PERBANDINGAN METODE *MACHINE LEARNING* UNTUK SENTIMEN ANALISIS *REVIEW* PENJUALAN PRODUK DI TOKOPEDIA

**TA**

**Sebagai Syarat Dalam Menyelesaikan Program Studi Strata Satu (S-1)**

**Program Studi Teknik Informatika**

A picture containing text, clipart

Description automatically generated

**Disusun Oleh:**

|  |  |  |
| --- | --- | --- |
| **Nama** | **:** | **Muhammad Reza** |
| **Nim** | **:** | **2019470055** |

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK**

**UNIVERSITAS MUHAMMADIYAH JAKARTA**

**2023**

PERBANDINGAN METODE *MACHINE LEARNING* UNTUK SENTIMEN ANALISIS *REVIEW* PENJUALAN PRODUK DI TOKOPEDIA

**TA**

**Sebagai Syarat Dalam Menyelesaikan Program Studi Strata Satu (S-1)**

**Program Studi Teknik Informatika**

A picture containing text, clipart

Description automatically generated

**Disusun Oleh:**

|  |  |  |
| --- | --- | --- |
| **Nama** | **:** | **Muhammad Reza** |
| **Nim** | **:** | **2019470055** |

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK**

**UNIVERSITAS MUHAMMADIYAH JAKARTA**

**2023**

**UNIVERSITAS MUHAMMADIYAH JAKARTA FAKULTAS**

**TEKNIK-PRODI TEKNIK INFORMATIKA**

**LEMBAR PERSETUJUAN**

**PERBANDINGAN METODE *MACHINE LEARNING* UNTUK SENTIMEN ANALISIS *REVIEW* PENJUALAN PRODUK DI TOKOPEDIA**

|  |  |  |
| --- | --- | --- |
| NAMA | : | Muhammad Reza |
| NIM | : | 2019470055 |
| PROGRAM STUDI | : | TEKNIK INFORMATIKA |

TA ini telah disetujui pada tanggal, 21 Juli 2023

Oleh

|  |
| --- |
| Pembimbing Utama |
|  |
| (Popy Meilina, S.T., M. Kom) |

|  |
| --- |
| Mengetahui |
| Ketua Program Studi Teknik Informatika |
| (Popy Meilina, S.T., M. Kom) |

**UNIVERSITAS MUHAMMADIYAH JAKARTA FAKULTAS**

**TEKNIK-PRODI TEKNIK INFORMATIKA**

**TANDA BUKTI PERSETUJUAN PEMBIMBING TA**

Pada Semester Ganjil/Genap

Tahun Akademik 2022/2023

Yang bertanda tangan dibawah ini Pembimbing TA menyetujui bahwa:

|  |  |  |
| --- | --- | --- |
| Nama | : | Muhammad Reza |
| Nim | : | 2019470055 |

PERBANDINGAN METODE *MACHINE LEARNING* UNTUK SENTIMEN ANALISIS *REVIEW* PENJUALAN PRODUK DI TOKOPEDIA

|  |  |  |
| --- | --- | --- |
| Dimulai bulan, tahun | : | ……………………… |
| Selesai bulan, tahun | : | ……………………… |

Untuk ikut serta Ujiuan Sidang Strata Satu (S1) yang diselenggarakan oleh Program Studi Teknik Informatika Universitas Muhammadiyah Jakarta

Jakarta, ……………………….

|  |
| --- |
| Pembimbing Utama |
|  |
| (Popy Meilina, S.T., M. Kom) |

**UNIVERSITAS MUHAMMADIYAH JAKARTA FAKULTAS  
TEKNIK-PRODI TEKNIK INFORMATIKA**

**LEMBAR PENGESAHAN**

**PERBANDINGAN METODE *MACHINE LEARNING* UNTUK SENTIMEN ANALISIS *REVIEW* PENJUALAN PRODUK DI TOKOPEDIA**

|  |  |  |
| --- | --- | --- |
| NAMA | : | Muhammad Reza |
| NIM | : | 2019470055 |
| PROGRAM STUDI | : | TEKNIK INFORMATIKA |

TA ini telah diuji pada tanggal...............

**Oleh Penguji**

|  |  |  |  |
| --- | --- | --- | --- |
| 1. | Nama Penguji I | : | .......................................... |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| 2. | Nama Penguji II | : | .......................................... |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| 3. | Nama Pembimbing Utama | : | Popy Meilina, S.T., M. Kom |

**UNIVERSITAS MUHAMMADIYAH JAKARTA FAKULTAS  
TEKNIK-PRODI TEKNIK INFORMATIKA**

**LEMBAR PERNYATAAN**

Bersama ini saya menyatakan bahwa isi yang terkandung dalam TA ini, dengan Judul:

PERBANDINGAN METODE *MACHINE LEARNING* UNTUK SENTIMEN ANALISIS *REVIEW* PENJUALAN PRODUK DI TOKOPEDIA

Adalah murni merupakan hasil penelitian dan pemikiran saya sendiri.

Demikian pernyataan ini saya buat dan siap menerima konsekuensi apapun di masa yang akan datang apabila ternyata TA ini merupakan salinan ataupun contoh karya-karya yang telah dibuat / diterbitkan sebelum tanggal TA ini

Jakarta, .......................

|  |
| --- |
| Penulis |
|  |
| (Muhammad Reza) |

**UNIVERSITAS MUHAMMADIYAH JAKARTA FAKULTAS  
TEKNIK-PRODI TEKNIK INFORMATIKA**

**SURAT PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS**

Yang bertanda tangan di bawah ini, saya:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nama | : | Muhammmad Reza |  |  |
| NIM | : | 2019470055 |  |  |
| Program Studi | : | Teknik Informatika |  |  |
| Jenjang | : | Strata Satu (S1) Jenis Karya | : | TA |

Demi pengembangan ilmu pengetahuan, dengan ini menyetujui untuk memberikan ijin kepada pihak Program Studi Teknik Informatika FT-UMJ **Hak Bebas Royalti NonEksklusif *(Non-exclusive Royalti-Free Right)*** atas karya ilmiah saya yang berjudul: “PERBANDINGAN METODE *MACHINE LEARNING* UNTUK SENTIMEN ANALISIS *REVIEW* PENJUALAN PRODUK DI TOKOPEDIA” beserta perangkat yang diperlukan (apabila ada).

Dengan **Hak Bebas Royalti Non-Eksklusif** ini pihak FT-UMJ berhak menyimpan,  
mengalih-media atau *bentuk*-kan, mengelolanya dalam pangkalan data *(database)*,  
mendistribusikannya dan menampilkan atau mempublikasikannya di *internet* atau media lain untuk kepentingan akademis tanpa perlu meminta ijin dari Saya selama tetap mencantumkan nama kami sebagai penulis/pencipta karya ilmiah tersebut.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak FT- UMJ,  
segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya  
ilmiah saya ini.

Demikian pernyataan ini saya buat  
dengan sebenarnya.

Jakarta, ……………

|  |
| --- |
| Yang menyatakan, |
|  |
| (Muhammad Reza) |

**UNIVERSITAS MUHAMMADIYAH JAKARTA FAKULTAS  
TEKNIK-PRODI TEKNIK INFORMATIKA**

# DAFTAR PRESENSI BIMBINGAN TA

PERBANDINGAN *METODE MACHINE* LEARNING UNTUK SENTIMEN ANALISIS *REVIEW* PENJUALAN PRODUK DI TOKOPEDIA

|  |  |  |
| --- | --- | --- |
| Nama | : | Muhammad Reza |
| Nim | : | 2019470055 |
| Program Studi | : | Teknik Informatika |

Dosen Pembimbing Utama: Ibu Popy Meilina, S.T., M. Kom

|  |  |  |  |
| --- | --- | --- | --- |
| No | Tanggal | Catatan Dosen Pembimbing | Paraf |
| 1 | 07 – Maret 2023 | Pengajuan judul untuk tugas akhir |  |
| 2 | 13 – Maret 2023 | 1. Judul di terima, yaitu PERBANDINGAN METODE MACHINE LEARNING UNTUK SENTIMEN ANALISIS REVIEW PENJUALAN PRODUK DI TOKOPEDIA 2. Revisi di bab 1 |  |
| 3 | 16 – Maret 2023 | 1. Revisi bab 1 terkait identifikasi masalah 2. Konsul mengenai tentang mencari data penelitian apakah dibanyak toko atau satu toko |  |
| 4 | 21 – Maret - 2023 | 1. Memperbaiki penulisan bab 1 terkait referensi jurnal agar lebih ringkas lagi 2. Revisi bab 1 sub bab manfaat penelitian |  |
| 5 | 30 - Maret - 2023 | 1. Bab 1 sudah tidak ada lagi revisi 2. Melanjutkan ke bab2 untuk sub bab pemodelan dan evaluasi |  |
| 6 | 04 – April -2023 | 1. Menjelaskan proposal bab 2 2. Tidak ada masalah di bab 2, maka lanjutkan untuk bab 3 |  |
| 7 | 09 – Mei – 2023 | 1. Memperbaiki di bab 3 tidak perlu teori lagi akan tetapi pengerjaan penelitian |  |
| 8 | 12- Mei – 2023 | 1. Melanjutkan penyelesaian bab 3 tentang sub bab pemodelan, dan evaluasi |  |

Dosen Pembimbing

(Popy Meilina, S.T., M. Kom)

# Internalisasi AIK

# ABSTRACT

# ABSTRAK

# KATA PENGANTAR

*Alhamdulillaahirabbil‟aalamiin*, puji syukur penyusun panjatkan atas kehadirat Allah SWT yang telah melimpahkan rahmat dan hidayah-Nya, serta tidak lupa shalawat serta salam selalu tercurah kepada junjungan umat, yaitu Nabi Muhammad SAW sebagai suri tauladan umat, sehingga penyusunan tugas akhir yang berjudul “Perbandingan Metode *Machine* *Learning* Untuk Sentimen Analisis *Review* Penjualan Produk Di TOKOPEDIA” sebagai syarat untuk kelulusan jenjang strata satu di Jurusan Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jakarta. Dalam penyusunan proposal tugas akhir penyusun banyak memperoleh petunjuk dan bimbingan dari berbagai pihak.

Untuk selanjutnya penyusun mengucapkan banyak terima kasih kepada pihak-pihak yang telah membantu dalam penyelesaianTA ini, yaitu:

1. Dekan Fakultas Teknik Bapak Irfan Purnawan, S.T., M.Chem.Eng.
2. Ketua Program Studi Teknik Informatika Ibu Popy Meilina, S.T., M. Kom
3. Dosen pembimbing Ibu Popy Meilina, S.T., M. Kom
4. Kedua orang tua penyusun yang selalu memberikan do’a dan motivasi

Jakarta, ……, ……

Penyusun,

# DAFTAR ISI

[DAFTAR PRESENSI BIMBINGAN TA vii](#_Toc141227338)

[Internalisasi AIK ix](#_Toc141227339)

[ABSTRACT x](#_Toc141227340)

[ABSTRAK xi](#_Toc141227341)

[KATA PENGANTAR xii](#_Toc141227342)

[DAFTAR ISI xiii](#_Toc141227343)

[DAFTAR TABEL xvi](#_Toc141227344)

[DAFTAR GAMBAR xvii](#_Toc141227345)

[DAFTAR LAMPIRAN xx](#_Toc141227346)

[BAB I 1](#_Toc141227347)

[PENDAHULUAN 1](#_Toc141227348)

[1.1. Latar Belakang Masalah 1](#_Toc141227349)

[1.2. Identifikasi Masalah 3](#_Toc141227350)

[1.3. Perumusan Masalah 3](#_Toc141227351)

[1.4. Batasan Masalah 3](#_Toc141227352)

[1.5. Tujuan dan Manfaat Penelitian 4](#_Toc141227353)

[1.6. Metodologi Penelitian 5](#_Toc141227354)

[1.7. Sistematika Penulisan 33](#_Toc141227355)

[BAB II 35](#_Toc141227356)

[TINJAUAN PUSTAKA 35](#_Toc141227357)

[2.1. Tokopedia 35](#_Toc141227358)

[2.2. Analisis Sentimen 35](#_Toc141227359)

[2.3. Text mining 35](#_Toc141227360)

[2.4. Text preprocessing 37](#_Toc141227361)

[2.5. *Worldcloud* 38](#_Toc141227362)

[2.6. Pemodelan 38](#_Toc141227363)

[2.7. Evaluasi 41](#_Toc141227364)

[BAB III 43](#_Toc141227365)

[METODE PENELITIAN 43](#_Toc141227366)

[3.1. Data penelitian 43](#_Toc141227367)

[3.2. Text preprocessinng 44](#_Toc141227368)

[3.2.1. Pelabelan 44](#_Toc141227369)

