PROPOSAL PROGRAM KREATIVITAS MAHASISWA

AI Tools dalam Mengklasifikasi Hasil Ulasan dalam Pelayanan e-Commerce di Indonesia

BIDANG KEGIATAN:
GEMASTIK_DATA MINING

Diusulkan Oleh:

Armadhani Hiro Juni Permana	NIM 1301190234	2019
Shabrina Retno Ningsih	NIM 1301194162	2019
Azka Zainur Azifa	NIM 1303194255	2019

TELKOM UNIVERSITY
BANDUNG
2021

ABSTRAK

Pada saat ini e-Commerce banyak digunakan oleh masyarakat untuk melakukan transaksi jual beli online yang sangat memudahkan masyarakat dalam berbelanja. Banyak e-Commerce yang menyediakan pilihan toko dan barang dengan menyertakan informasi berupa komentar review masing-masing produk yang dijual. Para pembeli bisa melakukan penilaian atau hanya sekedar melihat penilaian dari pembeli lain sebelum membeli produk yang diinginkan. Penilaian yang paling banyak dilihat adalah penilaian dengan bintang 5, sedangkan jika diperhatikan masih banyak review bintang 5 tetapi isi reviewnya menggambarkan ketidakpuasan pembeli terhadap produk tersebut. Oleh karena itu, analisis sentimen adalah salah satu solusi untuk mengelompokkan review yang sangat puas dan review tidak puas. Proses dilakukan dengan mengklasifikasikan review lalu pre-processing kemudian dilanjut untuk akurasi dengan algoritma Naive Bayes dan Logistic Regression. Nilai akurasi yang didapat dari penelitian ini adalah Naive Bayes mencapai akurasi hingga 90% dan Logistic Regression mencapai akurasi hingga 85%. Hasil akurasi Naive Bayes lebih baik dibandingkan Logistic Regression karena Naive Bayes bersifat sederhana dan memiliki waktu komputasi yang tinggi.

***Kata kunci** : Naive Bayes, Logistic Regression, e-Commerce, akurasi.*

BAB 1 PENDAHULUAN

1.1 Latar Belakang

E-Commerce adalah proses pembelian maupun penjualan produk secara elektronik [11]. Masyarakat membeli kebutuhan sehari-hari dengan *e-Commerce* dan mendapatkan produk yang diinginkan. Tentu pembeli menginginkan produk yang sampai adalah barang yang bagus dan berkualitas sesuai deskripsi yang ada di *e-Commerce* sehingga pembeli bisa mendapatkan gambaran dengan barang yang ingin mereka beli. Pembeli dapat melakukan penilaian terhadap produk yang telah dibeli setelah barang tersebut sampai dan sesuai dengan rasa kepuasan terhadap produk tersebut. Namun tidak sedikit dari pengguna *e-Commerce* yang salah mengkategorikan komentar yang diberikan.

Penilaian merupakan hal yang sangat penting dalam suatu aplikasi, khususnya pada aplikasi jual-beli. Kualitas produk yang ada pada *e-Commerce* dapat dilihat melalui penilaian yang telah diberikan. [3] Penilaian dapat berupa komentar dan rating yang diberikan oleh pembeli kepada penjual yang mempengaruhi penilaian produk pada platform *e-Commerce*. Hal ini memuat kualitas barang, keramahan penjual, kecepatan pengiriman, dan pengemasan yang baik. Pembeli memberikan rating berdasarkan kepuasan dari semua segi pelayanan yang diberikan, [12] rating produk memiliki skala 1 hingga 5 bintang. Rating bintang 5 dapat diartikan pembeli merasa sangat puas terhadap semua pelayanan dan barang dari toko tersebut, bintang 4 diartikan pembeli mendapatkan kepuasan, namun ada suatu pelayanan yang menurut pembeli kurang memuaskan, dan semakin kecil bintang yang diberikan, maka pelayanan serta barang dari toko tersebut sangat tidak memuaskan.

Pengguna terkadang memberikan komentar tidak bagus dan memberikan bintang 5 terhadap barang yang dibeli sedangkan pengguna lain akan melihat produk tersebut berdasarkan ratingnya saja yang bagus. Dari kasus tersebut dapat dibuktikan bahwa sistem penilaian yang diterapkan kurang efisien. Dari uraian tersebut maka diperlukan proses klasifikasi dan teknik *pre-processing* untuk data *review* tersebut [1]. [2] Pada penelitian ini akan memakai model algoritma *Naive Bayes* dan *Logistic Regression* untuk mencari tingkat akurasi tertinggi dari klasifikasi *review* yang sudah dilakukan.

