**Just IT : Jurnal Sistem Informasi, Teknologi Informasi dan Komputer**
Vol. 15, No. 2, Januari 2025, pp. 362 – 372
P-ISSN: 2089 – 0256, e-ISSN: 2598 – 3016

# PERBANDINGAN METODE *MACHINE LEARNING* UNTUK SENTIMEN ANALISIS *REVIEW* PENJUALAN PRODUK

**Muhammad Reza[1], Ardiansyah Dores[2], Sitti Nurbaya Ambo[3], Popy Meilina[4]**
Teknik Informatika Fakultas Teknik Universitas Muhammadiyah Jakarta
popy.meilina@umj.ac.id [4]

**Abstrak**

Toko *online* atau *e-commerce* menurut Moossa Giant dan Samuel Ikate adalah operasi bisnis yang dilakukan secara dunia maya atau *online*. Pada saat pandemi di tahun 2020 sampai Mei 2023 kegiatan masyarakat dibatasi masyarakat membeli barang di toko *online* agar tidak terkena virus *corona*, akan tetapi calon konsumen membeli barang melihat ulasan pada barang yang ingin mereka tuju apakah barang yang mereka beli bagus atau pengirimannya lambat. Salah satu toko *online* yang memiliki ulasan komentar pelanggan yang sudah membeli barang sebagai petunjuk calon konsumen untuk membeli atau tidak, maka peneliti melakukan analisis sentimen terhadap ulasan konsumen membeli barang di produk elektronik dan produk pakaian, data ulasan konsumen dikumpulkan dari data elektronik sebanyak 925 data, dan data pakaian sebanyak 1575 data, setelah mengumpulkan data dilakukan *preprocessing*, pembobotan kata, pemodelan dengan *supervised learning*, yaitu *naives bayes, decision tree, k nearest neighbor*, melakukan berbagai skenario dengan pembagian data dari 10% data uji 90% data latih, 20% data uji 80% data latih, 30% data uji 70% data latih, 40% data uji 60% data latih. Hasil terbaik pengujian dengan data pakaian menggunakan *decision tree* dengan *split* data 10% data uji dan 90% data latih menghasilkan hasil akurasi 66%, *recall* 66%, dan *precision* 65%, hasil terbaik pengujian dengan data elektronik menggunakan *decision tree* dengan *split* data 40% data uji dan 60% data latih menghasilkan hasil akurasi 66%, *recall* 66%, dan *precision* 65%

**Kata Kunci:** toko *online*, analisis sentimen, *naives bayes, decision tree, k nearest neighbor.*

**Abstract**

*Online Shops or e-commerce according to Moossa Giant and Samuel Ikate are business operations carried out in cyberspace or online. During the pandemic from 2020 to May 2023, people's activities were limited to buying goods at online stores so that they would not be exposed to the corona virus. However, potential consumers buying goods see reviews on the goods they want to go to, whether the goods they are buying are good or the delivery is slow. One of the online stores that has reviews of customer comments who have purchased goods as a guide for potential consumers to buy or not, the researchers conducted a sentiment analysis of consumer reviews buying goods in electronic products and clothing products, consumer review data were collected from electronic data as many as 925 data, and 1575 clothing data, after collecting the data preprocessing, word weighting, modeling with supervised learning, namely naive Bayes, decision trees, k nearest neighbor, perform various scenarios with data sharing from 10% test data 90% training data, 20% test data 80% training data, 30% test data 70% training data, 40% test data 60% training data. The best results of testing with clothing data using a decision tree with a split data of 10% test data and 90% training data yield results of 66% accuracy, 66%*

*recall and 65% precision, the best results for testing with electronic data using a decision tree with 40% split data test data and 60% training data produce 66% accuracy, 66% recall, and 65% precision.*
**Keywords:** *online Shop, sentiment analysis, naives bayes, decision tree, k nearest neighbor*

## 1. Pendahuluan

Toko *online* atau *e-commerce* menurut Moossa Giant dan Samuel Ikate adalah operasi bisnis yang dilakukan secara dunia maya atau *online* (Salsabila et al., 2022) (Gian & Ikate, 2021). Pada saat pandemi Covid-19 fenomena belanja secara maya mulai meningkat karena kegiatan masyarakat dibatasi (Ricky et al., 2021).

Pada saat kegiatan masyarakat dibatasi salah satu toko *online*, yaitu Tokopedia menyatakan dapat berbelanja dengan cara paling aman, tidak perlu takut akan kesehatan demi mempenuhi kebutuhannya (Vega et al., 2021). Tokopedia merupakan toko *online* yang paling banyak dikunjungi oleh masyarakat Indonesia, menerima 1,2 miliar pengunjung (Apriani et al., 2019).

Penelitian yang ditulis oleh Pratiwi Arbaini, ZakariahWahab, dan MarlinahWidiyanti menyatakan bahwa calon konsumen untuk membeli barang di Tokopedia dipengaruhi oleh ulasan pelanggan (Arbaini, 2020). Konsumen mempunyai minat untuk membeli suatu produk karena adanya ulasan *customer review* (Maulana & Santy, 2021). Ulasan di toko *online* tentu ada penilaian dari konsumen yang sudah membeli barang untuk memberikan opini berupa pengalaman atau evaluasi pelayanan yang sampai ke tangan pembeli (Zhang et al., 2020).