[3.2.2. *Casefolding* 47](#_Toc141227370)

[3.2.3. *Punctuation removal* 49](#_Toc141227371)

[3.2.4. *Stopwords removal* 51](#_Toc141227372)

[3.2.5. *Stemming* 53](#_Toc141227373)

[3.2.6. *Wordcloud* 55](#_Toc141227374)

[3.2.7. Pembobotan kata 57](#_Toc141227375)

[3.3. Pemodelan 64](#_Toc141227376)

[3.3.2. Naives bayes 64](#_Toc141227377)

[3.3.3. Decision tree 67](#_Toc141227378)

[1.3.3. *K-nearest neighbor* 69](#_Toc141227379)

[3.3 Evaluasi 73](#_Toc141227380)

[3.3.1. Evaluasi naives bayes 73](#_Toc141227381)

[3.3.3. Evaluasi decision tree 75](#_Toc141227382)

[3.3.4. Evaluasi *k nearest neighbor* 77](#_Toc141227383)

[BAB IV 79](#_Toc141227384)

[HASIL DAN PEMBAHASAN 79](#_Toc141227385)

[4.1. Data Penelitian 79](#_Toc141227386)

[4.1.1. Data Pakaian 79](#_Toc141227387)

[4.1.1. Data Elektronik 80](#_Toc141227388)

[4.2. *Naives bayes* 82](#_Toc141227389)

[4.2.1. Data Pakaian 82](#_Toc141227390)

[4.2.2. Data elektronik 83](#_Toc141227391)

[4.3. *Decision tree* 83](#_Toc141227392)

[4.3.1. Data Pakaian 83](#_Toc141227393)

[4.3.2. Data elektronik 84](#_Toc141227394)

[4.4. *K-nearest neighbor* 85](#_Toc141227395)

[4.4.1. Data Pakaian 85](#_Toc141227396)

[4.4.2. Data elektronik 87](#_Toc141227397)

[4.5. Komparasi evaluasi model 89](#_Toc141227398)

[BAB V 92](#_Toc141227399)

[KESIMPULAN DAN SARAN 92](#_Toc141227400)

[Kesimpulan 92](#_Toc141227401)

[Saran-saran 92](#_Toc141227402)

[DAFTAR PUSTAKA 93](#_Toc141227403)

[LAMPIRAN 101](#_Toc141227404)

# DAFTAR TABEL

[Tabel 1- 1 tabel data penelitian 5](#_Toc141233322)

[Tabel 1- 2 tabel pelabelan data 6](#_Toc141233323)

[Tabel 1- 3 tabel casefolding data 8](#_Toc141233324)

[Tabel 1- 4 tabel punctuation removal 9](#_Toc141233325)

[Tabel 1- 5 tabel removal stopwords 10](#_Toc141233326)

[Tabel 1- 6 stemming 11](#_Toc141233327)

[Tabel 1- 7 hasil text preprocessing 12](#_Toc141233328)

[Tabel 1- 8 tabel wordcloud 13](#_Toc141233329)

[Tabel 1- 9 rumus term frequency inverse document 14](#_Toc141233330)

[Tabel 1- 10 tabel term frequency inverse document 15](#_Toc141233331)

[Tabel 1- 11 tabel bobot kata untuk decision tree 20](#_Toc141233332)

[Tabel 1- 12 perhitungan normaliasi 20](#_Toc141233333)

[Tabel 1- 13 tabel hasil normalisasi 21](#_Toc141233334)

[Tabel 1- 14 perhitungan gini 0,6020599913279624 21](#_Toc141233335)

[Tabel 1- 15 tabel data untuk perhitungan navies bayes 22](#_Toc141233336)

[Tabel 1- 16 komputasi tabel probabilitas setiap label 23](#_Toc141233337)

[Tabel 1- 17 perhitungan kata terhadap label 23](#_Toc141233338)

[Tabel 1- 18 pengujian naives bayes 24](#_Toc141233339)

[Tabel 1- 19 tabel perhitungan pengujian menggunakan naives bayes 24](#_Toc141233340)

[Tabel 1- 20 tabel perhitungan TF-IDF untuk cosine similiairty 25](#_Toc141233341)

[Tabel 3- 1 data elektronik 38](#_Toc141233465)

[Tabel 3- 2 data pakaian 39](#_Toc141233466)

[Tabel 3- 3 hasil pelabelan data elektronik 40](#_Toc141233467)

[Tabel 3- 4 hasil pelabelan data pakaian 40](#_Toc141233468)

[Tabel 3- 5 hasil pelabelan penggabungan data menjadi 2 kategori data 41](#_Toc141233469)

[Tabel 3- 6 empat data penelitian 42](#_Toc141233470)

[Tabel 3- 7 casefolding 5 data elektronik dan data pakaian 43](#_Toc141233471)

[Tabel 3- 8 punctuation removal 44](#_Toc141233472)

[Tabel 3- 9 tabel stopword removal 4 data elektronik dan data pakaian 46](#_Toc141233473)

[Tabel 3- 10 tabel stemming 4 data elektronik dan data pakaian. 48](#_Toc141233474)

[Tabel 3- 11 tabel worldcloud 50](#_Toc141233475)

[Tabel 3- 12 rumus perhitungan term frequency inverse document 52](#_Toc141233476)

[Tabel 3- 13 perhitungan term frequency inverse document kata “sesuai” 52](#_Toc141233477)

[Tabel 3- 14 tabel pembobotan kata “sesuai” setiap dokumen 53](#_Toc141233478)

[Tabel 3- 15 tabel pembobotan kata 53](#_Toc141233479)

[Tabel 3- 16 tabel kata dan label untuk naives bayes 59](#_Toc141233480)

[Tabel 3- 17 komputasi tabel probabilitas setiap label 59](#_Toc141233481)

[Tabel 3- 18 perhitungan kata terhadap label 60](#_Toc141233482)

[Tabel 3- 19 pengujian naives bayes 60](#_Toc141233483)

[Tabel 3- 20 tabel perhitungan pengujian menggunakan naives bayes 61](#_Toc141233484)

[Tabel 3- 21 tabel kata dan label untuk decision tree 61](#_Toc141233485)

[Tabel 3- 22 perhitungan rata-rata 62](#_Toc141233486)

[Tabel 3- 23 normalisasi 62](#_Toc141233487)

[Tabel 3- 24 perhitungan gini 0,6020599913279624 63](#_Toc141233488)

[Tabel 3- 25 perhitungan gini 0,9030900000000001 63](#_Toc141233489)

[Tabel 3- 26 tabel bobot kata untuk k nearest neighbour 64](#_Toc141233490)

[Tabel 3- 27 tabel perhitungan query 64](#_Toc141233491)

[Tabel 3- 28 tabel bobot kata hasil kalkulasi jarak 65](#_Toc141233492)

[Tabel 3- 29 perhitungan cosine similiarity kata lebar 66](#_Toc141233493)

[Tabel 3- 30 perhitungan cosine similiarity kata lebar 67](#_Toc141233494)

[Tabel 3- 31 evaluasi data pakaian 68](#_Toc141233495)

[Tabel 3- 32 evaluasi data elektronik 69](#_Toc141233496)

[Tabel 3- 33 evaluasi data pakaian 70](#_Toc141233497)

[Tabel 3- 34 evaluasi data pakaian 71](#_Toc141233498)

[Tabel 3- 35 evaluasi data pakaian 72](#_Toc141233499)

[Tabel 3- 36 evaluasi data elektronik 73](#_Toc141233500)

[Tabel 4- 1 data pakaian dengan label 74](#_Toc141233560)

[Tabel 4- 2 tabel wordcloud data pakaian 74](#_Toc141233561)

[Tabel 4- 3 data elektronik dengan label 75](#_Toc141233562)

[Tabel 4- 4 tabel wordcloud data elektronik 76](#_Toc141233563)

[Tabel 4- 5 evaluasi naives bayes data pakaian 77](#_Toc141233564)

[Tabel 4- 6 evaluasi naives bayes data elektronik 78](#_Toc141233565)

# 

# DAFTAR GAMBAR

[Gambar 1- 1 rumus atribut gini impurity 19](#_Toc141233235)

[Gambar 1- 2 rumus total atribut gini impurity 19](#_Toc141233236)

[Gambar 1- 3 rumus naives bayes 22](#_Toc141233237)

[Gambar 1- 4 rumus cosine similiarity 25](#_Toc141233238)

[Gambar 1- 5 perhitungan cosine similiarity kata recommended. 27](#_Toc141233239)

[Gambar 1- 6 perhitungan cosine similiarity kata pesan 27](#_Toc141233240)

[Gambar 1- 7 rumus akurasi 28](#_Toc141233241)

[Gambar 1- 8 rumus recall 28](#_Toc141233242)

[Gambar 1- 9 rumus precision 29](#_Toc141233243)

[Gambar 1- 10 confusion matrix 29](#_Toc141233244)

[Gambar 2. 1 wordclooud 38](file:///E:\COLLEGE\SEMESTER%208\Skripsi\Repo%20Github\skripsi\Skripsi\Penulisan\Skripsi_MuhammadReza_2019470055.docx#_Toc141227503)

[Gambar 2. 2 rumus gini impurity 39](file:///E:\COLLEGE\SEMESTER%208\Skripsi\Repo%20Github\skripsi\Skripsi\Penulisan\Skripsi_MuhammadReza_2019470055.docx#_Toc141227504)

[Gambar 2. 3 rumus total gini impurity 39](file:///E:\COLLEGE\SEMESTER%208\Skripsi\Repo%20Github\skripsi\Skripsi\Penulisan\Skripsi_MuhammadReza_2019470055.docx#_Toc141227505)

[Gambar 2. 4 rumus naives bayes 39](file:///E:\COLLEGE\SEMESTER%208\Skripsi\Repo%20Github\skripsi\Skripsi\Penulisan\Skripsi_MuhammadReza_2019470055.docx#_Toc141227506)

[Gambar 2. 5 rumus perhitungan setiap kelas 40](file:///E:\COLLEGE\SEMESTER%208\Skripsi\Repo%20Github\skripsi\Skripsi\Penulisan\Skripsi_MuhammadReza_2019470055.docx#_Toc141227507)

[Gambar 2. 6 rumus cosine similiarity 41](file:///E:\COLLEGE\SEMESTER%208\Skripsi\Repo%20Github\skripsi\Skripsi\Penulisan\Skripsi_MuhammadReza_2019470055.docx#_Toc141227508)

[Gambar 3. 3 Rumus cosine similiarity 70](#_Toc141227509)

# DAFTAR LAMPIRAN

# BAB I

# PENDAHULUAN

## Latar Belakang Masalah

Toko online atau *e-commerce* menurut Moossa Giant dan Samuel Ikate adalah operasi bisnis yang dilakukan secara dunia maya atau *online* (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., 2019a).

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). Pembelajaran mesin *naives bayes, decision tree* (pohon keputusan), *k-nearest neighbor* adalah pembelajaran mesin berdasarkan label (Hardanto & Wulandhari, 2021) (Chamorro-Atalaya 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).

Penelitian sebelumnya mengenai analisis sentimen yang menggunakan metode pembelajaran mesin berdasarkan label, menggunakan *decision tree* seperti penelitian menggunakan data dari ulasan aplikasi *e-commerce* di Google play store dengan total data 940, melakukan *preprocessing text* (pembersihan data)*,* menggunakan 750 data latih,190 data uji menghasilkan rata-rata akurasi 92,63%, precision 69,58%, dan recall 69,99% (Ikhsanti et al., 2021).

Penelitian analisis sentimen menggunakan pembelajaran mesin berdasarkan label *naives bayes* menggunakan data penilaian pelanggan di Toko *Online* *Shop* sebanyak 1,899 data*,* melakukan *preprocessing text* (pembersihan data)*,* menghasilkan akurasi 97,16% (Stone & Fathoni, 2022).

Penelitian analisis sentimen ulasan aplikasi Shoppe di *website* Google Playstore, data yang diteliti sebanyak 500 data, menggunakan pembelajaran mesin berdasarkan label, yaitu *navies bayes, k-nearest neighbor* sebagai metode penelitian. Hasil evaluasi *navies bayes* *accuracy* 0,914, *precision* 0,915, *recall* 0,914 dan *F1 score* 0,916. 2. Hasil evaluasi *k-nearest neighbor* *accuracy* 0,928, *precision* 0,929, *recall* 0,928, dan *F1 score* 0,926 (Limbong et al., 2022).

Berdasarkan beberapa penelitian yang sudah dipaparkan, maka penelitian ini melakukan analisa sentimen di *e-commerce* Tokopedia mengenai ulasan pelanggan di beberapa produk, yaitu elektronik, pakaian menggunakan metode *machine learning supervised decision tree, k nearest neighbor, naives bayes*.

