PRISMA, Prosiding Seminar Nasional Matematika
https://proceeding.unnes.ac.id/prisma
ISSN 2613-9189

Analisis Sentimen Pada Ulasan Aplikasi Tokopedia Menggunakan Klasifikasi *Naïve Bayes*

Nanda Aurelia Salsabila[a,*], Umi Sa'adah [b], Fatkhurohman Fauzi [c]

[a,b] *Universitas Negeri Semarang, Semarang 50229,Indonesia*
[c] *Universitas Muhammadiyah Semarang, Semarang 50272,Indonesia*

** Alamat Surel: nanda27yom@students.unnes.ac.id*

Abstrak

Aplikasi *e-commerce* telah menjadi bagian integral dari kehidupan Masyarakat, membuat belanja online lebih simple dan nyaman. Tokopedia salah satu *platform e-commerce* terkemuka di Indonesia, menghadirkan aplikasi *mobile* yang memungkinkan pengguna untuk membeli dan menjual produk serta berinteraksi dengan komunitas. Pada tahun 2023 Tokopedia menduduki kuartal pertama dengan 117 juta pengunjung. Oleh karena itu penting untuk memahami pandangan dan sentimen pengguna terhadap aplikasi Tokopedia dalam meningkatkan pengalaman pengguna. Penelitian ini berfokus pada ulasan pengguna aplikasi Tokopedia dengan menggunakan metode *Naïve Bayes* dengan menggunakan *software* R studio. Data ulasan diperoleh dari *google play*. Data tersebut dilakukan *scraping* data untuk mengumpulkan maupun penyortiran ulasan agar mempermudah dalam memperoleh informasi dari data yang berjumlah sangat banyak selanjutnya data disimpan dalam bentuk dokumen dengan format csv sebanyak 2000 ulasan yang diurutkan berdasarkan the most relevant. Kemudian data tersebut dianalisis isi dan karakter (*Exploring and preparing*) selanjutnya mengklasifikasi sentiment dengan menggunakan *Naïve Bayes*. Data *review* dengan score 3 bintang (netral) dihilangkan sehingga tersisa 1819 data ulasan dengan score 1 dan 2 bervalue "Negatif", data score 4 dan 5 bervalue "Positif". Sehingga menghasilkan nilai akurasi dari klasifikasi Naïve Bayes sebesar 82,97 % dengan jumlah ulasan positif sebesar 338 dan ulasan negatif sebesar 1481.

Kata kunci:

Naïve Bayes, Teks Mining, Tokopedia

1. Pendahuluan

Setiap tahunnya teknologi semakin berkembang salah satunya perkembangan teknologi di Indonesia. Berkembangnya teknologi menjadikan banyak perubahan di seluruh aspek kehidupan. Diantaranya adalah aspek berkomunikasi, berbelanja, dan juga berdagang. perubahan aspek tersebut membuat banyak perusahaan mengubah proses jual beli yang dulunya dilakukan secara langsung menjadi berbasis digital atau disebut dengan *e-commerce*. Sebuah survei yang dilakukan oleh Asosiasi Pengusaha Jasa Internet (APJII) pada tahun 2023 menemukan bahwa jumlah pengguna internet di Indonesia mencapai 213 juta pada Januari 2023, naik 5,44% dari 202 juta pada tahun sebelumnya. Ini menunjukkan bahwa telepon genggam dan akses internet digunakan oleh sebagian besar penduduk Indonesia sebesar 98,3% (APPJII, 2023) . Salah satunya *e-commerce* saat ini yaitu Tokopedia.

Tokopedia merupakan salah satu *platform marketplace* terbesar di Indonesia. Tokopedia berhasil menjadi salah satu perusahaan internet indonesia dengan pertumbuhan yang pesat. Tokopedia adalah perusahaan internet yang memungkinkan setiap individu dan pemilik bisnis di Indonesia untuk mengembangkan dan mengelola bisnis *online* mereka secara mudah dan gratis, sekaligus memungkinkan pengalaman berbelanja *online* yang lebih aman dan nyaman. Tokopedia percaya bahwa *marketplace* adalah bisnis model paling indah di dunia, karena kesuksesan sebuah *marketplace* hanya dapat diraih dengan membuat orang lain menjadi lebih sukses. Aplikasi Tokopedia terdapat dalam situs *Google Play* merupakan layanan digital yang lahir dari *Google*, didalamnya mencakup toko maupun media yang digunakan dalam pemasaran produk-produk seperti aplikasi, *game*, musik, dan lain sebagainya. Dalam *google play* sendiri terdapat fitur ulasan (*review*) dari pengguna. Hal tersebut dapat dimanfaatkan oleh Tokopedia sebagai tolak ukur terhadap efektif dan efisien dalam menemukan informasi terhadap suatu