Opini menurut Irawan Noor Kabiru Puspita dan Kencana Sari adalah penilaian konsumen terdapat 2 kondisi, yaitu opini *positive* (bagus) atau opini negative (kurang bagus) (Kabiru & Sari, 2019). Menurut Rahmat Syahputra berbagai opini dapat dikelompokkan menggunakan sentimen analisis (Syahputra et al., 2022).

Analisis sentimen atau *sentiment analysis* adalah penggkategorian data teks yang berisi opini untuk mendapatkan pemahaman sikap pelanggan (Dang et al., 2021). Analisis sentimen di *e-commerce* melakukan pengambilan data berbentuk teks dari ulasan *customer review* kemudian dilakukan pelabelan rating 1-2 diberikan label negatif, rating 3 diberikan label netral, dan rating 4-5 diberikan label positif (Demircan et al., 2021).

Analisis sentimen yang sebagian besar data dikategorikan oleh manusia, algoritma pembelajaran mesin *supervised learning* (berdasarkan label) untuk mempelajari polaritas (positif, negatif, atau netral) dari ulasan (Bharathi et al., 2022).

*Decision tree* merupakan algoritma *supervised learning* yang bekerja seperti struktur pohon di setiap *node* atau simpul mewakili dari atribut yang dilatih (Panhalkar & Doye, 2022). *Naives bayes* adalah Naïve Bayes merupakan algoritma klasifikasi probabilitas berdasarkan label data untuk memprediksi peluang masa depan dengan data sebelumnya (Watrianthos et al., 2019). *K-nearest neighbor* Merupakan algoritma klasifikasi dengan menggunakan input fitur dan *output* fitur dengan melihat dari kelas atau fitur *neighbor*(tetangga) terdekat (Cunningham & Delany, 2021).

## 2. Landasan Teori

Analisis Sentimen merupakan opini yang bersifat positif, negatif berasal dari data teks (Septiani & Sibaroni, 2019). Sentimen analisis pada dasarnya adalah melakukan klasifikasi untuk memahami sudut pandang, interaksi, dan emosi dari data teks (Ramadhan & Ramadhan, 2022).

*Text mining* adalah kegiatan menambang data *unstructured* yang datanya berbeda dengan data berbentuk tabel atau *structured*, akan tetapi datanya berbentuk teks serta didapatkan di dokument, media sosial, serta *text mining* mengekstra informasi dari data teks (Hassani et al., 2020).

*Text preprocessing* adalah menurut penelitian Firdaus dan peneltian Filcha menjelaskan pembersihan data, seperti menghilangkan tanda baca, menghapus kata ganti agar data teks menjadi kata dasar (Firdaus et al., 2022) (Filcha & Hayaty, 2019). Berikut tahap *text preprocessing*:

Labelling merupakan tahap pelabelan berdasarkan rating (Demircan et al., 2021). Pada tahap pelabelan data berdasarkan rating, menurut penelitian Elmurngi (Elmurngi & Gherbi, 2018). Terdapat pembagian 3 kategori sentimen, sebagai berikut:

1. Bintang 1-2 diberikan label negatif.
2. Bintang 3 diberikan label netral.
3. Bintang 4-5 diberikan label positif.

*Case Folding* adalah transformasi data teks yang mempunyai huruf kapital menjadi huruf kecil (Pravina et al., 2019).

*Punctuation Removal* merupakan tahap menghapus tanda baca di data teks, seperti (.) (,) (?), dan (angka) (Dyo fatra et al., 2020).
*Removal stopwords* menurut penelitian Wasim Bourequat merupakan teknik menghilangkan kata yang tidak berarti (Bourequat & Mourad, 2021).Contoh kata hubung: "dan" "atau"

*Stemming* menurut penelitian Asvarizal Filcha merupakan teknik transformasi kata menjadi kata dasar sebenarnya (Filcha & Hayaty, 2019). Contoh *stemming* sebagai berikut: "menyapu" -> sapu.

Pembobotan kata menurut penelitian Jeremy Andre Septian dan penelitian Faizal Nur Rozi *term inverse document matrix* merupakan tahapan menghitung frekuensi kalimat yang dipecah menjadi kata untuk melihat jumlah frekuensi kata dari masing-masing dokumen atau disebut dengan *term frequency*, hasil dari frekuensi kata kemudian menghitung jumlah dokumen dan jumlah frekuensi kata di masing-masing dokumen disebut dengan *inverse document matrix*, kemudian dilakukan perhitungan berdasarkan kata yang berada di dokumen (term frekuensi) dikalikan dengan *inverse document matrix* (Septian et al., 2019) (Rozi & Sulistyawati, 2019).

*Wordcloud* merupakan visualisasi untuk melihat berbagai macam label data teks dari label positif, label negatif, dan label netral (Naury et al., 2021). *Wordcloud* pada sentimen analisis melakukan *text preprocessing* terlebih dahulu agar terlihat kata-kata tanpa adanya kata yang tidak diperlukan (Cahyaningrum et al., 2020).