## Identifikasi Masalah

Berdasarkan permasalahan di latar belakang, penelitian terdahulu **yaitu Limbong** mengenai analisis sentimen salah satunya adanya perbandingan dua pembelajaran mesin, maka penulis menggunakan tiga metode metode *machine learning* *supervised decision tree, k nearest neighbor, naives bayes,* yang kemudian dilakukan perbandingan agar dapat mengetahui akurasi dalam mengolah analisis sentimen.

## Perumusan Masalah

Berdasarkan permasalahan diatas, akan dilakukan perumusan atau kajian

sebagai berikut:

1. Bagaimana cara mengolah data teks untuk melakukan analisis sentimen, melakukan analisis sentimen dengan menghasilkan positif, nilai negatif, dan nilai netral, serta cara melakukan pelabelan sentimen berdasarkan rating bintang 1 sampai 5?
2. Bagaimana hasil dan akurasi perbandingan melakukan komparasi *machine learning* untuk analisis sentimen berdasarkan 3 kondisi positif, negatif, netral?

## Batasan Masalah

Proposal tugas akhir ini memiliki batasan agar lebih mengerucut dan tidak

melebar, maka diberikan batasan-batasan sebagai berikut:

1. Data yang diambil adalah ulasan produk elektronik (laptop, *handphone*), produk pakaian (kemeja, kaos).
2. Data Tokopedia kategori elektronik laptop diambil dari 16 Maret - 25 April 2023 sebanyak 398 data, kategori *handphone* diambil dari 19 April – 02 Mei 2023 sebanyak 495 data.
3. Data Tokopedia pakaian kategori kaos diambil dari 30 April - 03 Mei 2023 sebanyak 930 data, kategori kemeja diambil dari 30 April – 03 Mei 2023 sebanyak 645 data.
4. Melakukan penggabungan data kaos, kemeja mejadi data pakaian, dan melakukan penggabungan data *hp,* data laptop menjadi data elektronik, kemudian analisis sentimen dari data kategori pakaian, dan data kategori elektronik.
5. Mengolah data teks yang sudah didapatkan untuk dilakukan pelabelan, yaitu penilaian atau ulasan konsumen berdasarkan rating apakah opini tersebut positif, negatif, dan netral.
6. Melakukan pengolahan data dengan *case folding, Removal stopwords, stemming,* dan juga pembobotan *menggunakan Term Frequency Inverse Document*
7. Melakukan komparasi *machine learning decision tree, naives bayes, k - nearest neighbor* untuk sentimen analisis ulasan konsumen Tokopedia di fitur ulasan dan *review*.
8. Melakukan evaluasi dari tiga *machine learning,* menggunakan *confusion matrix, metrics accuracy, metrics recall, metrics precision.*
9. Hanya menampilkan hasil perhitungan dari ketiga metode dengan menggunakan data pakaian dan data elektronik.
10. *Dashboard* yang dibuat hanya menampilkan hasil senitmen opini.

## Tujuan dan Manfaat Penelitian

Proposal tugas akhir ini memiliki tujuan penelitian, manfaat sebagai berikut:

1. Melakukan sentimen analisis atau klasifikasi opini konsumen di fitur ulasan *review* pembelian Tokopedia menggunakan algoritma *decision tree, naives bayes, K - Nearest Neighbor.*
2. Membandingkan tiga *machine learning* algoritma, yaitu *decision tree, naives bayes, K - Nearest Neighbor* yang lebih baik berdasarkan akurasi.
3. Manfaat penelitian untuk mengetahui akurasi dari tiga metode *machine learning* yang dapat digunakan untuk sentimen analisis ulasan *review* pembelian produk di Tokopedia.

## Metodologi Penelitian

1. **Data penelitian**

Data yang digunakan dalam penelitian adalah data teks dari hasil mengambil *website* Tokopedia yang data didalamnya terdapat konten berupa komentar (opini pembeli) dan rating berupa bintang 1 s/d 5. Dari data yang diambil terdapat 2 kategori, yaitu data pakaian (kaos, kemeja) sebanyak 893 data, data elektronik (*handphone,* laptop) sebanyak 1,575 data, berikut beberapa data ulasan yang sudah diambil.

Tabel 1- 1 tabel data penelitian

|  |  |  |
| --- | --- | --- |
| Data | Komentar | Rating |
| Data pakaian (kemeja) | Nyaman model bagus jahitan rapi ukuran pas | 5 |
| Data pakaian  (kaos) | Barang tidak sesuai dengan yg z pesan pertama n kemarin z pesan warna navy pas dengan badan z kainx jga bagus terus z pesan lagi maroon ndh sama panjang dengan warna navy... | 1 |
| Data elektronik  (*handphone)* | Produk original, pengiriman lumayan cepet. Belum di coba.. Tp looks ok nih.. Recommended seller. | 4 |
| Data elektronik  (laptop) | kualitas barang bermasalah | 1 |

1. **Pengolahan data**

Pada tahap ini dilakukan proses sebagai berikut:

1. Pelabelan data

Pada tahap ini dilakukan pelabelan data berdasarkan rating, menurut penelitian Demircan (Demircan et al., 2021) terdapat pembagian 3 kategori sentimen, sebagai berikut:

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

Berikut pelabelan data yang dilakukan berdasarkan rating.

Tabel 1- 2 tabel pelabelan data

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Komentar | *Rating* | label |
| Data pakaian (kemeja) | Nyaman model bagus jahitan rapi ukuran pas | 5 | positif |
| Data pakaian  (kaos) | Barang tidak sesuai dengan yg z pesan pertama n kemarin z pesan warna navy pas dengan badan z kainx jga bagus terus z pesan lagi maroon ndh sama panjang dengan warna navy... | 1 | negatif |
| Data elektronik  (*handphone)* | Produk original, pengiriman lumayan cepet. Belum di coba.. Tp looks ok nih.. Recommended seller. | 4 | positif |
| Data elektronik  (laptop) | kualitas barang bermasalah | 1 | negatif |

1. *Case folding*

Merupakan tahap mengolah data teks jika memiliki huruf kapital atau *uppercase* makadiubah menjadi huruf kecil atau *lowercase* (KURNIAWAN & APRILIANI, 2020). Berikut contoh *casefolding* yang dilakukan di tabel 1-3.

Tabel 1- 3 tabel casefolding data

|  |  |  |
| --- | --- | --- |
| Data | Komentar | *casefolding* |
| Data elektronik  (handphone) | Produk original, pengiriman lumayan cepet. Belum di coba.. Tp looks ok nih.. Recommended seller. | produk original, pengiriman lumayan cepet. belum di coba.. tp looks ok nih.. recommended seller. |
| Data elektronik  (laptop) | kualitas barang bermasalah | kualitas barang bermasalah |
| Data pakaian  (kemeja) | Nyaman model bagus jahitan rapi ukuran pas | nyaman model bagus jahitan rapi ukuran pas |
| Data pakaian  (kaos) | Barang tidak sesuai dengan yg z pesan pertama n kemarin z pesan warna navy pas dengan badan z kainx jga bagus terus z pesan lagi maroon ndh sama panjang dengan warna navy... | barang tidak sesuai dengan yg z pesan pertama n kemarin z pesan warna navy pas dengan badan z kainx jga bagus terus z pesan lagi maroon ndh sama panjang dengan warna navy... |

1. *Punctuation Removal*

Merupakan tahap menghapus tanda baca pada data teks (Merinda Lestandy et al., 2021). Berikut *punctuation Removal*

yang dilakukan pada data penelitian:

Tabel 1- 4 tabel punctuation removal

|  |  |  |
| --- | --- | --- |
| Data | *cassefolding* | *Punctuation Removal* |
| Data elektronik  (*handphone)* | produk original, pengiriman lumayan cepet. belum di coba.. tp looks ok nih.. | produk original pengiriman lumayan cepet belum di coba tp looks ok nih recommended seller |
| Data elektronik  (laptop) | kualitas barang bermasalah | kualitas barang bermasalah |
| Data pakaian  (kemeja) | nyaman model bagus jahitan rapi ukuran pas | nyaman model bagus jahitan rapi ukuran pas |
| Data pakaian  (kaos) | barang tidak sesuai dengan yg z pesan pertama n kemarin z pesan warna navy pas dengan badan z kainx jga bagus terus z pesan lagi maroon ndh sama panjang dengan warna navy... | barang tidak sesuai dengan yg z pesan pertama n kemarin z pesan warna navy pas dengan badan z kainx jga bagus terus z pesan lagi maroon ndh sama panjang dengan warna navy |

1. *Removal stopwords*

Merupakan tahap mengolah data teks untuk menghapus kata hubung seperti kata “atau”, ”dan” karena tersebut merupakan kata yang sering muncul dan tidak memiliki arti apapun (Pradana & Hayaty, 2019) (Deviyanto & Wahyudi, 2018). Berikut *removal stopwords* yang dilakukan pada data penelitian:

Tabel 1- 5 tabel removal stopwords

|  |  |  |
| --- | --- | --- |
| Data | *Punctuation Removal* | *Stopword Removal* |
| Data elektronik  (*handphone)* | produk original pengiriman lumayan cepet belum di coba tp looks ok nih recommended seller | produk original pengiriman lumayan cepet coba tp looks ok nih recommended seller |
| Data elektronik  (laptop) | kualitas barang bermasalah | kualitas barang bermasalah |
| Data pakaian  (kemeja) | nyaman model bagus jahitan rapi ukuran pas | nyaman model bagus jahitan rapi ukuran pas |
| Data pakaian  (kaos) | barang tidak sesuai dengan yg z pesan pertama n kemarin z pesan warna navy pas dengan badan z kainx jga bagus terus z pesan lagi maroon ndh sama panjang dengan warna navy | barang sesuai z pesan n kemarin z pesan warna navy pas badan z kainx jga bagus z pesan maroon ndh warna navy |

1. *Stemming*

Merupakan tahap untuk mengurangi prefiks sebuah kata menjadi kata dasar (Pradana & Hayaty, 2019). Berikut *stemming* yang dilakukan pada data penelitian.

Tabel 1- 6 stemming

|  |  |  |
| --- | --- | --- |
| Data | *Stopword Removal* | *stemming* |
| Data elektronik  (*handphone)* | produk original pengiriman lumayan cepet coba tp looks ok nih recommended seller | produk original kirim lumayan cepet coba tp looks ok nih recommended seller |
| Data elektronik  (laptop) | kualitas barang bermasalah | kualitas barang masalah |
| Data pakaian  (kemeja) | nyaman model bagus jahitan rapi ukuran pas | nyaman model bagus jahit rapi ukur pas |
| Data pakaian  (kaos) | barang sesuai z pesan n kemarin z pesan warna navy pas badan z kainx jga bagus z pesan maroon ndh warna navy | barang sesuai z pesan n kemarin z pesan warna navy pas badan z kainx jga bagus z pesan maroon ndh warna navy |

1. Hasil text *preprcessing*

Hasil *text preprocessing* terdiri dari *casefolding, stopword removal, punctuation removal,* dan *stemming* dapat dilihat pada tabel 1-7.

Tabel 1- 7 hasil text preprocessing

|  |  |  |
| --- | --- | --- |
| Data | Komentar bersih | Label |
| Data pakaian (kemeja) | nyaman model bagus jahit rapi ukur pas | positif |
| Data pakaian  (kaos) | barang sesuai z pesan n kemarin z pesan warna navy pas badan z kainx jga bagus z pesan maroon ndh warna navy | negatif |
| Data elektronik  (*handphone)* | produk original kirim lumayan cepet coba tp looks ok nih recommended seller | positif |
| Data elektronik  (laptop) | kualitas barang masalah | negatif |
| Data elektronik  (*handphone)* | gambar c dtg a ya | netral |

*Worcloud* adalah mencakup daftar kata yang terkait dengan klasifikasi sentimen negatif dan positif atau kata yang sering muncul (Rahman et al., 2022). Berikut wordcloud pada data penelitian sebagai berikut:

Tabel 1- 8 tabel wordcloud

|  |  |
| --- | --- |
| *Wordcloud* | Kata-kata yang muncul |
| A picture containing text, font, graphics, logo  Description automatically generated | Kata-kata yang muncul pada data pakaian label positif:   * nyaman * model * bagus |
| A picture containing text, font, graphics, circle  Description automatically generated | Kata-kata yang muncul pada data elektronik label positif:   * produk * original * kirim |
| A picture containing font, text, graphics, logo  Description automatically generated | Kata-kata yang muncul pada data elektronik negatif:   * masalah * kualitas * barang |
| A picture containing font, graphics, text, logo  Description automatically generated | Kata-kata yang muncul pada data elektronik netral:   * ya * gambar * dtg |
|  | Kata-kata yang muncul data pakaian label negatif:   * pesan * warna * barang |

1. Pembobotan kata

Pada tahap ini setelah pengolahan data melakukan perhitungan kata dengan menggunakan metode *Term Frequency Inverse Document*, *Term Frequency Inverse Document* adalah metode perhitungan kata berdasarkan jumlah dokumen data dengan jumlah frekuensi kata yang muncul di setiap dokumen (Melita et al., 2018). Term Frequency Inverse Document mempunyai fungsi sebagai seleksi fitur untuk pemodelan machine learning klasifikasi (Prayoga et al., 2021). Berikut rumus *term frequency inverse document* pada tabel 1-9, dan perhitungan rumus *term Frequency Inverse Document* setiap kata pada tabel 1-10.