To cite this article:
Salsabila, N., Saadah, U & Fauzi, F. (2024). Analisis Sentimen Pada Ulasan Aplikasi Tokopedia Menggunakan Klasifikasi Naïve Bayes. *PRISMA, Prosiding Seminar Nasional Matematika 7*, 44-51

produk. Ulasan biasanya berisi kritik positif dan negatif yang ditulis oleh pengguna secara tidak langsung baik jumlah kecil atau besar, yang berpotensi berdampak pada calon pelanggan. Dibutuhkan metode untuk menyortir data ulasan secara cepat dan mengklasifikasinya dalam kategori positif dan negatif karena menyortirnya secara manual memakan waktu yang lama.

Salah satu cabang pemrosesan bahasa alami (NLP) adalah analisis sentimen, yang bertanggung jawab untuk membangun sistem yang dapat mengenali dan mengekstraksi pendapat dari teks. Saat ini, sebagian besar informasi teks dapat ditemukan di internet dalam bentuk forum, blog, media sosial, dan situs web yang berisi *review*. Analisis Sentimen dapat digunakan untuk mengubah informasi yang tidak terstruktur menjadi data yang lebih terstruktur (Darwis et al., n.d.)

Naïve Bayes Classifier adalah salah satu metode pembelajaran text mining untuk analisis sentimen yang dianggap lebih baik daripada metode klasifikasi lainnya dalam hal akurasi dan komputasi (Apriani et al., 2019). Algoritma *Naïve Bayes Classifier* dapat digunakan untuk memprediksi suatu nilai dari variabel dalam data testing (Berliana et al., 2018). Keuntungan penggunaan metode ini adalah bahwa itu hanya membutuhkan jumlah data pelatihan yang kecil untuk menentukan perkiraan parameter yang dibutuhkan dalam proses pengklasifikasian. Karena yang dianggap sebagai variabel independen, hanya varians dari suatu variabel dalam kelas yang diperlukan untuk untuk menentukan klasifikasi, bukan varians dari matriks kovarians secara keseluruhan (Mochammad Haldi Widianto, n.d.).

Penelitian ini menawarkan solusi untuk menghitung persentase komentar dan respons pengguna aplikasi berdasarkan ulasan yang diberikan pengguna atau pelanggan pada aplikasi Tokopedia di *Google Play Store*. Komentar dan respons ini akan membantu perusahaan meningkatkan layanan, kualitas produk dan juga perusahaan. Penelitian ini bermanfaat bagi manajemen Tokopedia agar tetap konsisten dalam pelayanan pada pelanggan untuk mempertahankan tingkat kepercayaan pelanggan, yang pada gilirannya dapat menghasilkan peningkatan keuntungan sesuai dengan tujuan yang ditetapkan.

2. Metode

2.1 Sumber Data

Data ulasan aplikasi Tokopedia diambil dari *google play store* dengan alamat URL: https://play.google.com/store/apps/details?id=com.tokopedia.tkpd&hl=en . Data ulasan dari pengguna aplikasi Tokopedia diambil dalam bentuk dokumen dengan format csv sebanyak 2000 ulasan yang diurutkan berdasarkan the most relevant. Data berisi ID pengguna, *username* dari pengguna, tanggal ulasan tersebut diunggah, *content* atau usulan yang diberikan pengguna, dan *score* yang merupakan nilai bintang diberikan oleh pengguna.

2.2 Pelabelan Data

Data yang diperoleh kemudian diproses dengan pemberian label pada setiap data atau ulasan. Label yang digunakan adalah positif dan negatif yang mengacu pada *score* yang diberikan oleh para pengguna. Nilai *score* yang kurang dari 3 (1 dan 2) akan diberi label negatif dan nilai *score* yang lebih dari 3 (4 dan 5) akan diberi label positif. Nilai *score* yang bernilai 3 dianggap netral sehingga tidak dijadikan sebagai bahan pertimbangan. Ulasan dikatakan positif apabila ulasan berisi tentang pengalaman pemakaian aplikasi Tokopedia yang menyenangkan pengguna, mempermudah transaksi, dan memuaskan pengguna. Ulasan dikatakan negatif apabila berisikan keluhan dan kekecewaan para pengguna dan juga problem pada aplikasi.