1.2 Tujuan

Tujuan yang akan dicapai dari penelitian ini berupa:

1. Menganalisis komentar pada *e-Commerce* berdasarkan kualitas barang dan pelayanan yang diberikan oleh toko bahwa komentar tersebut telah dikategorikan dengan benar ataupun tidak.
2. Mencari model algoritma terbaik dari *Naive Bayes* dibandingkan *Logistic Regression* berdasarkan tingkat akurasi tertinggi.
3. Mengklasifikasi komentar atau *review* yang diberikan pembeli.

1.3 Manfaat

Manfaat yang dapat diperoleh dari hasil penelitian ini berupa:

1. Pengguna dapat menilai suatu produk bukan hanya berdasarkan rating dalam bentuk bintang tetapi juga dengan berdasarkan kategori komentar.
2. Pengguna mendapatkan gambaran produk yang ingin dibeli melalui rating dan kategori komentar yang ada.

1.4 Batasan Masalah

Pada penelitian ini terdapat batasan masalah berupa:

1. Data yang dipakai merupakan *review* pada toko Berrybenka dari kategori bintang 5.
2. Jumlah dataset yang digunakan adalah 150 *record*, 200 *record* dan 250 *record*.
3. Aplikasi *e-Commerce* yang dianalisis adalah Shopee.

BAB 2 METODE PENELITIAN

2.1 Klasifikasi *Naive Bayes*

Naive Bayes merupakan salah satu metode pada tahap klasifikasi yang menggunakan probabilitas dan statistika. Algoritma *naive bayes* dapat memprediksi peluang di masa yang akan datang sehingga diketahui sebagai Teorema Bayes. [10] Keuntungan dari pendekatan *naive bayes* yaitu ketika diklasifikasi akan menghasilkan nilai error yang lebih kecil ketika dataset berjumlah besar. Hanya varians dari suatu variabel pada kelas yang dibutuhkan untuk menetapkan klasifikasi sehingga tidak memerlukan keseluruhan dari matriks *kovarians*. Formulasi *naive bayes* untuk klasifikasi adalah sebagai berikut:

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \qquad (1)$$

Keterangan :
X = *tuple* atau objek data, H = hipotesis bahwa *tuple* X berada di kelas C
P(H|X) = probabilitas posterior bahwa hipotesis H benar untuk *tuple* X
P(X|H) = probabilitas posterior X dengan syarat H
P(H) = probabilitas prior hipotesis H benar untuk setiap tuple
P(X) = probabilitas prior dari *tuple* X

2.2 Klasifikasi *Logistic Regression*

Logistic Regression adalah salah satu algoritma klasifikasi untuk menentukan hubungan antara fitur diskrit atau kontinu dengan hasil yang berupa output probabilitas diskrit. Metode ini secara umum digunakan untuk regresi binomial. Analisis regresi menggunakan variabel prediktor baik numerik maupun kategori.

$$\ln\left(\frac{p}{1-p}\right) = B_0 + B_1 X \qquad (2)$$

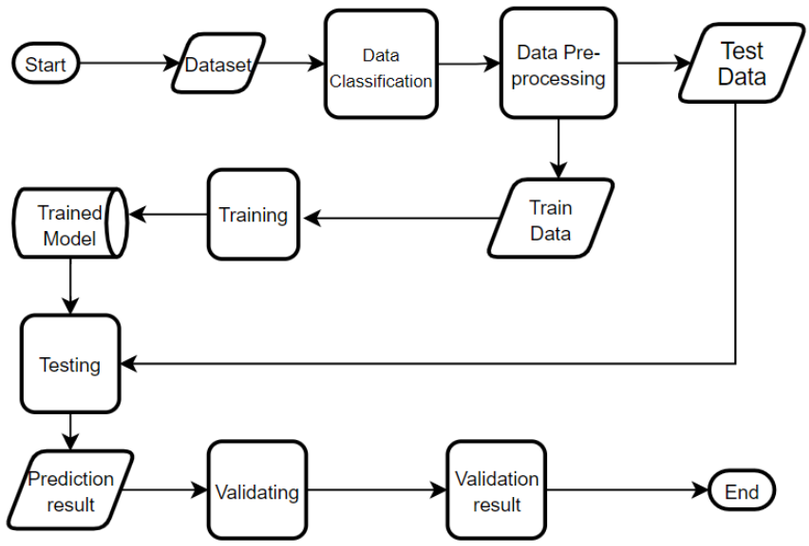
Keterangan :
Ln = logaritma natural
p = probabilitas logistik
B0 = konstanta, B1 = koefisien masing-masing variabel