Tabel 1- 9 rumus term frequency inverse document

|  |  |
| --- | --- |
| Rumus *term frequency inverse document* | Rumus *inverse document frequecny* |
|  |  |

Tabel 1- 10 tabel term frequency inverse document

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Term* | tf | | | | df | n/df | idf | tfidf | | | |
| D1 | D2 | D3 | D4 | D1 | D2 | D3 | D4 |
| seller | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 | 0 | 0 | 0 | 0,6020599913279624 |
| jga | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0 | 0,6020599913279624 | 0 | 0 |
| cepet | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 | 0 | 0 | 0,6020599913279624 | 0 |
| barang | 0 | 0 | 0 | 1 | 1 | 4 | 0,6020599913279624 | 0 | 0 | 0 | 0,6020599913279624 |
| ndh | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0 | 0,6020599913279624 | 0 | 0 |
| looks | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| warna | 0 | 1 | 0 | 0 | 2 | 4 | 0,6020599913279624 |  | 1,2041199826559248 |  |  |
| original | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| recommended | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| coba | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| produk | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| kainx | 0 | 0 | 1 | 0 | 1 | 1 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| model | 1 | 0 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0,6020599913279624 |  |  |  |
| pesan | 0 | 3 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 1.806179973983887 |  |  |
| pas | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| tp | 0 | 0 | 0 | 1 | 1 | 4 | 0,6020599913279624 |  |  | 0.6020599913279624 |  |
| kemarin | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| maroon | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| badan | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| kirim | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| rapi | 1 | 0 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0.6020599913279624 |  |  |  |
| navy | 0 | 2 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 1,2041199826559248 |  |  |
| lumayan | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| ok | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| bagus | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| kualitas | 0 | 0 | 0 | 1 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| masalah | 0 | 0 | 0 | 1 | 1 | 4 | 0,6020599913279624 |  |  |  | 0,6020599913279624 |
| ukur | 0 | 0 | 0 | 1 | 1 | 4 | 0,6020599913279624 |  |  |  | 0,6020599913279624 |
| sesuai | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| nyaman | 1 | 0 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0,6020599913279624 |  |  |  |
| nih | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| jahit | 1 | 0 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0,6020599913279624 |  |  |  |

1. **Pemodelan**

Dilakukan pemodelan dengan menggunakan *supervised learning*. *Supervised learning* adalah pembelajaran dalam *machine learning* yang membutuhkan label untuk melakukan pelatihan (El Mohadab et al., 2019).

Model yang digunakan, yaitu sebagai berikut:

1. *Decision tree*

Pada pembelajaran mesin *decision tree* melakukan perhitungan *gini impurity* untuk mencari nilai atribut masing-masing fitur, sesudah melakukan perhitungan *impurity* masing-masing fitur maka hitung total nilai *impurity* kemudian mencari nilai terkecil dari masing-masing atribut untuk menjadi *root* (Jananto et al., 2021). Berikut rumus mencari nilai *impurity* dan total nilai i*mpurity*:

m: adalah jumlah kelas atau fitur

: adalah jumlah *record* yang ada di fitur

|  |
| --- |
|  |

Gambar 1- 1 rumus atribut gini impurity

|  |
| --- |
|  |

Gambar 1- 2 rumus total atribut gini impurity

Pada Gambar 1-1 (Jananto et al., 2021). Pemilihan atribut digunakan dengan mencari nilai *impurity* masing-masing fitur, sesudah dilakukan perhitungan *impurity* masing-masing fitur maka hitung total nilai *impurity* dengan rumus pada gambar 1-2 (Jananto et al., 2021).

Tahapan perhitungan *decision tree* untuk data *numerical* atau data angka diperlukan normalisasi, dengan cara sebagai berikut:

* Urutkan data angka dari terkecil sampai ke terbesar
* Hitung rata-rata terdekat

Tabel 1- 11 tabel bobot kata untuk decision tree

|  |  |  |
| --- | --- | --- |
| Kata | *Term frequency inverse document* | Label |
| nyaman | 0,6020599913279624 | positif |
| masalah | 0,6020599913279624 | negatif |
| kainx | 0,6020599913279624 | negatif |
| recommended | 0,6020599913279624 | positif |

Tabel 1- 12 perhitungan normaliasi

|  |  |
| --- | --- |
| Kata | Perhitungan |
| nyaman dan masalah |  |
|  |
|  |
| kainx dan *recommended* |  |
|  |
|  |

Tabel 1- 13 tabel hasil normalisasi

|  |  |  |  |
| --- | --- | --- | --- |
| Kata | *Term frequency inverse document* | normalisasi | Label |
| nyaman | 0,6020599913279624 | 0,6020599913279624 | positif |
| masalah | 0,6020599913279624 | negatif |
| kainx | 0,6020599913279624 | 0,6020599913279624 | negatif |
| recommended | 0,6020599913279624 | positif |

Perhitungan sudah normalisasi dilakukan, maka selanjutnya menghitung gini impurity dapat dilihat pada tabel 1-14.

Tabel 1- 14 perhitungan gini 0,6020599913279624

|  |  |  |
| --- | --- | --- |
| Probabilitas | label | Perhitungan |
| Probabilitas(gini)<0,6020599913279624 | 1. 0 positif 2. 0 netral 3. 0 negatif |  |
| Probabilitas(gini)>0,6020599913279624 | 1. 0 positif 2. 0 netral 3. 0 negatif |  |
| Total *gini impurity* | |  |

*Gini impurity term frequency inverse document* <0,6020599913279624 karena menghasilkan hasil terkecil, yaitu 0.

1. *Naïve Bayes*

Merupakan metode pembelajaran mesin untuk melakukan klasifiaksi sentimen analisis (Samsir et al., 2021). Dalam menghitung pembelajaran mesin *naives bayes* ada beberapa langkah sebagai berikut:

1. Menghitung kemungkinan sebelumnya (Farisi et al., 2019) .
2. Menghitung probabilitas masing-masing kelas data (Farisi et al., 2019).

Setelah penjelasan runtutan perhitungan pembelajaran mesin *naives bayes* berikut rumus perhitungan pembelajaran mesin *navies* sebagai berikut di gambar 1-3:

|  |
| --- |
|  |

Gambar 1- 3 rumus naives bayes

Pada gambar 1-3 (Pintoko & Lhaksmana, 2018) merupakan rumus *naives* dengan penjelasan sebagai berikut:

1. X = Merupakan data dengan kelas yang tidak terklasifikasi x (Pintoko & Lhaksmana, 2018)
2. H = Hipotesis khusus kelas (Pintoko & Lhaksmana, 2018)
3. P(X|H) = Berdasarkan kondisi H, kemungkinan hipotesis X dihitung (Pintoko & Lhaksmana, 2018).
4. P(X) = kemungkinan X (Pintoko & Lhaksmana, 2018).

Tabel 1- 15 tabel data untuk perhitungan navies bayes

|  |  |  |
| --- | --- | --- |
| Kata | *Term frequency inverse document* | Label |
| nyaman | 0,6020599913279624 | positif |
| masalah | 0,6020599913279624 | negatif |
| kainx | 0,6020599913279624 | negatif |
| *recommended* | 0,6020599913279624 | positif |

Berikut perhitungan navies bayes, dan perhitungan probabilitas setiap label.

Tabel 1- 16 komputasi tabel probabilitas setiap label

|  |  |
| --- | --- |
| Label | Perhitungan |
| positif |  |
| negatif |  |

Setelah melakukan perhitungan probabilitas setiap label, langkah selanjutnya menghitung kata terhadap label dapat dilihat pada tabel 1-17.

Tabel 1- 17 perhitungan kata terhadap label

|  |  |  |
| --- | --- | --- |
| Kata | Label | Perhitungan |
| nyaman | positif |  |
| negatif |  |
| recommended | Positif |  |
| negatif |  |
| kainx | positif |  |
| negatif |  |
| masalah | positif |  |
| negatif |  |

Tabel 1- 18 pengujian naives bayes

|  |  |
| --- | --- |
| Kalimat | label |
| kainx nyaman masalah | ? |

Pada tabel 1-18 melakukan pengujian *naives bayes* dengan kalimat “kainx nyaman masalah”, kemudian melakukan komputasi sebagai berikut.

Tabel 1- 19 tabel perhitungan pengujian menggunakan naives bayes

|  |  |
| --- | --- |
| Peluang label positif |  |
|  |
|  |
| Peluang label negatif |  |
|  |
|  |

Hasil uji didapatkan dari tabel 1-19 dengan kata “kainx nyaman malasah” mendapat label negatif.

1. *K-Nearest-Neighbor*

K-Nearest-Neighbor menurut penelitian Kang, Seokho adalah machine learning untuk prediksi berdasarkan label dari nilai k tetangga atau neighbor terdekat (Kang, 2021). Perhitugan *k-nearest neighbor* memerlukan beberapa tahap, sebagai berikut:

1. Menghitung tingkat kemiripan menggunakan *cosine similiarity* pada gambar 1.17 dari hasil *term frequecny iverse document* (Dwiki et al., 2021)*.*
2. Menentukan jarak yang akan diambil berdasarkan nilai K (Dwiki et al., 2021).

|  |
| --- |
|  |

Gambar 1- 4 rumus cosine similiarity

Pada penelitian ini mencari kueri “*recommend* pesan” menghasilkan kelas, menghitung *k-nearest neighbor* menghitung similiaritas menggunakan *cosine similiarity*.

Pada tabel 1-20 melakukan perhitungan kata setiap dokumen perhitungan *cosine* similiairty .

Tabel 1- 20 tabel perhitungan TF-IDF untuk cosine similiairty

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Term* | Wij\*wiq | | | | Wij^2 | | | | Wiq^2 |
| D1 | D2 | D3 | D4 | D1 | D2 | D3 | D4 | Q |
| nyaman | 0,6020599913279624 | 0 | 0 | 0 | 0,36247623315782616029223613581376 | 0 | 0 | 0 | 0 |
| masalah | 0 | 0 | 0 | 0,6020599913279624 | 0 | 0 | 0 | 0,36247623315782616029223613581376 | 0 |
| kainx | 0 | 0,6020599913279624 | 0 | 0 | 0 | 0,36247623315782616029223613581376 | 0 | 0 | 0 |
| *recommended* | 0 | 0 | 0,6020599913279624 | 0 | 0 | 0 | 0,36247623315782616029223613581376 | 0 | 0,36247623315782616029223613581376 |
| pesan | 0 | 0 | 1,806179973983887 | 0 | 0 | 0 | 3,26228609842043544263012522232384 | 0 | 3,26228609842043544263012522232384 |
| Total | 0,6020599913279624 | 0,6020599913279624 | 2,4082399653118496 | 0,6020599913279624 | 0,36247623315782616029223613581376 | 0,36247623315782616029223613581376 | 3,6247623315782616029223613581376 | 0,36247623315782616029223613581376 | 3,6247623315782616029223613581376 |

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

Gambar 1- 5 perhitungan cosine similiarity kata recommended.

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

Gambar 1- 6 perhitungan cosine similiarity kata pesan

Dari hasil perhitungan mencari *similiarity* maka urutkan dari terbesar ke terkecil:

1. Dokumen 3 (positif)
2. Dokumen 1 (positif)

Jika memakai nilai K untuk *k-neareast neighbor* adalah 2, maka sentimen untuk kalimat “*recommend* pesan” berada di label positif.