2.3 *Naive Bayes*

Metode yang akan digunakan dalam pengeksekusian data diatas adalah dengan menggunakan metode yang bernama *Naïve Bayes*. Klasifikasi *Naïve Bayes* adalah metode klasifikasi data berdasarkan probabilitas yang mungkin terjadi di masa yang akan datang. Metode ini merupakan pendekatan statistik untuk melakukan inferensi induksi pada persoalan klasifikasi (Larone,2001). Dengan penggunaan metode tersebut, perusahaan dapat memprediksi sentiment ulasan yang akan ada di masa depan. Adapun Teorema *Naive Bayes* sebagai berikut:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Keterangan:

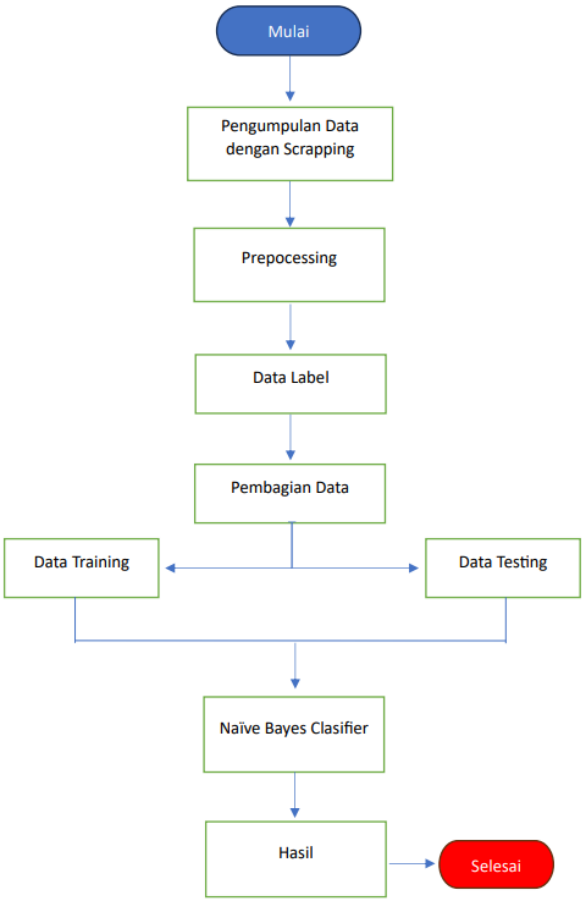
A, B = event

$P(A|B)$ = probability of A given B is true

$P(B|A)$ = probability of B given A is true

$P(A), P(B)$ = the independent probabilities of A and B

Naïve Bayes Classifer merupakan salah satu algoritma yang sederhana namun memiliki kemampuan dan akurasi yang tinggi dan termasuk dalam metode *machine learning* (Gumilang, 2018). Metode pengklasifikasian yang menggunakan probabilitas dan statistik diusulkan oleh ilmuwan Inggris Thomas Bayes. Teorema Bayes adalah prediksi peluang di masa depan berdasarkan pengalaman masa lalu. Salah satu ciri utama klasifikator Dr. Naive Bayes ini adalah keyakinan yang sangat kuat (naif) bahwa setiap kondisi atau kejadian independent (Rohith Gandhi, 2018). Keuntungan penggunaan metode ini adalah bahwa itu hanya membutuhkan jumlah data pelatihan yang kecil untuk menentukan perkiraan parameter yang dibutuhkan dalam proses pengklasifikasian. Karena yang dianggap sebagai variabel independen, hanya varians dari suatu variabel dalam kelas yang diperlukan untuk untuk menentukan klasifikasi, bukan varians dari matriks kovarians secara keseluruhan (Mochammad Haldi Widianto, n.d.).