$$p = \frac{e^{(B0+B1X)}}{\left(1 + e^{(B0+B1X)}\right)} \qquad (3)$$

Keterangan :
Peluang responden dapat dihitung dengan persamaan di atas.
e = fungsi eksponen

2.3 Perancangan Model

Model yang digunakan dalam penelitian ini adalah dengan metode algoritma *Naive Bayes* dan algoritma *Logistic Regression.*



Gambar 2.3. Proses Analisis Sentimen

Mengacu pada gambar 2.3 terdapat tahapan yang dilakukan di penelitian ini. Pada tahap pertama dilakukan pengumpulan data komentar salah satu produk di toko yang ada di *e-Commerce*. Komentar yang diambil yaitu pada bintang 5 suatu produk yang terdiri dari 150, 200, dan 250 komentar diambil secara acak. Kemudian data komentar diklasifikasikan ke dalam komentar puas dan tidak puas. Lalu dilakukan *pre-processing* dalam pengolahan data untuk mempermudah metode yang digunakan agar dapat berjalan dengan baik. Data kemudian dibagi menjadi data latih dan data uji dengan perbandingan 80% dan 20%, lalu dilakukan klasifikasi berdasarkan metode *naive bayes* dan *logistic regression*. Setelah itu dilakukan validasi dari hasil yang diperoleh.

2.4 Pre processing

Pada tahap *pre-processing*, data awal akan diolah untuk menjadi lebih terstruktur dan lebih bersih. Hal ini diperlukan agar dapat meningkatkan dan mempercepat proses klasifikasi. Tahap *pre-processing* yang digunakan pada analisis ini berupa :

1. *Case Folding*
   Proses ini memiliki tujuan untuk membersihkan kata yang tidak penting dan merapikan *string* yang ada pada *dataframe*, seperti mengubah semua kata yang ada pada *dataset* dalam bentuk huruf kecil.
2. *Tokenizing*
   Proses pemisahan kalimat menjadi kata-kata yang bisa disebut dengan *token* yang akan dianalisa. Namun metode *tokenizing* memiliki kelemahan karena dapat memisahkan tanda baca, maka dari itu sebaiknya teks harus melewati proses *case folding* agar lebih efisien.
3. *Filtering* atau S*topword Removal*
   Tahap ini merupakan proses pengambilan kata-kata penting yang bersifat informatif dari data yang telah melalui proses *tokenizing*. Kata-kata yang memiliki informasi yang rendah akan dihapus.

Dengan menggunakan data yang didapatkan dari shopee berupa komentar sebanyak 150 *record* , 200 *record* dan 250 *record* dimana sebanyak 80% data menjadi data latih dan 20% data menjadi data uji. Contoh komentar yang digunakan seperti pada gambar 2.4

Username	Review	Kelayakan
n*****i	Jam tangannya bagus bangeeett, cantik cuman gaada label berrybenkanya dan pengiriman agak lama, mungkin krn pandemi juga. Tp overall sukakk	Tidak
t*****n	Jamnya bagus banget aku suka model jam tangan yang kayak gini! ♡	Ya
silka.spou	Terloooovvvveee pokoknya bgaus banget,packing aman..mantap pokoknya	Ya

Gambar 2.4 *Dataset* Sebelum di *Pre-Processing*

Setelah melewati proses *pre-processing,* komentar akan berubah menjadi *dataset* baru atau data bersih nya. Contoh komentar yang didapatkan seperti pada gambar 2.5

Username	Review	Kelayakan
n*****i	jam tangannya bagus bangeeett cantik cuman gaada label berrybenkanya pengiriman krn pandemi tp overall sukakk	Tidak
t*****n	jamnya bagus banget suka model jam tangan kayak gini	Ya
silka.spou	terloooovvvveee pokoknya bgaus banget packing aman mantap pokoknya	Ya

Gambar 2.5 *Dataset* Sesudah di *Pre-Processing*

2.5 Phase Training

Proses *training* pada gambar 3.1 menggunakan data train sebagai *input* sebanyak 80%. Pelatihan dilakukan dengan dua metode yang berbeda, *naive bayes classifier* dan *logistic regression*. Kedua metode ini dijalankan menggunakan *input* data latih dengan label yang ada yaitu [0,1]. *Output* yang dikeluarkan dari proses *training* ini akan divalidasi untuk proses lebih lanjut.