1. **Evaluasi**

Evaluasi dilakukan dengan, menggunakan 4 metode, anatara lain sebagai berikut:

1. Akurasi

Menghitung akurasi skor berdasarkan hasil prediksi dari data *testing*, dengan memperhatikan *true positive, true negative, false positive, false negative* (Romli et al., 2021). Berikut cara menghitung skor akurasi sebagai berikut:

|  |
| --- |
|  |

Gambar 1- 7 rumus akurasi

1. *Recall*

Merupakan perhitungan dari hasil prediksi menggunakan data uji untuk menghasilkan skor nilai salah perhitungan *recall* dilakukan sebagi berikut:(Pintoko & Lhaksmana, 2018) (Romli et al., 2021).

|  |
| --- |
|  |

Gambar 1- 8 rumus recall

1. *Precision*

Merupakan perhitungan dari hasil prediksi menggunakan data uji untuk mengukur prediksi nilai positif dari berapa banyak *true positive dengan false positive dilakukan* sebagai berikut: (Yun, 2021) (Romli et al., 2021).

|  |
| --- |
|  |

Gambar 1- 9 rumus precision

1. *Confusion matrix*

Merupakan hasil dari evaluasi pemodelan *machine learning* yang berbentuk kotak, terdapat 2 kolom dan 2 baris yang didalamnya ada *false negative, true negative, true negative, false positive.* Berikut merupakan contoh *confusion matrix:* (Yun, 2021)*.*

|  |
| --- |
|  |

Gambar 1- 10 confusion matrix

## Sistematika Penulisan

# BAB II

# TINJAUAN PUSTAKA

## 2.1. Tokopedia

Tokopedia adalah ­*e-commerce* dengan pengguna terbanyak berjumlah 153,46 juta (Handayani, 2021). Tokopedia didalamnya ada berbagai macam produk yang dijual mulai dari elektronik, pakaian, kosmetik. Oleh karena itu dengan jumlah pengguna yang banyak, serta menjual berbagai macam produk, Tokopedia memberikan fitur untuk memberikan pengalaman atau opini kepada konsumen yang sudah membeli barang di Tokopedia, didalam fitur tersebut ada berbagai macam penilaian dari konsumen yang sudah membeli ada yang penilaian secara positif, penilaian secara negatif, penilaian secara positif (Apriani et al., 2019b).

## 2.2. Analisis Sentimen

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).

Sentimen analisis melakukan pengelompokkan atau pelabelan dari sentimen yang ada di teks apakah sentimen tersebut bernilai positif atau negatif (Zamzami et al., 2021). Menurut Mayur Wankhade sentimen analisis terdapat beberapa level, yaitu *aspect level, phrase level, sentence level, document level* (Wankhade et al., 2022)*.*

## 2.3. Text mining

*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).

## 2.4. Text preprocessing

*Text preprocessing* adalah tahap persiapan agar data dapat dilakukan pemodelan (Cahyaningtyas et al., 2021). Penelitian Firdaus dan peneltian Filcha menjelaskan *Text preprocessing* merupakan 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* sebagai berikut:

1. *Labelling*

Tahap pelabelan berdasarkan rating, karena dilakukan untuk untuk melatih algoritma pada kumpulan data berlabel dalam proyek penambangan data untuk membuat klasifikasi atau prediksi(Demircan et al., 2021). Pada tahap pelabelan data berdasarkan rating, menurut penelitian Elmurngi (Elmurngi & Gherbi, 2018) terdapat pembagian 3 kategori sentimen, sebagai berikut:

• Bintang 1-2 diberikan label negatif

• Bintang 3 diberikan label netral

• Bintang 4-5 diberikan label positif

1. *Case Folding*

Tahap *case folding* adalah transformasi data teks yang mempunyai huruf kapital menjadi huruf kecil (Pravina et al., 2019).

1. *Punctuation Removal*

Merupakan tahap menghapus tanda baca di data teks, seperti (.) (,) (?), dan (angka) (Dyo fatra et al., 2020).

1. *Removal stopwords*

*Removal stopwords* menurut penelitian Wasim Bourequat merupakan teknik menghilangkan kata yang tidak berarti (Bourequat & Mourad, 2021). Contoh kata hubung:

“dan” “atau”

1. *Stemming*

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

“menyapu” -> sapu

1. Pembobotan kata

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)*.*

## *Worldcloud*

Merupakan salah satu dari banyak cara untuk menampilkan data teks secara visual (Ainur Rahman & Suroyo, 2021). *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). Berikut pada gambar 2.1 adalah contoh *wordlcoud*:

Gambar 2. 1 wordclooud

data recommned cepet kirim

## Pemodelan

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).

Perhitungan *decision tree* dengan cara sebagai berikut:

|  |
| --- |
|  |

Gambar 2. 2 rumus gini impurity

|  |
| --- |
|  |

Gambar 2. 3 rumus total gini impurity

Pada gambar 2.2 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).

1. Naives bayes

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

|  |
| --- |
|  |

Gambar 2. 4 rumus naives bayes

Pada gambar 2.4 (Salsabila et al., 2022) merupakan rumus naives dengan penjelasan sebagai berikut:

1. X = variabel acak (Salsabila et al., 2022)
2. H = kelas atau label (Salsabila et al., 2022)
3. P(X|H) = Berdasarkan X, probabilitas C dihitung (Salsabila et al., 2022).
4. P(X) = kemungkinan X (Salsabila et al., 2022).

*Naives bayes* memperrlukan perhitungan setiap kelas, rumus untuk mencari kemungkinan setiap kelas sebagai berikut:

|  |
| --- |
|  |

Gambar 2. 5 rumus perhitungan setiap kelas

Pada gambar 2.5 (Salsabila et al., 2022) merupakan rumus *naives* mencari probalitas setiap kelas dengan penjelasan sebagai berikut:

1. P(x) = kemungkinan X (Salsabila et al., 2022).
2. Nx = Jumlah dari kelas x (Salsabila et al., 2022).
3. N = Jumlah gabungan semua kelas (Salsabila et al., 2022).
4. *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* pada gambar 2.6 dari hasil term *frequecny imverse document* (Dwiki et al., 2021).
3. *Sorting* data dari ukuran terbesar ke terkecil *dari* hasil komputasi *cosine similiarity*(Dwiki et al., 2021)
4. Menentukan jarak yang akan diambil berdasarkan nilai K(Dwiki et al., 2021).

|  |
| --- |
|  |

Gambar 2. 6 rumus cosine similiarity

## Evaluasi

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*

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

1. *Precision*

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

1. *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)*.*

1. *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)*.*

# BAB III

# METODE PENELITIAN

## Data penelitian

Data yang diambil dari opini pelanggan Tokopedia di *website* Tokopedia pada bagian ulasan dari pelanggan yang membeli produk elektronik (laptop *handphone*), produk pakaian (kaos, kemeja) menggunakan menghasilkan 893 data kategori elektronik (laptop *handphone*), berikut data kategori elektronik dapat dilihat pada tabel 3-1.

Tabel 3- 1 data elektronik

|  |  |
| --- | --- |
| Grafik data | Keterangan |
| A picture containing text, screenshot, plot, diagram  Description automatically generated | 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 |
| A picture containing text, screenshot, plot, diagram  Description automatically generated | rating 1 berjumlah 110 data  rating 2 berjumlah 32 data  rating 3 berjumlah 89 data  rating 4 berjumlah 123 data  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 pada tabel 3-2:

Tabel 3- 2 data pakaian

|  |  |
| --- | --- |
| Grafik data | Keterangan |
| A picture containing text, screenshot, diagram, plot  Description automatically generated | 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 |
| A picture containing text, screenshot, plot, diagram  Description automatically generated | 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 preprocessinng

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

### Pelabelan

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 dapat dilihat pada tabel 3-3 dan tabel 3-4.

Tabel 3- 3 hasil pelabelan data elektronik

|  |  |
| --- | --- |
| Grafik data | Keterangan |
|  | 1. Total data laptop label positif 235 data 2. Total data laptop label netral 74 data 3. Total data laptop label negatif 90 data |
| A picture containing text, screenshot, rectangle, diagram  Description automatically generated | 1. Total data *handphone* label positif 262 data 2. Total data *handphone* label netral 89 data 3. Total data *handphone* label negatif 142 data |

Tabel 3- 4 hasil pelabelan data pakaian

|  |  |
| --- | --- |
| Grafik data | Keterangan |
| A blue bar graph with white text  Description automatically generated | 1. Total data kaos label positif 495 data 2. Total data kaos label netral 190 data 3. Total data kaos label negatif 245 data |
| A picture containing text, screenshot, diagram, rectangle  Description automatically generated | 1. Total data kaos label positif 345 data 2. Total data kaos label netral 145 data 3. Total data kaos label negatif 146 data |

Pada tahap sesudah melakukan pelabelan maka melakukan penggabungan data penelitian, menjadi 2 kategori, yaitu data elektronik terdiri dari data laptop, data *handphone*, data pakaian terdiri dari data pakaian, data kaos, berikut hasil penggabungan dan total label yang dihasilkan dapat dilihat pada tabel 3-5.

Tabel 3- 5 hasil pelabelan penggabungan data menjadi 2 kategori data

|  |  |
| --- | --- |
| Grafik data | Keterangan |
|  | Total data elektronik:   * 1. Total label positif adalah 500 data.   2. Total label negatif adalah 232 data.   3. Total label netral adalah 163 data. |
|  | Total data pakaian:   * 1. Total label positif adalah 849 data.   2. Total label negatif adalah 391 data.   3. Total label netral adalah 335 data. |

### *Casefolding*

*Casefolding* merupakan tahap untuk transformasi data teks menjadi huruf kecil. Berikut beberapa data hasil *casefolding* dari data elektronik dan data pakaian.

Tabel 3- 6 empat data penelitian

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Komentar (ulasan) | Rating | Label |
| Data elektronik | terkecoh banget sama variannya ternyata yang di klik 9a. bukan 9c. semoga awet deh ya hpnya. thx seller | 4 | positif |
| Data elektronik | terima kasih gan barang sudah mendarat dengan selamat 👍🏻👍🏻 ada beberapa dent yang seperti nya bekas jatuh dan juga ada keyboard yang coak seperti kena rokok. | 4 | positif |
| Data pakaian | Kiriman cepat sampai. Bahan kain agak tebal. Ukurannya kurang lebar dikit, berasa bukan 52. Thanks | 3 | netral |
| Data  pakaian | pengiriman lama, pesanan tidak sesuai dgn apa yg dipesan. kecewa sama barang yg dtng tdk sesuai pdhal produk terkenal. | 2 | negatif |

Pada tabel 3-6 merupakan empat data penelitian yang sudah dilakukan pelabelan, kemudian dilakukan tahap *casefolding, punctuation removal, stemming, wordcloud,* dan pembobotan kata, menuju ke tahap pemodelan.

Tabel 3- 7 casefolding 5 data elektronik dan data pakaian

|  |  |  |
| --- | --- | --- |
| Data | Komentar (ulasan) | *Casefolding* |
| Data elektronik | terkecoh banget sama variannya ternyata yang di klik 9a. bukan 9c. semoga awet deh ya hpnya. thx seller | terkecoh banget sama variannya ternyata yang di klik 9a. bukan 9c. semoga awet deh ya hpnya. thx seller |
| Data elektronik | terima kasih gan barang sudah mendarat dengan selamat 👍🏻👍🏻 ada beberapa dent yang seperti nya bekas jatuh dan juga ada keyboard yang coak seperti kena rokok. | terima kasih gan barang sudah mendarat dengan selamat 👍🏻👍🏻 ada beberapa dent yang seperti nya bekas jatuh dan juga ada keyboard yang coak seperti kena rokok. |
| Data pakaian | Kiriman cepat sampai. Bahan kain agak tebal. Ukurannya kurang lebar dikit, berasa bukan 52. Thanks | kiriman cepat sampai. bahan kain agak tebal. ukurannya kurang lebar dikit, berasa bukan 52. thanks |
| Data  pakaian | pengiriman lama, pesanan tidak sesuai dgn apa yg dipesan. kecewa sama barang yg dtng tdk sesuai pdhal produk terkenal. | pengiriman lama, pesanan tidak sesuai dgn apa yg dipesan. kecewa sama barang yg dtng tdk sesuai pdhal produk terkenal. |

### *Punctuation removal*

*Punctuation removal* merupakan tahapan untuk menghapus tanda baca dan nomor karena agar tidak memperbanyak bobot kata pada tahap pembobotan kata, berikut beberapa data yang dilakukan *punctuation removal*:

