proses *warehousing* sehingga lebih mudah diproses oleh mesin/*computer*. Analisis teks lebih sulit karena biasanya hanya digunakan sebagai konsumsi manusia. Ditambah struktur teks yang kompleks, struktur yang tidak lengkap, Bahasa yang berbeda, dan arti yang tidak standar. Oleh karena itu pada umumnya digunakan *Natural Language Processing* untuk analisis teks yang tidak berstruktur tersebut. Tahap-tahap *text mining* umumnya adalah text preprocessing dan feature selection (Feldman & Sanger, 2007.

2.3 Web Scraping

Mitchell (2018) mendefinisikan bahwa yang dimaksudkan sebagai Web *Scraping* adalah kegiatan pengumpulan (gathering) data yang bersumber dari Internet. Tujuan dari Web *Scraping* adalah mendapatkan data untuk kemudian melakukan ekstraksi informasi yang dimiliki oleh data tersebut. Cara kerja Web *Scraping* adalah dengan mengakses halaman Web, memilih elemen data yang ada dalam halaman tersebut, melakukan ekstraksi dan transformasi bila diperlukan, dan terakhir menyimpan data tersebut menjadi datasetterstruktur (Boeing dan Waddell, 2017)

2.4 Text Preprocessing

Preprocessing merupakan proses untuk mempersiapkan data mentah sebelum dilakukan proses lain. Pada umumnya, *preprocessing* data dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah yang diproses oleh sistem. *Preprocessing* sangat penting dalam pembuatan *analisis sentimen*, terutama untuk media sosial yang sebagian besar berisi kata – kata atau kalimat yang tidak formal dan tidak terstruktur serta memiliki *noise* yang besar. Ada tiga model *preprocessing* untuk kalimat atau teks dengan *noise* yang besar (A Clark, 2003). Tiga model tersebut adalah :

1. *Orthographic Model*. Model ini dipergunakan untuk memperbaiki kata atau kalimat yang memiliki kesalahan dari segi bentuk kata atau kalimat. Contoh

kesalahan yang diperbaiki dengan *Orthographic model* adalah huruf kapital di tengah kata.

2. *Error Model*. Model ini dipergunakan untuk memperbaiki kesalahan dari segi kesalahan eja atau kesalahan penulisan. Ada dua jenis kesalahan yang dikoreksi dengan model ini yaitu kesalahan penulisan dan kesalahan eja. Kesalahan penulisan mengacu pada kesalahan pengetikan sedangkan kesalahan eja muncul ketika penulis tidak tahu ejaannya benar atau salah.
3. *White Space Model*. Model ke tiga ini mengacu pada pengoreksian tanda baca. Contoh kesalahan untuk model ini adalah tidak menggunakan tanda titik '.' di akhir kalimat. Namun, model ini tidak terlalu signifikan, terutama ketika berhadapan dengan media sosial yang jarang mengindahkan tanda baca. (Mujilahwati, 2016).

Secara umum proses yang dilakukan dalam tahapan *preprocessing* adalah sebagai berikut :

a. Spelling Normalization

Merupakan tahap awal yang harus dilakukan untuk mendapatkan dokumen data yang baik. Perlakuan yang dilakukan yaitu memperbaiki kata-kata yang terdapat salah ejaan atau disingkat menjadi bentuk tertentu. Proses ini dilakukan dengan bantuan Microsoft Excel dan software Rstudio.

b. Case folding

Merupakan proses penyeragaman bentuk huruf menjadi huruf kecil semua antara "a" sampai dengan "z". Dengan tujuan agar kata yang ditulis dengan

huruf awal kapital dan huruf kecil tidak terdeteksi mempunyai arti yang berbeda.

c. *Tokenizing*

Merupakan proses pemisahan teks menjadi potongan kata yang disebut dengan token. Bertujuan untuk mendapatkan potongan kata yang akan menjadi entitas serta memiliki nilai dalam matriks dokumen teks yang akan dianalisis.

d. *Filtering*

Kata dan tanda baca yang nantinya tidak bernilai atau tidak berarti akan dieliminasi seperti url, angka, tanda baca, hastag, kata hubung, kata ganti dan lainya. Pemilihan kata yang bermakna menggunakan Stopwords (menghilangkan kata yang kurang penting). Kata penghubung yang akan dihilangkan yaitu:

a. Penghubung antar kata, seperti dan, atau, serta.

b. Preposisi, seperti di, ke, pada.