Pemodelan menurut penelitian Sebastian Raschka adalah kata hipotesis dan model sering digunakan secara sinonim dalam bidang pembelajaran mesin (Raschka, 2018). Pemodelan pada tahap ini setelah memproses data teks menggunakan pemodelan *supervised learning*, sebagai berikut:

1. *Decision tree*
   *Decision tree* menurut penelitian Apriliani dan penelitian Chee Sun Lee merupakan algoritma *supervised learning* yang mempunyai struktur seperti pohon, yang mempunyai simpul untuk atribut pengujian, setiap cabang mewakili hasil pengujian, dan daun mewakili kelas (Apriliani et al., 2020) (Lee et al., 2022).

Persamaan *decision tree* sebagai berikut:

$$Gini(D) = 1\sum_{i=1}^{m} p_i^2 \quad (1)$$

M = Adalah kategori atau kelas (Xu et al., 2020).

$p_i^2$ =Adalah kemungkinan atau probabilitas kelas (Zhao, 2022)

Pada persamaan (1), (2) melakukan perhitungan *gini impurity* untuk mencari nilai atribut masing-masing fitur (Jananto et al., 2021).
Sesudah melakukan perhitungan impurity masing-masing fitur maka hitung total nilai *impurity* pada persamaannya di gambar 2-3 kemudian mencari nilai terkecil dari masing-masing atribut untuk menjadi root (Jananto et al., 2021).

$$Gini_A = \left|\frac{D_i}{D}\right| Gini(D_1) + |\frac{D_2}{D}|Gini(D_2) \quad (2)$$

$|\frac{D_i}{D}|Gini(D_1) =$
Nilai perhitungan dari masing – masing kelas (Xu et al., 2020)

$|\frac{D_2}{D}|Gini(D_2) =$
Nilai perhitungan dari masing – masing kelas (Xu et al., 2020)

2. *Naives bayes*
   algoritma yang seringkali digunakan dalam sentimen analisis karena pembelajaran dari fitur untuk pengujian data untuk menghasilkan kemungkinan atau probabilitas (Watrianthos et al., 2019). Persamaan *naives bayes* dengan cara sebagai berikut:

$$p(X|H) = \frac{p(X|H)p(H)}{p(X)} \quad (3)$$

X = variabel acak (Salsabila et al., 2022).

H = kelas atau label (Salsabila et al., 2022).

P(X|H) = Berdasarkan X, probabilitas H dihitung (Salsabila et al., 2022).

P(X) = kemungkinan X (Salsabila et al., 2022).

Pada persamaan (3) (Salsabila et al., 2022) merupakan rumus *naives bayes*. *Naives bayes* memperlukan perhitungan setiap kelas, rumus untuk mencari kemungkinan setiap kelas sebagai berikut:

$$p(x) = \frac{N_x}{N} \quad (4)$$

P(x) = kemungkinan X (Salsabila et al., 2022).

Nx = Jumlah dari kelas x (Salsabila et al., 2022).

N = Jumlah gabungan semua kelas (Salsabila et al., 2022).

Pada persamaan (4) (Salsabila et al., 2022) merupakan rumus *naives bayes* mencari probalitas setiap kelas.

3. *K-Nearest-Neighbor*
   Dalam penerapan *text mining* atau klasifikasi menggunakan data teks dengan *K-nearest-neighbor* harus menentukan nilai k dari bobot kata *term frequency inverse document* dikalkulasi untuk melihat kemiripan antar dokumen (Dwiki et al., 2021). Tahapan menghitung *K-Nearest-Neighbor* sebagai berikut:
   1. Memilih nilai K (Dwiki et al., 2021)
   2. Menghitung tingkat kemiripan menggunakan *cosine similiarity* hasil term *frequecny imverse document* (Dwiki et al., 2021).
   3. *Sorting* data secara *descending dari* hasil komputasi *cosine similiarity* (Dwiki et al., 2021)
   4. Mengambil sebanyak nilai K yang paling tinggi kemiripannya dengan dokumen yang sudah diurutkan (Dwiki et al., 2021)

Setelah penjelasan runtutan perhitungan pembelajaran mesin *k-nearest neighbor* berikut rumus perhitungan *cosine similiarity* yang digunakan untuk menghitung jarak:

Evaluasi adalah tahap untuk mengukur keakuratan model, untuk model klasifikasi memiliki metode presisi, *recall*, akurasi (Fidan, 2020). Untuk menghitung metode *precision, recall, accuration* harus memperhatikan tp (*true positive)*, fn (*false negative)*, fp (*false positive)*, tn (*true negative).* Berikut cara menghitung keempat metode:

1. *Recall*

$$CosSim(q\ d_j)\frac{\vec{d_j}\vec{q}}{\left|\vec{d_j}\right|\left|\vec{q}\right|} = \frac{\sum_{i=1}^{t}(w_{ij} * w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 * \sum_{i=1}^{t} w_q^2}} \quad (5)$$

$\sum_{i=1}^{t}(w_{ij} * w_{iq})$ = Adalah total dari *vector* setiap dokumen ($w_{ij}$) dan dikalikan dengan total *vector* setiap *query* atau total *vector inpu* ($w_{iq}$) (Darmalaksana et al., 2020)