Tabel 3- 8 punctuation removal

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Komentar (ulasan) | *Casefolding* | *Punctuation removal* |
| Data elektronik | terkecoh banget sama variannya ternyata yang di klik 9a. bukan 9c. semoga awet deh ya hpnya. thx seller | terkecoh banget sama variannya ternyata yang di klik 9a. bukan 9c. semoga awet deh ya hpnya. thx seller | terkecoh banget sama variannya ternyata yang di klik a bukan c semoga awet deh ya hpnya thx seller |
| Data elektronik | terima kasih gan barang sudah mendarat dengan selamat 👍🏻👍🏻 ada beberapa dent yang seperti nya bekas jatuh dan juga ada keyboard yang coak seperti kena rokok. | terima kasih gan barang sudah mendarat dengan selamat 👍🏻👍🏻 ada beberapa dent yang seperti nya bekas jatuh dan juga ada keyboard yang coak seperti kena rokok. | terima kasih gan barang sudah mendarat dengan selamat ada beberapa dent yang seperti nya bekas jatuh dan juga ada keyboard yang coak seperti kena rokok |
| Data pakaian | Kiriman cepat sampai. Bahan kain agak tebal. Ukurannya kurang lebar dikit, berasa bukan 52. Thanks | kiriman cepat sampai. bahan kain agak tebal. ukurannya kurang lebar dikit, berasa bukan 52. thanks | kiriman cepat sampai bahan kain agak tebal ukurannya kurang lebar dikit berasa bukan thanks |
| Data  pakaian | pengiriman lama, pesanan tidak sesuai dgn apa yg dipesan. kecewa sama barang yg dtng tdk sesuai pdhal produk terkenal. | pengiriman lama, pesanan tidak sesuai dgn apa yg dipesan. kecewa sama barang yg dtng tdk sesuai pdhal produk terkenal. | pengiriman lama pesanan tidak sesuai dgn apa yg dipesan kecewa sama barang yg dtng tdk sesuai pdhal produk terkenal |

### *Stopwords removal*

*Stopwords removal* merupakan tahapan untuk menghilangkan kata hubung, berikut merupakan data elektronik, data pakaian yang dilakukan *stopwords removal:*

Tabel 3- 9 tabel stopword removal 4 data elektronik dan data pakaian

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data | Komentar (ulasan) | *Casefolding* | *Punctuation removal* | *Stopwords removal* |
| Data elektronik | terkecoh banget sama variannya ternyata yang di klik 9a. bukan 9c. semoga awet deh ya hpnya. thx seller | terkecoh banget sama variannya ternyata yang di klik 9a. bukan 9c. semoga awet deh ya hpnya. thx seller | terkecoh banget sama variannya ternyata yang di klik a bukan c semoga awet deh ya hpnya thx seller | terkecoh banget variannya klik a c semoga awet deh ya hpnya thx seller |
| Data elektronik | terima kasih gan barang sudah mendarat dengan selamat 👍🏻👍🏻 ada beberapa dent yang seperti nya bekas jatuh dan juga ada keyboard yang coak seperti kena rokok. | terima kasih gan barang sudah mendarat dengan selamat 👍🏻👍🏻 ada beberapa dent yang seperti nya bekas jatuh dan juga ada keyboard yang coak seperti kena rokok. | terima kasih gan barang sudah mendarat dengan selamat ada beberapa dent yang seperti nya bekas jatuh dan juga ada keyboard yang coak seperti kena rokok | terima kasih gan barang mendarat selamat dent nya bekas jatuh keyboard coak kena rokok |
| Data pakaian | Kiriman cepat sampai. Bahan kain agak tebal. Ukurannya kurang lebar dikit, berasa bukan 52. Thanks | kiriman cepat sampai. bahan kain agak tebal. ukurannya kurang lebar dikit, berasa bukan 52. thanks | kiriman cepat sampai bahan kain agak tebal ukurannya kurang lebar dikit berasa bukan thanks | kiriman cepat bahan kain tebal ukurannya lebar dikit berasa thanks |
| Data  pakaian | pengiriman lama, pesanan tidak sesuai dgn apa yg dipesan. kecewa sama barang yg dtng tdk sesuai pdhal produk terkenal. | pengiriman lama, pesanan tidak sesuai dgn apa yg dipesan. kecewa sama barang yg dtng tdk sesuai pdhal produk terkenal. | pengiriman lama pesanan tidak sesuai dgn apa yg dipesan kecewa sama barang yg dtng tdk sesuai pdhal produk terkenal | pengiriman pesanan sesuai dgn dipesan kecewa barang dtng tdk sesuai pdhal produk terkenal |

### *Stemming*

*Stemming* merupakan tahapan transformasi teks data kata menjadi ke bentuk dasar, berikut merupakan beberapa data elektronik, data pakaian yang dilakukan *stemming*:

Tabel 3- 10 tabel stemming 4 data elektronik dan data pakaian.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data | Komentar (ulasan) | *Casefolding* | *Punctuation removal* | *Stopwords removal* | *Stemming* |
| Data elektronik | terkecoh banget sama variannya ternyata yang di klik 9a. bukan 9c. semoga awet deh ya hpnya. thx seller | terkecoh banget sama variannya ternyata yang di klik 9a. bukan 9c. semoga awet deh ya hpnya. thx seller | terkecoh banget sama variannya ternyata yang di klik a bukan c semoga awet deh ya hpnya thx seller | terkecoh banget variannya klik a c semoga awet deh ya hpnya thx seller | kecoh banget varian klik a c moga awet deh ya hpnya thx seller |
| Data elektronik | terima kasih gan barang sudah mendarat dengan selamat 👍🏻👍🏻 ada beberapa dent yang seperti nya bekas jatuh dan juga ada keyboard yang coak seperti kena rokok. | terima kasih gan barang sudah mendarat dengan selamat 👍🏻👍🏻 ada beberapa dent yang seperti nya bekas jatuh dan juga ada keyboard yang coak seperti kena rokok. | terima kasih gan barang sudah mendarat dengan selamat ada beberapa dent yang seperti nya bekas jatuh dan juga ada keyboard yang coak seperti kena rokok | terima kasih gan barang mendarat selamat dent nya bekas jatuh keyboard coak kena rokok | terima kasih gan barang darat selamat dent nya bekas jatuh keyboard coak kena rokok |
| Data pakaian | Kiriman cepat sampai. Bahan kain agak tebal. Ukurannya kurang lebar dikit, berasa bukan 52. Thanks | kiriman cepat sampai. bahan kain agak tebal. ukurannya kurang lebar dikit, berasa bukan 52. thanks | kiriman cepat sampai bahan kain agak tebal ukurannya kurang lebar dikit berasa bukan thanks | kiriman cepat bahan kain tebal ukurannya lebar dikit berasa thanks | kirim cepat bahan kain tebal ukur lebar dikit asa thanks |
| Data  pakaian | pengiriman lama, pesanan tidak sesuai dgn apa yg dipesan. kecewa sama barang yg dtng tdk sesuai pdhal produk terkenal. | pengiriman lama, pesanan tidak sesuai dgn apa yg dipesan. kecewa sama barang yg dtng tdk sesuai pdhal produk terkenal. | pengiriman lama pesanan tidak sesuai dgn apa yg dipesan kecewa sama barang yg dtng tdk sesuai pdhal produk terkenal | pengiriman pesanan sesuai dgn dipesan kecewa barang dtng tdk sesuai pdhal produk terkenal | kirim pesan sesuai dgn pes kecewa barang dtng tdk sesuai pdhal produk kenal |

### *Wordcloud*

*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:

Tabel 3- 11 tabel worldcloud

|  |  |
| --- | --- |
| *Wordcloud* | Kata-kata yang muncul |
|  | Kata-kata yang muncul di data elektronik label positif:   * klik * banget * awet * Kecoh * deh * varian |
|  | Kata-kata yang muncul di data pakaian label netral:   * cepat * kirim * kain |
|  | Kata-kata yang muncul di data pakaian label negatif:   * sesuai * pdhal * barang * kecewa |

### Pembobotan kata

Pada tahap ini melakukan pembobotan kata dilakukan setelah *casefolding, punctuation removal, stopwords removal, stemming.* Pembobotan kata dilakukan untuk pemodelan *machine learning,* 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:

n = jumlah data

df = Total kemunculan frekuensi kata di setiap dokumen

tf = kemunculan frekuensi kata di setiap dokumen

Tabel 3- 12 rumus perhitungan term frequency inverse document

|  |
| --- |
|  |
|  |

Berikut merupakan perhitungan *term frequency inverse document* kata “sesuai”:

Tabel 3- 13 perhitungan term frequency inverse document kata “sesuai”

|  |  |  |
| --- | --- | --- |
| kata | Idf | *Term frequency inverse document* |
| sesuai |  | 1,2041199826559248 |

Tabel 3- 14 tabel pembobotan kata “sesuai” setiap dokumen

|  |
| --- |
| 1,2041199826559248 |
| 0 |
| 0 |
| 0 |

Tabel 3- 15 tabel pembobotan kata

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Term* | tf | | | | df | n/df | idf | tfidf | | | |
| D1 | D2 | D3 | D4 | D1 | D2 | D3 | D4 |
| sesuai | 0 | 0 | 0 | 2 | 1 | 4 | 0,6020599913279624 | 0 | 0 | 0 | 1,2041199826559248 |
| klik | 1 | 0 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0,6020599913279624 | 0 | 0 | 0 |
| deh | 1 | 0 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0,6020599913279624 | 0 | 0 | 0 |
| banget | 1 | 0 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0,6020599913279624 | 0 | 0 | 0 |
| kirim | 0 | 0 | 1 | 1 | 2 | 2 | 0,3010299956639812 |  |  | 0,3010299956639812 | 0,3010299956639812 |
| kain | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| terima | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| bahan | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| ya | 1 | 0 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0,6020599913279624 |  |  |  |
| dikit | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| varian | 1 | 0 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0,6020599913279624 |  |  |  |
| barang | 0 | 1 | 0 | 1 | 2 | 2 | 0,3010299956639812 |  | 0,3010299956639812 |  | 0,3010299956639812 |
| dgn | 0 | 0 | 0 | 1 | 1 | 4 | 0,6020599913279624 |  |  |  | 0,6020599913279624 |
| darat | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| lebar | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| kecewa | 0 | 0 | 0 | 1 | 1 | 4 | 0,6020599913279624 |  |  | 0.6020599913279624 |  |
| dent | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| keyboard | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| tebal | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| hpnya | 1 | 0 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0,6020599913279624 |  |  |  |
| seller | 1 | 0 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0.6020599913279624 |  |  |  |
| jatuh | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| kasih | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| gan | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| awet | 1 | 0 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0,6020599913279624 |  |  |  |
| kenal | 0 | 0 | 0 | 1 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| thx | 1 | 0 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0,6020599913279624 |  |  |  |
| produk | 0 | 0 | 0 | 1 | 1 | 4 | 0,6020599913279624 |  |  |  | 0,6020599913279624 |
| cepat | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| coak | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0,6020599913279624 |  |  |  |
| tdk | 0 | 0 | 0 | 1 | 1 | 4 | 0,6020599913279624 |  |  |  | 0,6020599913279624 |
| thanks | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| asa | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| rokok | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| kecoh | 1 | 0 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0,6020599913279624 |  |  |  |
| selamat | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| pesan | 0 | 0 | 0 | 1 | 1 | 4 | 0,6020599913279624 |  |  |  | 0,6020599913279624 |
| kena | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| nya | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| ukur | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| moga | 1 | 0 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0,6020599913279624 |  |  |  |
| bekas | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| dtng | 0 | 0 | 0 | 1 | 1 | 4 | 0,6020599913279624 |  |  |  | 0,6020599913279624 |
| pdhal | 0 | 0 | 0 | 1 | 1 | 4 | 0,6020599913279624 |  |  |  | 0,6020599913279624 |
| pes | 0 | 0 | 0 | 1 | 1 | 4 | 0,6020599913279624 |  |  |  | 0,6020599913279624 |

## Pemodelan

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

### Naives bayes

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

Tabel 3- 16 tabel kata dan label untuk naives bayes

|  |  |  |
| --- | --- | --- |
| Kata | *Term frequency inverse document* | Label |
| lebar | 0,6020599913279624 | netral |
| awet | 0,6020599913279624 | positif |
| selamat | 0,6020599913279624 | positif |
| sesuai | 1,2041199826559248 | negatif |

Berikut perhitungan *navies bayes*, dan perhitungan probabilitas setiap label.

Tabel 3- 17 komputasi tabel probabilitas setiap label

|  |  |
| --- | --- |
| Label | Perhitungan |
| positif |  |
| negatif |  |
| netral |  |

Setelah melakukan perhitungan probabilitas setiap label, langkah selanjutnya menghitung kata terhadap label dapat dilihat pada tabel 3-18.

Tabel 3- 18 perhitungan kata terhadap label

|  |  |  |
| --- | --- | --- |
| Kata | Label | Perhitungan |
| lebar | positif |  |
| negatif |  |
| netral |  |
| awet | Positif |  |
| negatif |  |
| netral |  |
| selamat | positif |  |
|  | negatif |  |
|  | netral |  |
| sesuai | positif |  |
|  | negatif |  |
|  | netral |  |

Sesudah menghitung hasil kata setiap label, maka melakukan pengujian, sebagai berikut.

Tabel 3- 19 pengujian naives bayes

|  |  |
| --- | --- |
| Kalimat | label |
| lebar sesuai | ? |

Pada tabel 3-20 melakukan pengujian *naives bayes* dengan kalimat “lebar sesuai”, kemudian melakukan komputasi sebagai berikut.