Gambar 1. Alur Penelitian

3. Pembahasan

3.1 Pengambilan Data

Data ulasan aplikasi Tokopedia diambil dari *google play store* dengan alamat URL: https://play.google.com/store/apps/details?id=com.tokopedia.tkpd&hl=en. (Tokopedia – Android Apps on Google Play, 2014). Data tersebut dilakukan *scraping* data. *Scraping* data adalah proses mengumpulkan maupun penyortiran ulasan agar mempermudah dalam memperoleh informasi dari data yang berjumlah sangat banyak. Data ulasan dari penggunaan Tokopedia diambil dalam bentuk dokumen dengan format csv sebanyak 2000 ulasan yang diurutkan berdasarkan *the most relevant*. Data yang diperoleh ini diambil atribut nama (*username*), *at*, *content* (ulasan), dan *score*. Kemudian data ulasan yang memiliki score 3 bintang dihilangkan untuk menghindari penilaian bias (netral), sehingga tersisa 1819 data ulasan. Data ulasan dengan score 1 dan 2 diberikan label "NEGATIF", sedangkan data dengan score 4 dan 5 diberikan label "POSITIF". Selanjutnya data disimpan dalam format csv, pada penelitian ini diberi nama "data_tokped1.csv".

3.2 Exploring dan Preparing Data

Exploring dan preparing data dilakukan untuk menganalisis data awal agar dapat memahami isi dan karakteristik data. Setelah mengimport data "data_tokped1.csv" dengan perintah read.csv, selanjutnya dihitung jumlah ulasan yang bernilai "POSITIF" dan "NEGATIF".

```
#Mengecek jumlah data positif dan negatif
table(data_tokped1$label)
```

Diperoleh hasil run terdapat ulasan bernilai "POSITIF sebanyak 338 dan ulasan bernilai "NEGATIF" sebanyak 1481. Kemudian dilanjutkan langkah membuat corpus dari kalimat ulasan dan membuat *document-term spare matrix* dari *corpus* tersebut dengan perintah berikut.

```
# Create a corpus from the sentences
data_corpus <- VCorpus(VectorSource(data_tokped1$content))

# create a document-term sparse matrix directly from the corpus
data_dtm <- DocumentTermMatrix(data_corpus, control = list(
  tolower = TRUE,
  removeNumbers = TRUE,
  stopwords = TRUE,
  removePunctuation = TRUE,
  stemming = TRUE
))
```

Langkah selanjutnya yaitu memisahkan *data training* dan *data testing*, pembagian *data training* dan *testing* sebesar 80% untuk *data training* dan sisanya *data testing*. Dimana data ke-1 sampai ke-1456 merupakan *data training* dan data ke -1457 sampai ke-1819 merupakan *data testing* yang digunakan untuk menguji performa model yang sudah dilatih.

```
# creating training and test datasets
data_dtm_train <- data_dtm[1:1456, ]
data_dtm_test  <- data_dtm[1457:1819, ]
```

Lalu diperiksa perbandingan dari data bernilai "POSITIF" dan "NEGATIF" yang diperoleh berdasarkan pada *data testing* dan *data training* yang telah dipisahkan.

```
# check that the proportion of spam is similar
prop.table(table(data_train_labels))
prop.table(table(data_test_labels))
```

Dari hasil run perintah di atas diperoleh perbandingan ulasan bernilai "POSITIF" dan "NEGATIF" pada *data training* adalah 0,206044 banding 0,793956, sedangkan pada *testing* adalah 0,1046832 banding 0,8953168. Perbandingan yang diperoleh tidak sama, sehingga perlu menggunakan *sampling*.

```
rm(data_dtm_train)
rm(data_dtm_test)
rm(data_train_labels)
rm(data_test_labels)

# Create random samples
set.seed(123)
train_index <- sample(1819, 1456)

data_train <- data_tokped1[train_index, ]
data_test  <- data_tokped1[-train_index, ]

# check the proportion of class variable
prop.table(table(data_train$label))
prop.table(table(data_test$label))
```

Dari *sampling* diperoleh perbandingan ulasan bernilai "POSITIF" dan "NEGATIF" pada *data training* menjadi 0,1799451 dibanding 0,8200549, sedangkan pada *testing* menjadi 0,2093664 banding 0,7906336 Selanjutnya dibuat *corpus* dari *data training* dan data tersebut. Untuk melihat kata yang sering muncul pada ulasan bernilai "POSITIF" dan "NEGATIF" divisualisasikan dengan *word cloud*.

```
#WORDCLOUD VISUALIZATION
# subset the training data into spam and ham groups
positive <- subset(data_train, label == "POSITIF")
negative  <- subset(data_train, label == "NEGATIF")

wordcloud(positive$content, max.words = 75, scale = c(3, 0.5))
wordcloud(negative$content, max.words = 75, scale = c(3, 0.5))
```