Setelah dilakukan preprocessing dan didapatkan data terstruktur tersebut yang telah diberi label positif dan negatif.

2.5 Analisis Sentimen

Analisis sentimen atau bisa juga disebut *opinion mining* merupakan sebuah cabang penelitian di *domain text mining* yang mulai banyak dilakukan pada

tahun 2013. Lee dan Pang menjelaskan analisis sentimen atau dikenal sebagai *opinion mining* adalah proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi (Lee & Pang, 2008).

Secara umum, opinion mining diperlukan untuk mengetahui sikap seorang pembicara atau penulis sehubungan dengan beberapa topik atau polaritas kontekstual keseluruhan dokumen.Sikap yang diambil mungkin menjadi pendapat atau penilaian atau evaluasi (teori appraisal), keadaan afektif (keadaan emosional penulis saat menulis) atau komunikasi emosional (efek emosional penulis yang ingin disampaikan pada pembaca) (Saraswati, 2011).

Analisis sentimen dapat digunakan dalam berbagai kemungkian *domain*, dari produk konsumen, jasa kesehatan, jasa keuangan, peristiwa sosial dan politik pada pemilu. Kecendrungan penelitian tentang analisis sentimen berfokus pada pendapat yang menyatakan atau menyiratkan suatu sentimen positif atau negatif. Pendapat mewakili hampir semua aktivitas manusia, karena pendapat dapat mempengaruhi terhadap perilaku seseorang. Setiap kali kita perlu membuat keputusan, kita ingin tahu pendapat orang lain. Dalam dunia nyata, bisnis dan organisasi selalu ingin melihat opini publik tentang suatu produk atau jasa (Liu, 2012). Dengan analisis sentimen, suatu bisnis dapat melacak produk-produk, merek dan orang-orang misalnya dan menentukan apakah dilihat positif atau negatif di web. Hal ini memungkinkan bisnis untuk mengetahui komentar buruk, persepsi produk baru dan persepsi terhadap suatu merek tertentu.

2.6 Klasifikasi

Teknik klasifikasi adalah salah satu dari teknik *data mining* yang termasuk *supervised learning. Supervised learning* artinya proses pembentukan sebuah korespondensi menggunakan sebuah *training dataset.* Tujuannya adalah untuk memprediksi target dari beberapa atribut (Zaki & Meira, 2014). Terdapat pada dua pekerjaan utama pada klasifikasi yaitu melakukan *training* untuk disimpan sebagai prediksi dan melakukan *testing* untuk proses klasifikasi agar diketahui di label mana objek data tersebut (Liu, Loh, & Sun, 2009).

2.7 Naïve Bayes *Classifier* (NBC)

Teorema Bayes merupakan teorema yang mengacu pada probabilitas bersyarat (Siang, 2005). Secara umum teorema Bayes dapat dinotasikan pada persamaan berikut.

$$P(Aj \mid Bi) = \frac{P(Bj|Ai)P(Aj)}{P(Bi)} \qquad (2.4)$$

Dimana :

P(A|B) : Peluang kategori j, ketika terdapat kemunculan kata i

P(A|B) : Peluang kata i masuk ke dalam kategori j

P (A) : Peluang kemunculan kategori j

P(B) : Peluang kemunculan kata

Naïve Bayes *Classifier* merupakan salah satu algoritma yang digunakan dalam klasifikasi data *mining*. Klasifikasi sendiri merupakan penentuan sebuah

record data baru ke salah satu dari beberapa kategori (atau kelas) yang telah didefinisikan. Dan pada Naive Bayes *Classifier* ialah mengadopsi teorema Bayessian. Bayes merupakan teknik prediksi berbasis probabilitas sederhana yang berdasar pada penerapan teorema bayes dengan asumsi independensi (ketidaktergantungan yang kuat (naif)). Dengan kata lain ,dalam Naive Bayes yang digunakan adalah "model fitur independen".