$\sum_{i=1}^{t} w_{ij}^2 * \sum_{i=1}^{t} w_{iq}^2$ = Adalah total dari *vector* setiap dokumen ($w_{ij}$) di pangkatkan 2 dan dikalikan dengan total *vector* setiap *query* ($w_{iq}$) dipangkatkan 2

Merupakan rasio data yang bernilai relevan dari data uji yang diambil (Bahassine et al., 2020).

$$recall = \frac{TP}{TP + FN} \quad (6)$$

*Tp = true positive*
*Fn = false negative*

2. *Precision*

Menurut penelitian Hongwon Yun untuk mengukur hasil dari data uji seberapa banyak sampel yang meghasilkan menjadi *true positive* (Yun, 2021).

$$precision = \frac{TP}{TP + FP} \quad (7)$$

*Tp = true positive*
*Fn = false negative*

3. *Accuracy*

Menurut penelitian Hongwon Yun

diperoleh dari dengan cara membagi jumlah yang diprediksi dengan data uji dengan menambah jumlah hasil *true positive dan true negative* (Yun, 2021).

$$acc = \frac{TP + FN}{TP + TN + FP + FN} \quad (8)$$

*Tp = true positive*
*Fn = false negative*
*Fp = false positive*
*Tn = true negative*

4. *Confusion matrix*

Merupakan hasil dari evaluasi dengan model yang diuji menggunakan data *testing* menghasilkan output berupa baris dan kolom yang didalamnya ada *true negative, true positive, false positive, false negative* (Hasnain et al., 2020).

$$\begin{bmatrix} True\ Positive & False\ Negative \\ False\ Negative & True\ Negative \end{bmatrix}$$

Gambar 1 confusion matrix

## 3. Metodologi penelitian

Data yang diambil dari opini pelanggan Tokopedia di website Tokopedia pada bagian ulasan dari pelanggan yang membeli produk elektronik (laptop handphone), dan produk pakaian (kaos, kemeja). Data elektronik yang digunakan sebagai penelitian sebanyak 925 data, berikut grafik data kategori elektronik (laptop, handphone).

Data Hp
1. rating 1 berjumlah 70 data
2. rating 2 berjumlah 20 data
3. rating 3 berjumlah 74 data
4. rating 4 berjumlah 170 data
5. rating 5 berjumlah 65 data

Data elekronik
1. rating 1 berjumlah 110 data
2. rating 2 berjumlah 32 data
3. rating 3 berjumlah 89 data
4. rating 4 berjumlah 123 data
5. rating 5 berjumlah 142 data

Data kategori pakaian terdiri dari kemeja, kaos menghasilkan 1575 Data, berikut data kategori pakaian yang tediri dari kemeja, dan kaos.

Data kaos
1. rating 1 berjumlah 161 data
2. rating 2 berjumlah 84 data
3. rating 3 berjumlah 190 data
4. rating 4 berjumlah 245 data
5. rating 5 berjumlah 250 data

Data kemeja
1. rating 1 berjumlah 100 data
2. rating 2 berjumlah 46 data
3. rating 3 berjumlah 145 data
4. rating 4 berjumlah 154 data
5. rating 5 berjumlah 200 data

*Text preprocessing*

Tahap *preprocessing text* adalah tahap untuk menyiapkan data teks sebelum dilakukan pelatihan ke pemodelan *machine learning*, berikut tahapan preprocessing yang dilakukan pada penelitian ini.

Pada tahap pengumpulan data maka dilakukan pelabelan, untuk rentang rating [1,2] diberikan label negatif, rating 3 diberikan label netral, rating [4,5] diberikan label positif, hasil dari pelabelan data dijelaskan sebagai berikut:

Data elektronik
1. Total label positif adalah 500 data.
2. Total label negatif adalah 232 data.
3. Total label netral adalah 163 data.

Data pakaian
1. Total label positif adalah 849 data.
2. Total label negatif adalah 391 data.
3. Total label netral adalah 335 data.

*Casefolding* merupakan tahap untuk transformasi data teks menjadi huruf kecil. *Punctuation removal* merupakan tahapan untuk menghapus tanda baca dan nomor karena agar tidak memperbanyak bobot kata pada tahap pembobotan kata. *Stopwords removal* merupakan tahapan untuk menghilangkan kata hubung, *stopword*s yang digunakan ["yg","tg","nya","deh","dan", "atau", "dengan", "bahwa", "namun", "meskipun", "sedangkan"]. *Stemming* merupakan tahapan transformasi teks data kata menjadi ke bentuk dasar, berikut hasil *text preprocessing.*

Tabel 1 hasil *text preprocessing*

| Data | Komentar (ulasan) | Casefolding | Punctuation removal | Stopwords removal |
|---|---|---|---|---|
| Data elektronik | terkecoh banget | terkecoh banget | terkecoh banget | terkecoh banget |

| | | | | |
|---|---|---|---|---|
| | sama variannya ternyata yang di klik 9a. bukan 9c. semo | sama variannya ternyata yang di klik 9a. bukan 9c. semoga awet deh ya | sama variannya ternyata yang di klik a bukan c semoga awet deh ya | varian klik a semoga awet hp tmx seller |