2.6 Phase Testing

Proses pengujian pada gambar 3.1 menggunakan data uji dari *dataset* sebagai *input* sebanyak 20%. Pada tahap pengujian menggunakan algoritma *naive bayes* dan *logistic regression* dimana datanya telah terlatih sebelumnya. Hasil *testing* kemudian akan di evaluasi dengan metode *cross-validation* pada *naive bayes* dan metode *linear regression model* pada *logistic regression*.

2.7 Evaluasi Performansi

Setelah melewati proses pengujian, nilai akurasi pun berhasil didapatkan. Maka dari itu tingkat akurasi dapat diuji terhadap kinerja dari hasil pengujian algoritma *naive bayes* dengan bantuan *cross-validation* dan *logistic regression* dengan bantuan *linear regression model*.

BAB 3 HASIL DAN ANALISIS

3.1 Pengujian Klasifikasi Data

Dataset yang didapat akan diklasifikasikan untuk memisahkan komentar puas dan komentar yang terdapat kata-kata membandingkan yang mengarah ke komentar kontra terhadap suatu barang yang direview. Dataset akan di filter dengan kata-kata yang bersifat membandingkan yaitu “tapi”, “cuma”, “sayang”, “minus”, “kurang”, dll. Contoh komentar sebelum diklasifikasikan adalah sebagai berikut.

username	Review
n*****i	Jam tangannya bagus bangeeett, cantik cuman gaada label berrybenkanya dan pengiriman agak lama, mungkin krn pandemi juga. Tp overall sukakk
t*****n	Jamnya bagus banget aku suka model jam tangan yang kayak gini! ❤️
silka.spoutlet	Terloooovvvveee pokoknya bgaus banget,packing aman..mantap pokoknya

Gambar 3.1 *Dataset* Sebelum di Klasifikasi

Setelah dilakukan proses klasifikasi maka diperoleh dataset baru yang berisi kolom kelayakan untuk setiap komentar yang diklasifikasi. Contoh komentar yang sudah diklasifikasi ke dalam komentar puas dan komentar membandingkan yang mengacu ke komentar kontra terhadap suatu barang yang direview, dapat dilihat pada gambar dibawah ini.

Username	Review	Kelayakan
n*****i	Jam tangannya bagus bangeeett, cantik cuman gaada label berrybenkanya dan pengiriman agak lama, mungkin krn pandemi juga. Tp overall sukakk	Tidak
t*****n	Jamnya bagus banget aku suka model jam tangan yang kayak gini! ♡	Ya
silka.spou	Terloooovvvveee pokoknya bgaus banget,packing aman..mantap pokoknya	Ya

Gambar 3.2 *Dataset* Sesudah di Klasifikasi

3.2 Hasil Klasifikasi Menggunakan *Naive Bayes*

Pada tahap klasifikasi, Data akan diolah dengan menggunakan algoritma *Naive bayes* dengan bantuan *cross-validation* untuk mencari nilai keakuratan data. Hasil yang didapatkan dapat dilihat pada gambar 3.3

150 Data

	precision	recall	f1-score	support
0	0.75	0.75	0.75	8
1	0.91	0.91	0.91	22
accuracy			0.87	30
macro avg	0.83	0.83	0.83	30
weighted avg	0.87	0.87	0.87	30

200 Data

	precision	recall	f1-score	support
0	0.92	0.80	0.86	15
1	0.89	0.96	0.92	25
accuracy			0.90	40
macro avg	0.91	0.88	0.89	40
weighted avg	0.90	0.90	0.90	40

250 Data

	precision	recall	f1-score	support
0	0.75	0.75	0.75	4
1	0.94	0.94	0.94	16
accuracy			0.90	20
macro avg	0.84	0.84	0.84	20
weighted avg	0.90	0.90	0.90	20

Gambar 3.3 Akurasi *Naive Bayes*

Berdasarkan gambar 3.3 terlihat bahwa hasil pengolahan 150 data, 200 data, dan 250 dataset menunjukan tingkat akurasi yang dihasilkan sebesar 87% untuk 150 data, 90% untuk 200 data, 90% untuk 250 data. Hal ini menunjukan persentase keakuratan *naive bayes* sangat baik dan kecil kemungkinan *error* jika data semakin banyak.