Berikut adalah hasil *word cloud*:

Gambar 2. (a) *Word Cloud* Sentimen Positif; (b) *Word Cloud* Sentimen Negatif

Langkah selanjutnya yaitu membuat *document-term spare matrix* (DTM) dari *data training* dan *data testing*.

```
# create a document-term sparse matrix directly for train and test
train_dtm <- DocumentTermMatrix(train_corpus, control = list(
  tolower = TRUE,
  removeNumbers = TRUE,
  stopwords = TRUE,
  removePunctuation = TRUE,
  stemming = TRUE
))

test_dtm <- DocumentTermMatrix(test_corpus, control = list(
  tolower = TRUE,
  removeNumbers = TRUE,
  stopwords = TRUE,
  removePunctuation = TRUE,
  stemming = TRUE
))
```

Untuk melihat *document-term matrix* dari *data training* menggunkan perintah *train_dtm* dan diperoleh hasil sebagai berikut.

```
> train_dtm
<<DocumentTermMatrix (documents: 1456, terms: 5787)>>
Non-/sparse entries: 39681/8386191
Sparsity           : 100%
Maximal term length: 38
Weighting          : term frequency (tf)
```

Sedangkan untuk melihat *document-term matrix data testing* menggunkan perintah *test_dtm* dan diperoleh hasil seperti di bawah berikut.

```
> test_dtm
<<DocumentTermMatrix (documents: 363, terms: 2424)>>
Non-/sparse entries: 9545/870367
Sparsity           : 99%
Maximal term length: 29
Weighting          : term frequency (tf)
```

Karena jumlah data tidak terlalu besar, maka spare terms tidak dihapus. Buat fungsi untuk mengkonversi *counts* ke *factor*.

```
# create function to convert counts to a factor
convert_counts <- function(x) {
  x <- ifelse(x > 0, "Yes", "No")
}
```

Dilanjutkan dengan mengaplikasikan fungsi ini ke kolom *data training* maupun *data testing*.

```
# apply() convert_counts() to columns of train/test data
train_dtm_binary <- apply(train_dtm, MARGIN = 2, convert_counts)
test_dtm_binary  <- apply(test_dtm, MARGIN = 2, convert_counts)
```

3.3 *Training* data pada *Naive Bayes*

Dilakukan klasifikasi *Naive Bayes* menggunakan fungsi naiveBayes() pada *data tarining*.

```
#TRAINING MODEL ON THE DATA
data_classifier <- naiveBayes(as.matrix(train_dtm_binary), data_train$label)
```

3.4 Evaluasi Model

Evaluasi model dilakukan untuk melihat seberapa akurat dari hasil prediksi yang dilakukan menggunakan *data training*.

```
> data_test_pred <- predict(data_classifier, as.matrix(test_dtm_binary))
> head(data_test_pred)
[1] NEGATIF NEGATIF POSITIF NEGATIF NEGATIF POSITIF
Levels: NEGATIF POSITIF
```

Dibuat tabel untuk menunjukkan perbandingan data prediksi dan data aktual.

```
CrossTable(data_test_pred, data_test$label,
           prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
           dnn = c('predicted', 'actual'))
```

Berikut tabel yang dihasilkan

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Col Total |
|-------------------------|


Total Observations in Table:  363


             | actual
   predicted |   NEGATIF |   POSITIF | Row Total |
-------------|-----------|-----------|-----------|
     NEGATIF |       261 |        19 |       280 |
             |     0.909 |     0.250 |           |
-------------|-----------|-----------|-----------|
     POSITIF |        26 |        57 |        83 |
             |     0.091 |     0.750 |           |
-------------|-----------|-----------|-----------|
Column Total |       287 |        76 |       363 |
             |     0.791 |     0.209 |           |
-------------|-----------|-----------|-----------|
```

Kemudian hasil evaluasi divisualisasikan dalam bentuk *confusion matrix*

```
install.packages("caret")
library(caret)
conf.mat <- confusionMatrix(data_test_pred, data_test$label)
conf.mat
```

Berikut *confusion matrix* yang terbentuk