*Wordcloud* atau awan kata digunakan setelah melakukan *text preprocessing*, untuk melihat kata yang muncul dari label positif, netral, negatif. *Wordcloud* yang ditampilan pada penelitian sebagai berikut:


Gambar *1* kumpulan *wordcloud*

Pada gambar 1 merupakan *wordlcoud*
Pertama data elektronik label positif terdiri dari:

- klik
- banget
- awet
- Kecoh
- deh
- varian

Kedua data pakaian label netral terdiri dari:

- cepat
- kirim
- kain

Ketiga data pakaian label negatif tediri dari:

- sesuai
- pdhal
- barang
- kecewa

Pada tahap ini melakukan pembobotan kata cara kerja tahap ini memecah kalimat data teks menjadi per kata atau *term*, mengitung kemunculan *term* disetiap dokumen, menghitung *inverse document frequency* dengan rumus komputasi sebagai berikut:

Tabel 2 pembobotan kata

| *Term* | IDF | TFIDF | | | |
|---|---|---|---|---|---|
| | | | | | |
| lebar | 0,602 | | | 0,602 | |
| awet | 0,602 | 0,602 | | | |
| selamat | 0,602 | | 0,602 | | |
| sesuai | 0,602 | 0 | 0 | 0 | 1,204 |

Pada tahap pemodelan merupakan tahap untuk melatih data menggunakan *machine learning* pada penelitian ini menggunakan *decision tree, naives bayes, k-neareast neighbor*.

1. *Naives bayes*
   Pada tahap ini menggunakan *machine learninng naives*, berikut melakukan perhitungan *navies bayes*:

Melakukan perhitungan setiap label

Tabel 3 komputasi tabel probabilitas setiap label

| | |
|---|---|
| positif | $p(positif) = \frac{2}{4} = 0{,}5$ |
| negatif | $p(negatif) = \frac{1}{4} = 0{,}25$ |
| netral | $p(netral) = \frac{1}{4} = 0{,}25$ |

Setelah melakukan perhitungan probabilitas setiap label, langkah selanjutnya menghitung kata terhadap label dapat dilihat pada tabel 5.

Tabel 4 perhitungan kata terhadap label tabel

| Kata | Label | Perhitungan |
|---|---|---|
| lebar | positif | $p(lebar\|positif) = \frac{0+1}{2+4} = 0{,}166$ |
| | negatif | $p(lebar\|negatif) = \frac{0+1}{1+4} = 0{,}2$ |
| | netral | $p(lebar\|netral) = \frac{0{,}602+1}{1+4} = 0{,}320$ |

2. *Decision tree*
   Pada tahap pembelajaran mesin *decision tree* atau pohon keputusan, dilakukan perhitungan menggunakan hasil pembobotan kata. Melakukan normalisasi dari *term frequency inverse document* dapat dilihat pada tabel 5.

Tabel 5 perhitungan normalisasi

| Kata-kata | Perhitungan |
|---|---|
| lebar dan awet | $rata - rata\ tfidf_{(lebardanawet)}$ |
| | $\frac{0,602599913279624 + 0,602599913279624}{2}$ |

| | |
|---|---|
| | $= 0,6020599913279624$ |
| selamat dan sesuai | $rata - rata\ tf\,idf_{[selamat\ dan\ sesuai]}$ |
| | $\frac{0,6020599913279624 + 1,2041199826559248}{2}$ |
| | $= 0,9030900000000001$ |

Perhitungan normalisasi dilakukan, maka selanjutnya menghitung *gini impurity* dapat dilihat pada tabel 7 dan tabel 8.

Tabel 6 perhitungan *gini* 0,602

| Probabilitas | label | Perhitungan |
|---|---|---|
| Probabilitas(gini)<0,602 | 0 positif<br>0 netral<br>0 negatif | $1 - (\frac{0}{0})^2 - (\frac{0}{0})^2 - (\frac{0}{0})^2 = 0$ |
| Probabilitas(gini)>0,602 | 0 positif<br>0 netral<br>0 negatif | $1 - (\frac{0}{1})^2 - (\frac{0}{1})^2 - (\frac{1}{1})^2 = 0$ |
| Total *gini impurity* | | $(\frac{0}{1}) * 0 + (\frac{1}{1}) = 0$ |

Tabel 7 perhitungan *gini* 0,9031

| Probabilitas | label | Perhitungan |
|---|---|---|
| Probabilitas(gini)<0,9031 | 2 positif<br>1 netral<br>0 negatif | $1 - (\frac{2}{3})^2 - (\frac{1}{3})^2 - (\frac{0}{3})^2 = 0,44$ |
| Probabilitas(gini)>0,9031 | 0 positif<br>0 netral<br>1 negatif | $1 - (\frac{0}{1})^2 - (\frac{0}{1})^2 - (\frac{1}{1})^2 = 0$ |
| Total *gini impurity* | | $(\frac{3}{4}) * 0,44 + (\frac{1}{4}) * 0 = 0,33$ |

*Gini impurity term frequency inverse document* <0,6020599913279624 lebih kecil dari *gini impurity* 0,9030900000000001.