Menurut Olson Delen (2008) menjelaskan Naïve Bayes unt setiap kelas keputusan, menghitung probabilitas dg syarat bahwa kelas keputusan adalah benar, mengingat vektor informasi obyek. Algoritma ini mengasumsikan bahwa atribut obyek adalah independen. Probabilitas yang terlibat dalam memproduksi perkiraan akhir dihitung sebagai jumlah frekuensi dr " master " tabel keputusan. Naive Bayes *Classifier* bekerja sangat baik dibanding dengan model *classifier* lainnya. Hal ini dibuktikan oleh Xhemali , Hinde Stone dalam jurnalnya *"Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages"* mengatakan bahwa "*Naïve Bayes Classifier* memiliki tingkat akurasi yg lebih baik dibandingmodel *classifier* lainnya". Keuntungan penggunan adalah bahwa metoda ini hanya membutuhkan jumlah data pelatihan *(training data)* yg kecil unt menentukan estimasi parameter yg diperlukan dalam proses pengklasifikasian. Karena yg diasumsikan sebagai variable independent, maka hanya varians dr suatu variable dalam sebuah kelas yg dibutuhkan unt menentukan klasifikasi, bukan keseluruhan dr matriks kovarians.

Terdapat dua tahap dalam klasifikasi tweet. Tahap pertama adalah pelatihan terhadap tweet yang telah diketahui kategorinya. Sedangkan tahap kedua adalah proses klasifikasi tweet yang belum diketahui kategorinya (Falahah dan Nur, 2015). Dalam algoritma NBC setiap dokumen direpresentasikan dengan pasangan atribut "a_1, a_3,...,, a_n" dimana a_1 adalah kata pertama, a_2 adalah kata kedua dan seterusnya. Sedangkan V adalah himpunan kategori tweet. Pada saat klasifikasi algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan ($\boldsymbol{V_{MAP}}$). Adapun persamaan $\boldsymbol{V_{MAP}}$ adalah sebagai berikut

$$\boldsymbol{VMAP} = \underset{\boldsymbol{Aj}}{\boldsymbol{argmax}}\boldsymbol{P}(\boldsymbol{Vj}) \prod_{i-1}^{n} \boldsymbol{P}(\boldsymbol{ai}|\boldsymbol{vj}) \qquad (2.5)$$

Nilai ($\boldsymbol{vj}$) dihitung pada saat training, didapat dengan rumus sebagai berikut:

$$\boldsymbol{P}(\boldsymbol{vj}) = \frac{|\boldsymbol{doc\ j}|}{|\boldsymbol{training}|} \qquad (2.6)$$

$|\boldsymbol{doc\ j}|$ merupakan jumlah *review* pada kategori j dalam training. Sedangkan |training|merupakan jumlah *review* dalam data yang digunakan untuk training. Setiap probabilitas kata $\boldsymbol{ai}$ pada setiap kategori $\boldsymbol{P}(\boldsymbol{ai}|\boldsymbol{vj})$, dihitung pada saat training.

$$\boldsymbol{P}(\boldsymbol{ai}|\boldsymbol{vj}) = \frac{\boldsymbol{ni+1}}{|\boldsymbol{n+kosakata}|} \qquad (2.7)$$

Di mana,

ni : jumlah kemunculan kata $\boldsymbol{ai}$ dalam *review* yang berkategori $\boldsymbol{vj}$

n : banyaknya seluruh kata *review* dengan kategori $\boldsymbol{vj}$

|kosakata| : banyaknya kata dalam data training

2.8 Seleksi Fitur

Seleksi fitur adalah salah satu teknik terpenting dan sering digunakan dalam *preprocessing*. Teknik ini mengurangi jumlah fitur yang terlibat dalam menentukan suatu nilai kelas target, mengurangi fitur irelevan, berlebihan dan data yang menyebabkan salah pengertian terhadap kelas target yang membuat efek segera bagi aplikasi. Tujuan utama dari seleksi fitur ialah memilih fitur terbaik dari suatu kumpulan fitur data. (Maulida, Suyatno, & Hatta, 2016)

2.9 Information Gain (IG)

Information Gain merupakan teknik seleksi fitur yang memakai metode scoring untuk nominal ataupun pembobotan atribut kontinu yang didiskritkan menggunakan maksimal entropy.Suatu entropy digunakan untuk mendefinisikan nilai *Information Gain*. Entropy menggambarkan banyaknya informasi yang dibutuhkan untuk mengkodekan suatu kelas. *Information Gain* (IG) dari suatu term diukur dengan menghitung jumlah bit informasi yang diambil dari prediksi kategori dengan ada atau tidaknya term dalam suatu dokumen. (Maulida, Suyatno, & Hatta, 2016)