```
Confusion Matrix and Statistics

          Reference
Prediction NEGATIF POSITIF
   NEGATIF     261      19
   POSITIF      26      57

               Accuracy : 0.876
                 95% CI : (0.8377, 0.9081)
    No Information Rate : 0.7906
    P-Value [Acc > NIR] : 1.573e-05

                  Kappa : 0.6378

 Mcnemar's Test P-Value : 0.3711

            Sensitivity : 0.9094
            Specificity : 0.7500
         Pos Pred Value : 0.9321
         Neg Pred Value : 0.6867
             Prevalence : 0.7906
         Detection Rate : 0.7190
   Detection Prevalence : 0.7713
      Balanced Accuracy : 0.8297

       'Positive' Class : NEGATIF
```

Dari *confusion matrix* di atas dapat dilihat bahwa tingkat akurasi dari hasil prediksi menggunakan *Naïve Bayes* pada makalah ini sebesar 82,97% dengan positive class NEGATIF.

4. Kesimpulan

Perkembangan teknologi yang semakin maju membawa perubahan dalam berbelanja. Tokopedia salah satu aplikasi *e-commerce* yang ada di Indonesia. Untuk mengetahui bagaimana sentiment atau pendapat pengguna aplikasi ini dapat digunakan analisis ulasan dengan *Naive Bayes*. Langkah-langkah yang digunakan antara lain pengambilan data, *exploring and preparing the data*, *training a model on the data*, dan *evaluating model performance*. Pembagian data set menjadi *data training* dan *data testing* dapat mempengaruhi kinerja sistem dalam mengklasifikasi data, namun karena pada penelitian ini pembagian dilakukan dengan sistem dimana kemungkinan *data training* setiap kelas tidak seimbang sehingga mempengaruhi kinerja sistem yang dapat dilihat pada hasil akurasi pada skenario pembagian *data training* dan *data testing* dengan *presentase* yang berbeda. Jika dilihat dari hasil akurasi, pembagian dataset yang dapat menghasilkan akurasi tinggi adalah *data training* sebesar 80% dan *data training* sebesar 20%. Berdasarkan hasil analisis diperoleh hasil tingkat keakurasian dalam analisis sentimen data dari aplikasi Tokopedia di *Google Play Store* dengan menggunakan metode *Naïve Bayes Classifer* adalah sebesar 82,97% sebanyak 338 ulasan bernilai "POSITIF" dan 1481 ulasan bernilai "NEGATIF".

Daftar Pustaka

APPJII. (2023). *Survei Penetrasi & Perilaku Internet 2023*. www.survei.appjii.or.id

Apriani, R., Gustian, D., Program, S., Sistem, I., Putra, U. N., Indonesia, S., Raya, J., Kaler, C., 21, N., & Sukabumi, K. (2019). ANALISIS SENTIMEN DENGAN NAÏVE BAYES TERHADAP KOMENTAR APLIKASI TOKOPEDIA. In *Jurnal Rekayasa Teknologi Nusa Putra* (Vol. 6, Issue 1).

Berliana, G., Shaufiah, & Sa'adah, S. (2018). Klasifikasi Posting Tweet mengenai Kebijakan Pemerintah Menggunakan Naive Bayesian Classification. *E-Proceeding of Engineering*, *5*, 1562–1569.

Darwis, D., Siskawati, N., & Abidin, Z. (n.d.). *Penerapan Algoritma Naive Bayes untuk Analisis Sentimen Review Data Twitter BMKG Nasional*. *15*(1).

Gumilang, Z. A. N. (2018). IMPLEMENTASI NAÏVE BAYES CLASSIFIER DAN ASOSIASI UNTUK ANALISIS SENTIMEN DATA ULASAN APLIKASI E-COMMERCE SHOPEE PADA SITUS GOOGLE PLAY. *Tugas Akhir*.

Mochammad Haldi Widianto. (n.d.). *Algoritma Naive Bayes*. Binus University. Retrieved September 25, 2023, from https://binus.ac.id/bandung/2019/12/algoritma-naive-bayes/

Rohith Gandhi. (2018, May 5). *Pengkasifikasi Naive Bayes*. Menuju Ilmu Data. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

tokopedia – Android Apps on Google Play. (2014). Google.com. https://play.google.com/store/search?q=tokopedia&c=apps&hl=en-ID