3. *K-nearest neighbor*

Pada tahap pembelajaran mesin *k-nearest neighbor. K-nearest neighbor* bekerja berdasarkan label dari nilai K tetangga terdekat. Berikut perhitungan *k-nearest neighbor*:

Tabel 8 *cosine similiarity* kata lebar

Tabel 9 *cosine similiarity* kata sesuai

Diurutkan dengan nilai terbesar, maka Dokumen 4 mempunyai nilai paling besar dibanding dokumen 3, kemudian dilakukan perankingan:

Tabel 10 perhitungan mencari hasil dengan nilai K

| |
|---|
| $D4 = 0,8027466551039498$ |
| $D3 = 0,44798065223573263$ |
| Ambil nilai K = 1 |
| $D4$ $= kelas\ atau\ labelnya\ adalah\ negatif$ |

Kesimpulan bahwa kueri "lebar sesuai" menghasilkan kelas negatif.

## 4. Hasil dan pembahasan

Pada hasil penelitian didapatkan hasil pemodelan pembelajaran mesin yang tediri dari *decision tree, k nearest neighbor, naives bayes*, berikut hasil tiga pemodelan dengan data pakaian dan data elektronik:

A. Data pakaian

***p-ISSN 2089-0265***
***e-ISSN 2598-3016***

Tabel 11 haisl evaluasi data pakain

| pemodelan | Data skenario | *accuracy* | *Recall* | *precision* |
|---|---|---|---|---|
| Naives bayes | 40% data uji dan 60% data latih, | 37% | 47% | 52% |
| Decision tree | 10% data uji dan 90% latih | 66% | 66% | 65% |
| *K nearest neighbor* (nilai K=5) | 10% data uji dan 90% data latih | 63% | 63% | 62% |

B. Data elektronik

Tabel 12 haisl evaluasi data elektronik

| pemodelan | Data skenario | *accuracy* | *Recall* | *precision* |
|---|---|---|---|---|
| Naives bayes | 40% data uji dan 60% data latih, | 41% | 41% | 53% |
| Decision tree | 40% data uji dan 60% latih | 66% | 66% | 65% |
| *K nearest neighbor* (nilai K=4) | 10% data uji dan 90% data latih | 66% | 63% | 62% |

## 5. Kesimpulan

Dari hasil peneltitian yang telah dilakukan dengan melakukan komparasi algoritma pembelajaran mesin *supervised learning* menggunakan data sentimen ulasan pelanggan di Tokopedia dari ulasan pelanggan yang membeli elektronik, pelanggan membeli pakaian, sebagai berikut:

A. Data penelitian

Data yang diambil sebanyak 1575 data untuk data pakaian dan 895 data untuk data elektronik

Kemudian melakukan pelabelan untuk data pakaian menghasilkan label positif sebanyak 849 data, label netral sebanyak 391 data, label negatif sebanyak 335 data, untuk pelabelan data elektronik menghasilkan label positif sebanyak 500 data, label netral sebanyak 168 data, label negatif sebanyak 232 data.

B. Setelah melakukan pelabelan, maka lakukan *preprocessing text* terdiri dari *casefolding, stopword removal, punctuation removal, stemming.*

C. Evaluasi pada saat sesudah melakukan pemodelan, maka lakukan pengujian pemodelan, berikut hasil uji terbaik.

Data pakaian

*Decision tree*

Hasil *decision tree* dari ke empat skenario menghasilkan hasil terbaik, yaitu 10% data uji dan 90% data latih, menghasilkan akurasi 66%, *recall* 66%, *precision* 65%.

Data elektronik

*Decision tree*

Hasil *decision tree* dari ke empat skenario menghasilkan hasil terbaik, yaitu 40% data uji dan 60% data latih, menghasilkan akurasi 66%, *recall* 66%, *precision* 65%.

Kesimpulan sentimen kepada penjual dan pembeli. Melihat dari sentimen yang dihasilkan dari gambar 1, penjual harus memperhatikan dari segi pelayanan pengiriman, dan juga barang yang dijual agar tidak mengecewakan konsumen dan membuat calon konsumen tertarik untuk membeli barang, untuk calon konsumen bisa melihat berbagai pengalaman yang diberikan konsumen ke toko penjual seperti barang yang datang dalam keadaan aman (opini positif), konsumen yang memberikan opini netral tentang pengiriman yang diberikan oleh penjual, dan opini bersifat negatif seperti kekecewaan konsumen membeli barang di toko tersebut.

**UCAPAN TERIMA KASIH**

Mengucapkan terima kasih kepada LPPM UMJ telah mendanai penelitian yang

dilakukan.