Teknik seleksi fitur dengan *information gain* artinya adalah memilih simpul fitur dari pohon keputusan berdasar nilai *information gain*. Nilai *information gain*

sebuah fitur diukur dari pengaruh fitur tersebut terhadap keseragaman kelas pada data yang dipecah menjadi subdata dengan nilai fitur tertentu. Keseragaman kelas (entropy) dihitung pada data sebelum dipecah dengan persamaan 2.1 dan pada data setelah dipecah dengan persamaan 2.2 berikut ini.

$$Entropy\ (S) = \sum_{i=1}^{k} (Pi)\ log2(Pi) \qquad (2.1)$$

Dengan nilai Pi adalah proporsi data S dengan kelas i. K adalah jumlah kelas pada output S.

$$Entropy\ (S, A) = \sum_{i=1}^{v} (\frac{Sv}{v} * Enteropy\ (Sv)) \qquad (2.2)$$

Dengan nilai v adalah semua nilai yang mungkin dari atribut A, Sv adalah subset sari S dimana atribut A bernilai v. Nilai *information gain* dihitung dengan persamaan 2.3 berikut ini:

$$Gain\ (S, A) = Entropy\ (S) - Entropy\ (S, A) \qquad (2.3)$$

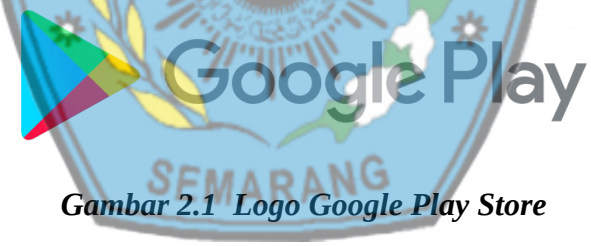
Dengan nilai Gain (S,A) adalah nilai *information gain*. Entropy (S) adalah nilai entropy sebelum pemisah. Entropy(S,A) adalah nilai entropy setelah pemisah. Besarnya nilai *information gain* menunjukkan seberapa besar pengaruh suatu atribut terhadap pengklasifikasian data. (Rasywir & Purwarianti, 2015)

2.10 *Google Play Store*

Google Play Store adalah layanan konten digital milik Google yang melingkupi toko online untuk produk-produk seperti musik/lagu, buku, aplikasi,

permainan, ataupun pemutar media berbasis cloud. Layanan ini dapat diakses baik melalui web, aplikasi android (Play Store) dan Google TV. *Google Play Store* mulai dikenalkan pada bulan Maret 2012 sebagai pengganti dari Google Play dan layanan musik Google.

Google Play milik Google saat ini telah menyediakan sekitar 700.000 aplikasi mobile menurut AppBrain7. Setelah beberapa bulan, mungkin ada lebih dari sepuluh ribu komentar tekstual dari aplikasi baru yang diluncurkan di GooglePlay (Liu et al, 2013). Hal ini sangat sesuai untuk penerapan analisis sentimen terhadap komentar-komentar pengguna tentang aplikasi pada *Google Play Store* untuk membantu pengguna dalam memberikan pertimbangan untuk menginstall aplikasi dan developer untuk menganalisis aplikasinya.



Gambar 2.1 Logo Google Play Store

2.11 E-Commerce

Electronic commerce (disingkat *E-Commerce)* sebagai sarana berbisnis menggunakan jaringan komputer, sebenarnya adalah dikenal sejak 20 tahun lalu sejak akhir tahun 70-an dan awal tahun 80-an. Generasi pertama *E-Commerce* dilakukan hanya antar perusahaan berupa transaksi jual beli yang difasilitasi oleh *Electronic Data Intechange (EDI)* dalam transaksi jual beli

elektronik ini banyak aspek-aspek yang bersentuhan langsung maupun tidak langsung (Firdaus, 2015). *E-Commerce* adalah proses pembelian dan penjualan antara dua belah pihak di dalam suatu perusahaan dengan adanya pertukaran barang, jasa, atau informasi melalui media internet (Indrajit, 2001). Onno (2000) memberikan pengertian tentang *E-Commerce* yaitu asset dinamis teknologi, aplikasi, dan proses bisnis yang menghubungkan perusahaan, konsumen dan komunitas melalui elektronik dan perdagangan barang, pelayanan dan informasi yang dilakukan secara elektronik. Sedangkan menurut Berkatulloh dan Prasetyo (2005) menjelaskan bahwa *E-Commerce* adalah kegiatan-kegiatan bisnis yang menyangkut konsumen *(consumers),* manufaktur (manufaktures), *service providers* dan pedagang perantara *(intermediaries)*, dengan menggunakan jaringan-jaringan computer *(computer networks)* yaitu internet.

a. Komponen *E-Commerce*

Pada *E-Commerce* terdapat mekanisme-mekanisme tertentu yang unik dan berbeda dibandingkan dengan mekanisme-mekanisme yang terdapat pada traditional commerce. Dalam mekanisme pasar *E-Commerce,* terdapat beberapa komponen yang terlibat, yakni (Turban & King, 2002):

1. Customer

Customer merupakan para pengguna Internet yang dapat dijadikan sebagai target pasar yang potensial untuk diberikan penawaran berupa produk, jasa, atau informasi oleh para penjual.