3.3 Hasil Klasifikasi Menggunakan *Logistic Regression*

Pada tahap klasifikasi, Data akan diolah dengan menggunakan algoritma *logistic regression* dengan menggunakan bantuan *linear regression model* untuk mencari nilai keakuratan data. Hasil yang didapatkan dapat dilihat pada gambar 3.4

150 Data

	precision	recall	f1-score	support
0	0.75	0.38	0.50	8
1	0.81	0.95	0.88	22
accuracy			0.80	30
macro avg	0.78	0.66	0.69	30
weighted avg	0.79	0.80	0.78	30

200 Data

	precision	recall	f1-score	support
0	1.00	0.53	0.70	15
1	0.78	1.00	0.88	25
accuracy			0.82	40
macro avg	0.89	0.77	0.79	40
weighted avg	0.86	0.82	0.81	40

250 Data

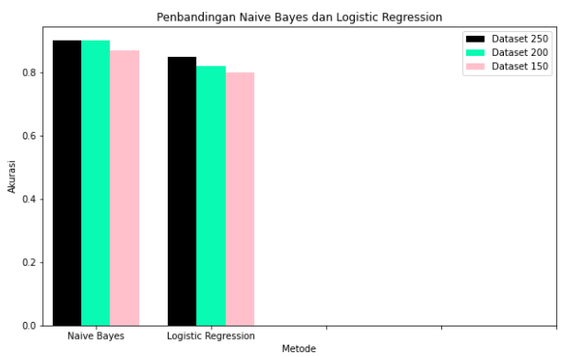
	precision	recall	f1-score	support
0	1.00	0.25	0.40	4
1	0.84	1.00	0.91	16
accuracy			0.85	20
macro avg	0.92	0.62	0.66	20
weighted avg	0.87	0.85	0.81	20

Gambar 3.4 Akurasi *Logistic Regression*

Berdasarkan gambar 3.4 terlihat bahwa hasil pengolahan 150 data, 200 data, dan 250 dataset menunjukan tingkat akurasi yang dihasilkan sebesar 80% untuk 150 data, 82% untuk 200 data, 85% untuk 250 data. Hal ini menunjukan persentase keakuratan *logistic regression* itu tinggi dan memiliki kualitas yang baik apabila mengeksekusi data yang lebih banyak.

3.4 Perbandingan *Naive Bayes* dan *Logistic Regression*

Setelah mencari nilai akurasi dari kedua algoritma, maka tingkat pengujian model kedua algoritma dapat dilihat. Berikut adalah grafik nilai akurasi dari kedua algoritma dapat dilihat pada gambar 3.5



Gambar 3.5 Grafik Perbandingan Akurasi

Dapat dilihat dari gambar 3.5 grafik diatas bahwa kedua algoritma memiliki nilai keakuratan yang tinggi. Algoritma *naive bayes* memiliki akurasi yang lebih tinggi dibandingkan dengan algoritma *logistic regression* karena mencapai nilai 90%, sedangkan algoritma *logistic regression* mencapai nilai 85%. Maka dari itu dapat dibuktikan bahwa algoritma *naive bayes* memiliki waktu komputasi yang lebih cepat dan lebih efisien dibandingkan *logistic regression.* Metode *naive bayes* digunakan karena sederhana dan menghasilkan akurasi yang tinggi sedangkan metode *logistic regression* digunakan karena variabelnya independen dan tidak memiliki keterbatasan(Kuncoro, 2001).

BAB 4 KESIMPULAN

Berdasarkan hasil pengujian model menggunakan algoritma *Naive Bayes* dan *Logistic Regression* pada penelitian yang telah dilakukan, terbukti bahwa Algoritma *Logistic Regression* menghasilkan tingkat akurasi hingga 85% dan algoritma *Naive Bayes* menghasilkan tingkat akurasi hingga 90%. Dari penelitian tersebut metode *Naive Bayes* dan *Logistic Regression* sudah baik atau sudah akurat.