**Daftar Pustaka**

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. *Jurnal Rekayasa Teknologi Nusa Putra*, *6*(1), 54–62. https://doi.org/10.52005/REKAYASA.V6I1.86

Apriliani, D., Abidin, T., Sutanta, E., Hamzah, A., & Somantri, O. (2020). Sentiment analysis for assessment of hotel services review using feature selection approach based-on *decision tree*. *International Journal of Advanced Computer Science and Applications*, *11*(4), 240–245. https://doi.org/10.14569/IJACSA.2020.0110432

Arbaini, P. (2020). PENGARUH CONSUMER ONLINE RATING DAN REVIEW TERHADAP KEPUTUSAN PEMBELIAN PADA PENGGUNA MARKETPLACE TOKOPEDIA. *Jurnal Bisnis Dan Manajemen*, *7*(1). https://doi.org/10.26905/jbm.v7i1.3897

Bahassine, S., Madani, A., Al-Sarem, M., & Kissi, M. (2020). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University - Computer and Information Sciences*, *32*(2), 225–231. https://doi.org/10.1016/j.jksuci.2018.05.010

Bharathi, R., Bhavani, R., & Priya, R. (2022). TWITTER TEXT SENTIMENT ANALYSIS OF AMAZON UNLOCKED MOBILE REVIEWS USING SUPERVISED LEARNING TECHNIQUES. *Indian Journal of Computer Science and Engineering*, *13*(4), 1242–1253. https://doi.org/10.21817/indjcse/2022/v13i4/221304100

Bourequat, W., & Mourad, H. (2021). Sentiment Analysis Approach for Analyzing iPhone Release using Support Vector Machine. *International Journal of Advances in Data and Information Systems*, *2*(1), 36–44. https://doi.org/10.25008/ijadis.v2i1.1216

Cahyaningrum, N. I., Yoshida Fatima, D. W., Kusuma, W. A., Ramadhani, S. A., Destanto, M. R., & Nooraeni, R. (2020). Analysis of User Sentiment of Twitter to Draft KUHP. *Jurnal Matematika, Statistika Dan Komputasi*, *16*(3), 273. https://doi.org/10.20956/jmsk.v16i3.8239

Cunningham, P., & Delany, S. J. (2021). K-Nearest Neighbour Classifiers-A Tutorial. In *ACM Computing Surveys* (Vol. 54, Issue 6). Association for Computing Machinery. https://doi.org/10.1145/3459665

Dang, C. N., Moreno-García, M. N., & De la Prieta, F. (2021). An approach to integrating sentiment analysis into recommender systems. *Sensors*, *21*(16). https://doi.org/10.3390/s21165666

Darmalaksana, W., Slamet, C., Zulfikar, W. B., Fadillah, I. F., Maylawati, D. S. adillah, & Ali, H. (2020). Latent semantic analysis and cosine similarity for hadith search engine. *Telkomnika (Telecommunication Computing Electronics and Control)*, *18*(1), 217–227. https://doi.org/10.12928/TELKOMNIKA.V18I1.14874

Demircan, M., Seller, A., Abut, F., & Akay, M. F. (2021). Developing Turkish sentiment analysis models using machine learning and e-commerce data. *International Journal of Cognitive Computing in Engineering*, *2*, 202–207. https://doi.org/10.1016/j.ijcce.2021.11.003

Dwiki, A., Putra, A., & Juanita, S. (2021). Analisis Sentimen pada Ulasan pengguna Aplikasi Bibit Dan Bareksa dengan Algoritma KNN. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *8*(2), 636–646. https://doi.org/10.35957/JATISI.V8I2.962

Dyo fatra, A. H., Hayatin, N. H., & Aditya, C. S. K. (2020). Analisa Sentimen Tweet Berbahasa Indonesia Dengan Menggunakan Metode Lexicon Pada Topik Perpindahan Ibu Kota Indonesia.

*Jurnal Repositor*, *2*(7), 977. https://doi.org/10.22219/repositor.v2i7.937

Elmurngi, E. I., & Gherbi, A. (2018). Unfair reviews detection on Amazon reviews using sentiment analysis with supervised learning techniques. *Journal of Computer Science*, *14*(5), 714–726. https://doi.org/10.3844/jcssp.2018.714.726

Fidan, H. (2020). Grey Relational Classification of Consumers' Textual Evaluations in E-Commerce. *Journal of Theoretical and Applied Electronic Commerce Research*, *15*(1), 48–65. https://doi.org/10.4067/S0718-18762020000100105

Filcha, A., & Hayaty, M. (2019). Implementasi Algoritma Rabin-Karp untuk Pendeteksi Plagiarisme pada Dokumen Tugas Mahasiswa. *JUITA : Jurnal Informatika*, *7*(1), 25. https://doi.org/10.30595/juita.v7i1.4063

Firdaus, M. F. El, Nurfaizah, N., & Sarmini, S. (2022). Analisis Sentimen Tokopedia Pada Ulasan di Google Playstore Menggunakan Algoritma Naïve Bayes Classifier dan K-Nearest Neighbor. *JURIKOM (Jurnal Riset Komputer)*, *9*(5), 1329–1336. https://doi.org/10.30865/JURIKOM.V9I5.4774

Gian, M., & Ikate, S. (2021). Development of Electronic Business From the Historical Point of View of an E-Commerce Concept. *Journal Dimensie Management and Public Sector*, *2*(2), 19–24. https://doi.org/10.48173/jdmps.v2i2.91

Hasnain, M., Pasha, M. F., Ghani, I., Imran, M., Alzahrani, M. Y., & Budiarto, R. (2020). Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking. *IEEE Access*, *8*, 90847–90861. https://doi.org/10.1109/ACCESS.2020.2994222

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. *Big Data and Cognitive Computing*, *4*(1), 1–34. https://doi.org/10.3390/bdcc4010001

Jananto, A., Sulastri, S., Nur Wahyudi, E., & Sunardi, S. (2021). Data Induk Mahasiswa sebagai Prediktor Ketepatan Waktu Lulus Menggunakan Algoritma CART Klasifikasi Data Mining. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *10*(1), 71–78. https://doi.org/10.32736/sisfokom.v10i1.991

Kabiru, I. N., & Sari, P. K. (2019). Analisa Konten Media Sosial E-commerce Pada Instagram Menggunakan Metode Sentiment Analysis Dan Lda-based Topic Modeling (studi Kasus: Shopee Indonesia). *EProceedings of Management*, *6*(1).