2. Penjual

Penjual maerupakan pihak yang menawarkan produk, jasa, atau informasi kepada para customer baik individu maupun organisasi. Proses penjualan dapat dilakukan secara langsung melalui *website* yang dimilik i oleh penjual tersebut melalui *marketplace.*

3. Produk

Salah satu perbedaan antara *E-Commerce* dengan *traditional commerce* terletak pada produk yang dijual. Pada dunia maya, penjual dapat menjual produk digital yang dapat dikirimkan secara langsung melalui Internet.

4. Infrastruktur

Infrastruktur pasar yang menggunakan media elektronik meliputi perangkat keras, lunak dan juga sistem jaringannya.

b. Jenis *E-Commerce* di Indonesia

Dua situs *marketplace* di Indonesia yang memperbolehkan penjual langsung berjualan barang di website ialah Tokopedia dan Shopee. Ada juga situs marketplace lainnya yang mengharuskan penjual menyelesaikan proses verifikasi terlebih dahulu seperti Blanja dan Elevenia. Cara model bisnis *E-Commerce* ini meraup keuntungan

adalah dengan memberlakukan layanan penjual premium, iklan premium, dan komisi dari setiap transaksi. Situs marketplace seperti ini lebih cocok bagi penjual yang lebih serius dalam berjualan online. Biasanya penjual memiliki jumlah stok barang yang cukup besar dan mungkin sudah memiliki toko fisik.

2.11.1 Tokopedia

Tokopedia merupakan perusahaan perdagangan elektronik atau sering disebut toko daring. Sejak didirikan pada tahun 2009, Tokopedia telah bertransformasi menjadi sebuah *unicorn* yang berpengaruh tidak hanya di Indonesia tetapi juga di Asia Tenggara. Hingga saat ini, Tokopedia termasuk marketplace yang paling banyak dikunjungi oleh masyarakat Indonesia.

Tujuan Tokopedia yaitu turut mendukung para pelaku Usaha Mikro Kecil dan Menengah (UMKM) dan perorangan untuk mengembangkan usaha mereka dengan memasarkan produk secara daring dengan Pemerintah dan pihak-pihak lainnya. Salah satu program kolaborasi yang diinisasi oleh Tokopedia adalah acara tahunan MAKERFEST yang diadakan sejak bulan Maret 2018.

Sejak tahun 2018, Tokopedia juga menghadirkan Tokopedia *Center*. Melalui Tokopedia *Center,* pengunjung dapat melakukan transaksi secara *online-to-offline* (O2O), membayar tagihan, membeli tiket, mendapatkan

informasi mengenai cara menggunakan aplikasi Tokopedia, belanja secara interaktif, sampai mencari inspirasi untuk memulai usaha daring secara gratis.

Gambar 2.2 Logo Tokopedia

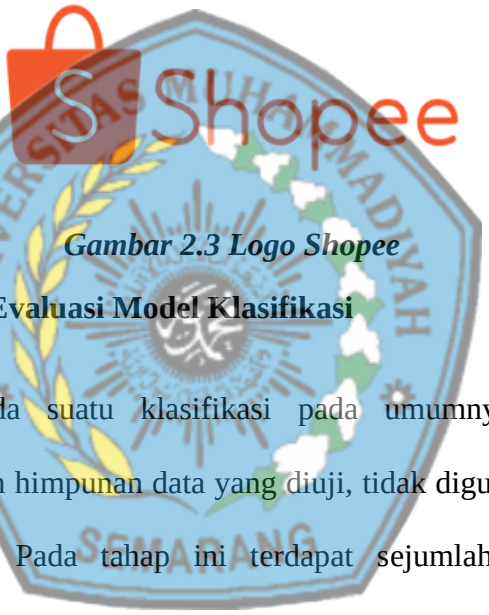
2.11.2 Shopee

Shopee adalah *platform E-Commerce* terkemuka di Asia Tenggara dan Taiwan.Ini adalah *platform* yang dirancang untuk kawasan ini, memberikan pelanggan pengalaman belanja online yang mudah, aman, dan cepat melalui pembayaran dan dukungan logistik yang kuat.