Tabel 3- 20 tabel perhitungan pengujian menggunakan naives bayes

|  |  |
| --- | --- |
| Peluang label positif |  |
|  |
|  |
| Peluang label negatif |  |
|  |
|  |
| Peluang label netral |  |
|  |
|  |

Pada tabel 3-17 perhitungan dari pengujian *naives*. Hasil probbilitas yang sudah didapatkan klasifikasi dari “lebar sesuai” berlabel negatif, karena probalitias negatif lebih besar daripada label netral maupun label positif. Hasil uji didapatkan “lebar sesuai” mendapat label negatif.

### Decision tree

Pada tahap pembelajaran mesin *decision tree* atau pohon keputusan, dilakukan perhitungan menggunakan hasil dari pembobotan kata pada tabel 3-13 sebagai berikut:

Tabel 3- 21 tabel kata dan label untuk decision tree

|  |  |  |
| --- | --- | --- |
| Kata | *Term frequency inverse document* | Label |
| lebar | 0,6020599913279624 | netral |
| awet | 0,6020599913279624 | positif |
| selamat | 0,6020599913279624 | positif |
| sesuai | 1,2041199826559248 | negatif |

Karena *term frequency inverse document* adalah tipe data numerikal maka dilakukan normalisasi dengan cara mencari rata-rata, perhitungan rata-rata dapat dilihat pada tabel 3-22, serta hasil normalisasi dapat dilihat pada tabel 3-23:

Tabel 3- 22 perhitungan rata-rata

|  |  |
| --- | --- |
| Kata-kata | perhitungan |
| lebar dan awet |  |
|  |
|  |
| selamat dan sesuai |  |
|  |
|  |

Tabel 3- 23 normalisasi

|  |  |  |  |
| --- | --- | --- | --- |
| Kata | *Term frequency inverse document* | normalisasi | Label |
| lebar | 0,6020599913279624 | 0,6020599913279624 | netral |
| awet | 0,6020599913279624 | positif |
| selamat | 0,6020599913279624 | 0,9030900000000001 | positif |
| sesuai | 1,2041199826559248 | negatif |

Perhitungan normalisasi dilakukan, maka selanjutnya menghitung gini impurity dapat dilihat pada tabel 3-24 dan tabel 3-25:

Tabel 3- 24 perhitungan gini 0,6020599913279624

|  |  |  |
| --- | --- | --- |
| Probabilitas | label | Perhitungan |
| Probabilitas(gini)<0,6020599913279624 | 1. 0 positif 2. 0 netral 3. 0 negatif |  |
| Probabilitas(gini)>0,6020599913279624 | 1. 0 positif 2. 0 netral 3. 0 negatif |  |
| Total *gini impurity* | |  |

Tabel 3- 25 perhitungan gini 0,9030900000000001

|  |  |  |
| --- | --- | --- |
| Probabilitas | label | Perhitungan |
| Probabilitas(gini)<0,9030900000000001 | 1. 2 positif 2. 1 netral 3. 0 negatif |  |
| Probabilitas(gini)>0,9030900000000001 | 1. 0 positif 2. 0 netral 3. 1 negatif |  |
| Total *gini impurity* | |  |

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

### *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 3- 26 tabel bobot kata untuk k nearest neighbour

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Term* | tf | | | | df | n/df | idf | tfidf | | | |
| D1 | D2 | D3 | D4 | D1 | D2 | D3 | D4 |
| lebar | 0 | 0 | 1 | 0 | 1 | 4 | 0,6020599913279624 |  |  | 0,6020599913279624 |  |
| awet | 1 | 0 | 0 | 0 | 1 | 4 | 0,6020599913279624 | 0,6020599913279624 |  |  |  |
| selamat | 0 | 1 | 0 | 0 | 1 | 4 | 0,6020599913279624 |  | 0,6020599913279624 |  |  |
| sesuai | 0 | 0 | 0 | 2 | 1 | 4 | 0,6020599913279624 | 0 | 0 | 0 | 1,2041199826559248 |

Mencari kueri “lebar sesuai” menghasilkan kelas, menghitung similiaritas menggunakan *cosine similiarity*:

Tabel 3- 27 tabel perhitungan query

|  |
| --- |
|  |
|  |

|  |
| --- |
|  |

Gambar 3. 3 Rumus cosine similiarity

Tabel 3- 28 tabel bobot kata hasil kalkulasi jarak

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Term* | Wij\*wiq | | | | Wij^2 | | | | Wiq^2 |
| D1 | D2 | D3 | D4 | D1 | D2 | D3 | D4 | Q |
| lebar | 0 | 0 | 0,6020599913279624 | 0 |  |  | 0,36247623315782616029223613581376 |  | 0,36247623315782616029223613581376 |
| awet | 0 | 0 | 0 | 0 | 0,36247623315782616029223613581376 |  |  |  |  |
| selamat | 0 | 0 | 0 | 0 |  | 0,36247623315782616029223613581376 |  |  |  |
| sesuai | 0 | 0 | 0 | 1,44990493263130464116894454325504 | 0 | 0 | 0 | 1,44990493263130464116894454325504 | 1,44990493263130464116894454325504 |
| Total | 0 | 0 | 0,6020599913279624 | 1,44990493263130464116894454325504 | 0,36247623315782616029223613581376 | 0,36247623315782616029223613581376 | 0,36247623315782616029223613581376 | 1,44990493263130464116894454325504 | 1,8123811657891308014611806790688 |

Tabel 3- 29 perhitungan cosine similiarity kata lebar

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

Tabel 3- 30 perhitungan cosine similiarity kata lebar

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

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

Ambil nilai K = 1

Kesimpulan bahwa kueri “lebar sesuai” menghasilkan kelas negatif.

## 3.3 Evaluasi

Pada tahap ini melakukan evaluasi dari tahap pemodelan, berikut perhitungan evaluasi

### Evaluasi naives bayes

Evaluasi *naives bayes* menggunakan 30% dari data kategori elektronik, data kategori pakaian.

#### Evaluasi data pakaian

Evaluasi skenario naives bayes dengan data kategori elektronik hanya menggunakan 30% dari total 1575 data menjadi 472 data. Pada hasil evaluasi menghasilkan sebagai berikut.

Pada pengujian *naives bayes* menghasilkanhasil evaluasi dari perhitungan dengan data latih sebesar 330 data, dan data uji sebesar 142 data, kemudian hasil perhitungan *accucary, precision, recall* dapat dilihat pada tabel 3-31.:

Tabel 3- 31 evaluasi data pakaian

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Confusion matrix* | *Tp* | *Tn* | *Fp* | *Fn* |
|  | 33, 18, 6 | 18, 3, 19, 6,33, 6, 5, 6, 33, 45, 7, 18 | 45, 6, 7, 3,19, 5 | 7, 5, 45, 19,  3, 6 |
| *Total tp, tn, fp, fn* | 57 | 199 | 85 | 85 |
| *Hasil evaluasi* | | | | |
| *Accuracy* | 60,09% | | | |
| *Recall* | 40,14% | | | |
| *precision* | 40,14% | | | |

#### Evaluasi data elektronik

Evaluasi skenario naives bayes dengan data kategori elektronik hanya menggunakan 30% dari total 895 menjadi 265 data. Pada hasil evaluasi menghasilkan sebagai berikut.

Pada pengujian *naives bayes* menghasilkanhasil evaluasi dari perhitungan dengan data latih sebesar 187 data, dan data uji sebesar 81 data, kemudian hasil perhitungan *accucary, precision, recall* dapat dilihat pada tabel 3-32.

Tabel 3- 32 evaluasi data elektronik

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Confusion matrix* | *Tp* | *Tn* | *Fp* | *Fn* |
|  | 28, 7 , 4 | 11, 7, 2, 4  28, 4, 4, 6, 28, 7, 12, 7 | 4, 12, 2, 7, 11, 6 | 7,6 12, 11, 4, 2 |
| *Total tp, tn, fp, fn* | 39 | 120 | 42 | 42 |
| *Hasil evaluasi* | | | | |
| *Accuracy* | 51,304 % | | | |
| *Recall* | 55,554% | | | |
| *precision* | 25,423% | | | |

### Evaluasi decision tree

Evaluasi *naives bayes* menggunakan 30% dari data kategori elektronik, dan data kategori pakaian.

#### Evaluasi data pakaian

Evaluasi skenario *decision tree* dengan data kategori elektronik hanya menggunakan 30% dari total 1575 data menjadi 472 data. Pada hasil evaluasi menghasilkan sebagai berikut.

Pada pengujian *decision tree* menghasilkanhasil evaluasi dari perhitungan dengan data latih sebesar 330 data, dan data uji sebesar 142 data, kemudian hasil perhitungan *accucary, precision, recall* dapat dilihat pada tabel 3-33.

Tabel 3- 33 evaluasi data pakaian

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Confusion matrix* | *Tp* | *Tn* | *Fp* | *Fn* |
|  | 53, 8, 14 | 7, 8, 6, 14, 53, 17, 14, 9, 53, 14, 14, 14 | 17, 14,  6, 14, 7, 9 | 14, 9,  14, 7,  17, 6 |
| *Total tp, tn, fp, fn* | 75 | 223 | 67 | 67 |
| *Hasil evaluasi* | | | | |
| *Accuracy* | 68,98% | | | |
| *Recall* | 52,81% | | | |
| *precision* | 52,81% | | | |

#### Evaluasi data elektronik

Evaluasi skenario naives bayes dengan data kategori elektronik hanya menggunakan 30% dari total 895 menjadi 265 data. Pada hasil evaluasi menghasilkan sebagai berikut:

Pada pengujian *decision tree* menghasilkanhasil evaluasi dari perhitungan dengan data latih sebesar 187 data, dan data uji sebesar 81 data, kemudian hasil perhitungan *accucary, precision, recall* dapat dilihat pada tabel 3-34.

Tabel 3- 34 evaluasi data pakaian

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Confusion matrix* | *Tp* | *Tn* | *Fp* | *Fn* |
| A white square with black text  Description automatically generated | 30, 3, 12 | 7, 8, 6, 14, 53, 17, 14, 9, 53, 14, 14, 14 | 17, 14,  6, 14, 7, 9 | 14, 9,  14, 7,  17, 6 |
| *Total tp, tn, fp, fn* | 75 | 223 | 67 | 67 |
| *Hasil evaluasi* | | | | |
| *Accuracy* | 68,98% | | | |
| *Recall* | 52,81% | | | |
| *precision* | 52,81% | | | |

### Evaluasi *k nearest neighbor*

Evaluasi *k nearest neighbor* menggunakan 30% dari data kategori elektronik, dan data kategori pakaian.

#### Evaluasi data pakaian

Evaluasi skenario *k nearest neighbor* dengan data kategori elektronik hanya menggunakan 30% dari total 1575 data menjadi 472 data. Pada hasil evaluasi menghasilkan sebagai berikut:

Tabel 3- 35 evaluasi data pakaian

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Confusion matrix* | *Tp* | *Tn* | *Fp* | *Fn* |
| A white rectangular grid with black text  Description automatically generated | 63, 11, 9 | 10, 11, 6, 9, 63, 11, 9, 11, 63, 11, 10, 11 | 11, 10,  6, 11, 10, 11 | 11, 11,  10, 10,  11, 6 |
| *Total tp, tn, fp, fn* | 83 | 225 | 59 | 59 |
| *Hasil evaluasi* | | | | |
| *Accuracy* | 72,30% | | | |
| *Recall* | 58,45% | | | |
| *precision* | 58,45% | | | |

Pada tabel 3-35 merupakan hasil pengujian pengujian *k nearest neighbor*,hasil evaluasi dari perhitungan dengan data latih sebesar 330 data, dan data uji sebesar 142 data, kemudian hasil perhitungan *accucary, precision, recall* menggunakan *k nearest neighbor* dengan nilai K=1, menghasilkan akurasi 72,30%, *recall* 58,45%, *precsision* 58,45% *.*

#### Evaluasi data elektronik

Evaluasi skenario *k nearest neighbor* dengan data kategori elektronik hanya menggunakan 30% dari total 895 menjadi 265 data. Pada hasil evaluasi menghasilkan sebagai berikut:

Tabel 3- 36 evaluasi data elektronik

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Confusion matrix* | *Tp* | *Tn* | *Fp* | *Fn* |
|  | 36, 7, 9 | 1, 7, 3, 9, 36, 2, 9, 11, 36, 6, 6, 11 | 2, 6, 3, 6, 1, 11 | 6, 11,  6, 1,  2, 3 |
| *Total tp, tn, fp, fn* | 52 | 137 | 29 | 29 |
| *Hasil evaluasi* | | | | |
| *Accuracy* | 76,51% | | | |
| *Recall* | 64,19% | | | |
| *precision* | 64,19% | | | |

Pada tabel 3-36 merupakan hasil pengujian *k nearest neighbor* menghasilkanhasil evaluasi dari perhitungan dengan data latih sebesar 187 data, dan data uji sebesar 81 data, kemudian hasil perhitungan *accucary, precision, recall* menggunakan *k nearest neighbor* dengan nilai K=1, menghasilkan akurasi 76,51%, *recall* 64,19%, *precsision* 64,19%

# BAB IV

# HASIL DAN PEMBAHASAN

## Data Penelitian

### Data Pakaian

Pada penelitian ini menggunakan data pakaian sebanyak 1575 data yang telah dilabeli oleh penulis dan dilakukan *text preprocessing*, berikut hasil pelabelan yang dilakukan pada tabel 4-1:

Tabel 4- 1 data pakaian dengan label

|  |  |
| --- | --- |
| Label | Total |
| Positif | 849 |
| Netral | 391 |
| Negatif | 335 |

Tabel 4-1 merupakan hasil dari pelabelan yang dilakukan, kemudian *wordcloud* atau awan kata untuk melihat kata-kata yang sering muncul pada masing-masing label data pakaian. Berikut tabel 4-2 tampilan *wordcloud* masing-masing label.