Penelitian ini berhasil membuktikan bahwa model algoritma terbaik pada permasalahan ini dengan menggunakan algoritma *naive bayes classifier*. Hasil penelitian ini dianggap tidak maksimal karena memiliki kemungkinan bahwa terdapat kalimat kepuasan yang terhapus. Berdasarkan *hyperparameter*, hal yang mempengaruhi output dalam penelitian ini yaitu review dan nilai kelayakan.

Meskipun telah tercapainya model algoritma terbaik yaitu *naive bayes classifier* yang menghasilkan nilai akurasi yang tinggi, namun terdapat beberapa hal yang perlu diperbaiki. Terdapat saran-saran yang diusulkan untuk meningkatkan mengoptimalkan penelitian seperti menggunakan *stemming* dan *TF-IDF* untuk mendapatkan hasil yang lebih efisien. Penelitian ini diharapkan agar pengguna dapat menyesuaikan *rating* dengan komentar yang akan dibuat sehingga pengguna lain tidak salah menilai saat ingin membeli produk tersebut.

REFERENSI

[1] Mujilahwati, Siti. 2016 “Pre-Processing Text Mining Pada Data Twitter”, Seminar Nasional Teknologi Infomasi dan Komunikasi 2016 (SENTIKA 2016), Maret 2016, 2089-9815.

[2] Indrayuni, Elly. 2019 “Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes”, Jurnal Khatulistiwa Informatika, VOL. VII, NO.1, Juni 2019, 2339-192.

[3] Agustina, Lidya., Fayardi, Alifia., Irwansyah. (2018) “Online Review: Indikator Penilaian Kredibilitas Online dalam Platform *e-Commerce*”, Jurnal Ilmu Komunikasi, Vol.15, No.2, Desember 2016, 141-154.

[4] vincentmichael089. (2019, May 8). *Machine Learning: Mengenal Logistic Regression - vincentmichael089 - Medium*. Medium; Medium. https://vincentmichael089.medium.com/machine-learning-2-logistic-regression-96b3d4e7b603.

[5] Budi, Setyo. 2017. “Text Mining Untuk Analisis Sentimen Review Film Menggunakan Algotima K-Means”, Techno.COM, Vol.16, No.1, Februari 2017.

[6] Kontributor dari proyek Wikimedia. (2011, January 3). *Regresi logistik*. Wikipedia.org; Wikimedia Foundation, Inc. https://id.wikipedia.org/wiki/Regresi_logistik.

[7] informatikalogi. (2017, April 8). *Algoritma Naive Bayes | INFORMATIKALOGI*. Informatikalogi. https://informatikalogi.com/algoritma-naive-bayes/.

[8] Kuncahyo Setyo Nugroho. (2019, June 18). *Dasar Text Preprocessing dengan Python - Kuncahyo Setyo Nugroho - Medium*. Medium; Medium. https://medium.com/@ksnugroho/dasar-text-preprocessing-dengan-python-a4fa52608ffe

[9] Lidya, Kartika., Sitompul, Opim., Efendi, Syahril. 2015. “Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (SVM) dan K-Nearest Neighbour (K-NN)”, Seminar Nasional Teknologi Informasi dan Komunikasi 2015 (SENTIKA 2015), Maret 2015, 2089-9815.

[10] Putri, Eko Riyan, Suparti dan Rita Rahmawati. 2014 “Perbandingan Metode Klasifikasi Naive Bayes dan K-Nearest Neighbour Pada Analisis Data Status Kerja di Kabupaten Demak Tahun 2012”, Jurnal Gaussian, VOL. III, NO.4, 2014, 2339-2541

[11] https://idcloudhost.com/author/marketing. (2020, February 27). *Pengertian E-Commerce dan Contohnya, Komponen, Jenis, dan Manfaat E-Commerce | IDCloudHost*. IDCloudHost; IDCloudHost. https://idcloudhost.com/pengertian-e-commerce-dan-contohnya-komponen-jenis-dan-manfaat-e-commerce/

[12] *Seller Education Hub*. (2021). Shopee.co.id. https://seller.shopee.co.id/edu/article/467