Shopee bertujuan untuk terus meningkatkan platformnya dan menjadi tujuan pilihan *E-Commerce* kawasan. Shopee memiliki berbagai pilihan kategori produk mulai dari elektronik konsumen hingga rumah & hidup, kesehatan & kecantikan, bayi & mainan, mode dan peralatan kebugaran.

Perusahaan Sea, pertama kali diluncurkan di Singapura pada 2015, dan sejak itu memperluas jangkauannya ke Malaysia, Thailand, Taiwan, Indonesia, Vietnam, dan Filipina. Sea adalah pemimpin dalam hiburan digital, *E-Commerce,* dan layanan keuangan digital di seluruh Asia Tenggara. Misi *Sea* adalah memperbaiki kehidupan konsumen dan usaha kecil dengan teknologi, dan terdaftar di NYSE dengan simbol SE.

Tujuan dari Shopee adalah menyediakan *platform* untuk menghubungkan pembeli dan penjual dalam satu komunitas. Karena berbelanja di perangkat seluler menjadi kebiasaan baru, Shopee bertujuan untuk terus meningkatkan platformnya untuk menghadirkan pengalaman belanja yang mulus dan menyenangkan bagi semua pengguna dan menjadi platform pilihan *E-Commerce* di era ini (Shopee, 2019).



Gambar 2.3 Logo Shopee

2.12 Ukuran Evaluasi Model Klasifikasi

Evaluasi pada suatu klasifikasi pada umumnya dilakukan dengan menggunakan sebuah himpunan data yang diuji, tidak digunakan dalam pelatihan klasifikasi tersebut. Pada tahap ini terdapat sejumlah ukuran yang dapat digunakan untuk menilai kembali atau mengevaluasi model klasifikasi, yaitu accuracy atau tingkat pengenalan, tingkat kesalahan atau kekeliruan klasifikasi, recall atau sensitivity atau true positif, specificity atau true negatif dan precision.

Model klasifikasi yang telah dibuat yaitu pemetaan dari suatu baris data dengan keluaran sebuah hasil prediksi kelas atau target dari data tersebut. Pada klasifikasi ini terdapat dua kelas sebagai luarannya yang disebut klasifikasi

biner. Kedua kelas tersebut biasa diinterpretasikan dalam {0,1}, {+1,-1} atau {positif,negatif}.

Pada proses evaluasi klasifikasi terdapat empat kemungkinan yang terjadi yaitu proses pengklasifikasian pada suatu baris data. Jadi, jika data positif dan diprediksi positif maka akan dihitung sebagai true positif, bahkan jika data itu diprediksi negatif maka akan dihitung sebagai false negatif. Jika data negatif dan diprediksi negatif maka akan dihitung sebagai true negatif, tetapi jika data tersebut diprediksi positif maka akan dihitung sebagai false positif. Hasil klasifikasi biner pada suatu dataset yang dipresentasikan dalam bentuk matriks 2×2 yaitu dinamakan confusion matrix. Berikut merupakan contoh dari matriks

Tabel 2.1 Confusion Matrix

Confusion Matrix		True Class	
		Positive	Negatif
Predicted Class	Positive	*True Postive* (TP)	*False Positive* (FP)
	Negatif	*False Negatif* (FN)	True Negatif (TN)

Confusion Matrix bermanfaat untuk menganalisis kualitas *classifier* dalam mengenali tuple-tuple dari kelas yang ada. TP dan TN menyatakan pada *classifier* mengenali tuple dengan benar , artinya tuple positif dikenali sebagai positif dan tuple negatif dikenali sebagai negatif. Sedangkan, FP dan FN menyatakan bahwa *classifier* salah dalam mengenali tuple, tuple negatif dikenali

sebagai positif dan tuple negatif dikenali sebagai positif. Ada pula dalam formula perhitungan performa klasifikasi yaitu nilai akurasi biasa ditampilkan dalam presentase.

2.13 Akurasi

Akurasi adalah nilai ketepatan dimana pengguna memprediksi suatu kata sesuai dengan jawaban suatu sistem. Berikut perhitungan nilai akurasi

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