Lee, S., Lee, C., Mun, K. G., & Kim, D. (2022). *Decision tree* Algorithm Considering Distances between Classes. *IEEE Access*, *10*, 69750–69756. https://doi.org/10.1109/ACCESS.2022.3187172

Maulana, F., & Santy, R. D. (2021). Pengaruh Ulasan Online Terhadap Niat Beli Dengan Kepercayaan Sebagai Intervening (Studi Kasus Terhadap Pengguna Aplikasi Tokopedia di Kota Bandung). *Journal of Economics, Management, Business and Accounting (JEMBA)*, *1*(1), 84–92. https://doi.org/10.34010/JEMBA.V1I1.5022

Naury, C., Fudholi, D. H., & Hidayatullah, A. F. (2021). Topic Modelling pada Sentimen Terhadap Headline Berita Online Berbahasa Indonesia Menggunakan LDA dan LSTM. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(1), 24. https://doi.org/10.30865/mib.v5i1.2556

Pravina, A. M., Cholissodin, I., & Adikara, P. P. (2019). Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(3), 2789–2797. http://j-ptiik.ub.ac.id

Ramadhan, N. G., & Ramadhan, T. I. (2022). Analysis Sentiment Based on IMDB Aspects from Movie Reviews using SVM. *Sinkron*, *7*(1), 39–45. https://doi.org/10.33395/sinkron.v7i1.11

204

Raschka, S. (2018). *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*. https://arxiv.org/abs/1811.12808v3

Ricky, R. D. M., Kawung, E., & Goni, S. Y. V. . (2021). Dampak Aplikasi Belanja Online (Online Shop) di Masa Pandemi Covid-19 Terhadap Minat Belanja Masyarakat di Kelurahan Girian Weru Ii Kecamatan Girian Kota Bitung Provinsi Sulawesi Utara. *Jurnal Ilmiah*, *1*(ilmiah).

Salsabila, S. M., Alim Murtopo, A., & Fadhilah, N. (2022). Analisis Sentimen Pelanggan Tokopedia Menggunakan Metode Naïve Bayes Classifier. *Jurnal Minfo Polgan*, *11*(2), 30–35. https://doi.org/10.33395/jmp.v11i2.11640

Septiani, L., & Sibaroni, Y. (2019). Sentiment Analysis Terhadap Tweet Bernada Sarkasme Berbahasa Indonesia. *Jurnal Linguistik Komputasional*, *2*(2), 62–67. https://doi.org/10.26418/JLK.V2I2.23

Syahputra, R., Yanris, G. J., & Irmayani, D. (2022). SVM and Naïve Bayes Algorithm Comparison for User Sentiment Analysis on Twitter. *Sinkron*, *7*(2), 671–678. https://doi.org/10.33395/sinkron.v7i2.11430

Vega, A., Delima, ), Putry, N., Azima, F., Wulandari, T., Dian, ), Puspita, P., Akuntansi, S. P., Ekonomi, F., Bisnis, D., & Riau, U. M. (2021). Analisis Strategi Pemasaran Tokopedia di Masa Pandemi Covid-19. *Jurnal Pendidikan Tambusai*, *5*(2), 3249–3254. https://doi.org/10.31004/JPTAM.V5I2.1378

Watrianthos, R., Suryadi, S., Irmayani, D., Nasution, M., & Simanjorang, E. F. S. (2019). Sentiment Analysis Of Traveloka App Using Naïve Bayes Classifier Method. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, *8*, 7. www.ijstr.org

Xu, W., Ning, L., & Luo, Y. (2020). Wind speed forecast based on post-processing of numerical weather predictions using a gradient boosting *decision tree* algorithm. *Atmosphere*, *11*(7). https://doi.org/10.3390/atmos11070738

Yun, H. (2021). Prediction model of algal blooms using logistic regression and confusion matrix. *International Journal of Electrical and Computer Engineering*, *11*(3), 2407–2413. https://doi.org/10.11591/ijece.v11i3.pp2407-2413

Zhang, S., Zhang, D., Zhong, H., & Wang, G. (2020). A multiclassification model of sentiment for e-commerce reviews. *IEEE Access*, *8*, 189513–189526. https://doi.org/10.1109/ACCESS.2020.3031588

Zhao, A. (2022). Financial Risk Evaluation of Digital Currency Based on CART Algorithm Blockchain. *Mobile Information Systems*, *2022*. https://doi.org/10.1155/2022/1356480