Tabel 4- 2 tabel wordcloud data pakaian

|  |  |
| --- | --- |
| *wordcloud* | Kata-kata yang muncul |
|  | Kata-kata yang muncul di data pakaian label positif:   1. bahan 2. bagus 3. mantap |
|  | Kata-kata yang muncul di data pakaian label netral:   1. barang 2. bahan 3. kirim |
|  | Kata-kata yang muncul di data pakaian label netral:   1. warna 2. kirim 3. ukur |

### Data Elektronik

Pada penelitian ini menggunakan data elektronik sebanyak 895 data yang telah dilabeli oleh penulis dan dilakukan *text preprocessing*, berikut hasil pelabelan yang dilakukan pada tabel 4-3:

Tabel 4- 3 data elektronik dengan label

|  |  |
| --- | --- |
| Label | Total |
| Positif | 500 |
| Netral | 168 |
| Negatif | 232 |

Tabel 4-4 merupakan hasil dari pelabelan yang dilakukan, kemudian *wordcloud* atau awan kata untuk melihat kata-kata yang sering muncul pada masing-masing label data elektronik.

Tabel 4- 4 tabel wordcloud data elektronik

|  |  |
| --- | --- |
| *wordcloud* | Kata-kata yang muncul |
|  | Kata-kata yang muncul di data elektronik label positif:   1. kirim 2. barang 3. amannya |
|  | Kata-kata yang muncul di data elektronik label netral:   1. kirim 2. barang 3. sesuai |
|  | Kata-kata yang muncul di data elektronik label negaitf:   1. beli 2. kirim 3. nya |

## *Naives bayes*

*Naives bayes* yang digunakan pada penelitian ini ada 2 kategori data, yaitu data pakaian, data elektronik.

### Data Pakaian

Data pakaian dilakukan pemodelan menggunakan pembelajaran mesin *naives bayes*, sebelum melakukan pemodelan melakukan *text preprocessing,* pembobotan kata menggunakan *term frequency inverse document,* hasil pemodelan menggunakan berupa beberapa skenario evaluasi:

* Terdiri dari 10% (158 data) data test dan 90% (1417) data latih
* Terdiri dari 20% data uji atau 315 data uji dan 80% data latih atau (1260) data latih
* Terdiri dari 30% data uji atau 473 data uji dan 70% data latih atau 1102 data latih
* Terdiri dari 40% data uji atau 630 data uji dan 60% data latih atau 945 data latih

Dari empat skenario diharapkan menghasilkan nilai akurasi, *recall, precision,* yang terbaik dari salah satu skenario, berikut hasil evaluasi dan skenario pada tabel 4.5.

Tabel 4- 5 evaluasi naives bayes data pakaian

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Evaluasi | 40% data test dan 60% data latih | 30% data test dan 70% data latih | 20% data test dan 80% data latih | 10% data test dan 90% data latih |
| *accuracy* | 37% | 37% | 36% | 32% |
| *recall* | 37% | 37% | 36% | 39% |
| *precision* | 52% | 53% | 55% | 50% |

### Data elektronik

Data elektronik berjumlah 895 data dilakukan pemodelan menggunakan pembelajaran mesin *naives bayes*, sebelum melakukan pemodelan melakukan *text preprocessing,* pembobotan kata menggunakan *term frequency inverse document,* hasil pemodelan pada tabel 4-6 berupa evaluasi, yaitu akurasi, *recall, precision,* penelitian ini melakukan beberapa skenario, yaitu dengan pembagain data latih dan data uji sebagai berikut:

Tabel 4- 6 evaluasi naives bayes data elektronik

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Evaluasi | 40% data test dan 60% data latih | 30% data test dan 70% data latih | 20% data test dan 80% data latih | 10% data test dan 90% data latih |
| *accuracy* | 41% | 41% | 41% | 41% |
| *recall* | 41% | 41% | 41% | 41% |
| *precision* | 53% | 54% | 52% | 52% |

## *Decision tree*

*Decision tree* yang digunakan pada penelitian ini ada 2 kategori data, yaitu data pakaian, data elektronik.

### Data Pakaian

Data pakaian berjumlah 1575 data dilakukan pemodelan menggunakan pembelajaran mesin *decision tree*, sebelum melakukan pemodelan melakukan *text preprocessing,* pembobotan kata menggunakan *term frequency inverse document,* hasil pemodelan pada tabel 4-7. berupa evaluasi, yaitu akurasi, *recall, precision,* penelitian ini melakukan beberapa skenario, yaitu dengan pembagain data latih dan data uji sebagai berikut:

Tabel 4- 7 evaluasi decision tree data pakaian

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Evaluasi | 40% data test dan 60% data latih | 30% data test dan 70% data latih | 20% data test dan 80% data latih | 10% data test dan 90% data latih |
| *accuracy* | 62% | 63% | 65% | 68% |
| *recall* | 62% | 64% | 65% | 68% |
| *precision* | 61% | 61% | 64% | 67% |

### Data elektronik

Data elektronik berjumlah 895 data dilakukan pemodelan menggunakan pembelajaran mesin *decision tree*, sebelum melakukan pemodelan melakukan *text preprocessing*, pembobotan kata menggunakan *term frequency inverse document*, kemudian melakukan pemodelan, hasil pemodelan yang sudah dilakukan evaluasi, yaitu akurasi, *recall, precision*, penelitian ini melakukan beberapa skenario, yaitu dengan pembagain data latih dan data uji pada tabel 4.6 sebagai berikut:

Tabel 4- 8 evaluasi decision tree data elektronik

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Evaluasi | 40% data test dan 60% data latih | 30% data test dan 70% data latih | 20% data test dan 80% data latih | 10% data test dan 90% data latih |
| *accuracy* | 79% | 59% | 62% | 61% |
| *recall* | 79% | 60% | 63% | 61% |
| *precision* | 79% | 58% | 60% | 58% |

## *K-nearest neighbor*

*K-nearest neighbor* yang digunakan pada penelitian ini ada 2 kategori data, yaitu data pakaian, data elektronik.

### Data Pakaian

Data pakaian dilakukan pemodelan menggunakan pembelajaran mesin *k-nearest neighbor*, sebelum melakukan pemodelan melakukan *text preprocessing,* pembobotan kata menggunakan *term frequency inverse document,* hasil pemodelan berupa evaluasi, yaitu akurasi, *recall, precision,* penelitian ini melakukan beberapa skenario, yaitu dengan pembagian nilai K= [1,5] dengan data latih dan data uji pada tabel 4.9 sebagai berikut.

Tabel 4- 9 evaluasi k-nearest neighbor data pakaian.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nilai K | Pembagian data | *accuracy* | *recall* | *precision* |
| K =1 | 40% data test dan 60% data latih |  |  |  |
| 30% data test dan 70% data latih |  |  |  |
| 20% data test dan 80% data latih |  |  |  |
| 10% data test dan 90% data latih |  |  |  |
| K = 2 | 40% data test dan 60% data latih |  |  |  |
| 30% data test dan 70% data latih |  |  |  |
| 20% data test dan 80% data latih |  |  |  |
| 10% data test dan 90% data latih |  |  |  |
| K = 3 | 40% data test dan 60% data latih |  |  |  |
| 30% data test dan 70% data latih |  |  |  |
| 20% data test dan 80% data latih |  |  |  |
| 10% data test dan 90% data latih |  |  |  |
| K = 4 | 40% data test dan 60% data latih |  |  |  |
| 30% data test dan 70% data latih |  |  |  |
| 20% data test dan 80% data latih |  |  |  |
| 10% data test dan 90% data latih |  |  |  |
| K = 5 | 40% data test dan 60% data latih |  |  |  |
| 30% data test dan 70% data latih |  |  |  |
| 20% data test dan 80% data latih |  |  |  |
| 10% data test dan 90% data latih |  |  |  |

### Data elektronik

Data elektronik berjumlah 895 data dilakukan pemodelan menggunakan pembelajaran mesin *k-nearest neigbhbor*, sebelum melakukan pemodelan melakukan *text preprocessing*, pembobotan kata menggunakan *term frequency inverse document*, hasil pemodelan berupa evaluasi, yaitu akurasi, *recall, precision*, penelitian ini melakukan beberapa skenario, yaitu dengan pembagain data latih dan data uji sebagai berikut:

Tabel 4- 10 evaluasi k-nearest neighbor nilai K= 1 data elektronik

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nilai K | Pembagian data | *accuracy* | *recall* | *precision* |
| K =1 | 40% data test dan 60% data latih |  |  |  |
| 30% data test dan 70% data latih |  |  |  |
| 20% data test dan 80% data latih |  |  |  |
| 10% data test dan 90% data latih |  |  |  |
| K = 2 | 40% data test dan 60% data latih |  |  |  |
| 30% data test dan 70% data latih |  |  |  |
| 20% data test dan 80% data latih |  |  |  |
| 10% data test dan 90% data latih |  |  |  |
| K = 3 | 40% data test dan 60% data latih |  |  |  |
| 30% data test dan 70% data latih |  |  |  |
| 20% data test dan 80% data latih |  |  |  |
| 10% data test dan 90% data latih |  |  |  |
| K = 4 | 40% data test dan 60% data latih |  |  |  |
| 30% data test dan 70% data latih |  |  |  |
| 20% data test dan 80% data latih |  |  |  |
| 10% data test dan 90% data latih |  |  |  |
| K = 5 | 40% data test dan 60% data latih |  |  |  |
| 30% data test dan 70% data latih |  |  |  |
| 20% data test dan 80% data latih |  |  |  |
| 10% data test dan 90% data latih |  |  |  |

## Komparasi evaluasi model

Hasil pemodelan yang sudah dipaparkan dari *decision tree, naives bayes* sampai dengan *k-nearest neighbor*.Dari tabel evaluasi pemodelan 4.3 sampai tabel 4.17 dilakukan seleksi dari hasil akurasi, *precision, recall* yang menghasilkan hasil paling baik atau paling besar, berikut hasilnya:

Tabel 4. 19 komparasi hasil evaluasi decision tree

|  |  |  |
| --- | --- | --- |
| *Decision tree* menggunakan data pakaian | | *Decision tree* menggunakan data elektronik |
| Pembagian data uji dan data latih | | Pembagian data uji dan data latih |
| 10% data test dan 90 data latih | | 40% data test dan 60% data latih |
| *accuracy* | 68% | 79% |
| *recall* | 68% | 79% |
| *precision* | 67% | 79% |

Tabel 4. 20 komparasi hasil evaluasi naives bayes

|  |  |  |
| --- | --- | --- |
| *Naives bayes* menggunakan data pakaian | | *Naives ba* menggunakan data elektronik |
| Pembagian data uji dan data latih | | Pembagian data uji dan data latih |
| 30% data test dan 70 data latih | | 30% data test dan 70 data latih |
| *accuracy* | 37% | 41% |
| *recall* | 37% | 41% |
| *precision* | 53% | 54% |

Tabel 4. 21 komparasi hasil evaluasi k-nearest neighbor

|  |  |  |
| --- | --- | --- |
| *k-nearest neighbor* menggunakan data pakaian | | *k-nearest neighbor* menggunakan data elektronik |
| Pembagian data uji dan data latih | | Pembagian data uji dan data latih |
| 10% data test dan 90 data latih | | 10% data test dan 90 data latih |
| Nilai K =5 | | Nilai K =4 |
| *accuracy* | 65% | 70% |
| *recall* | 65% | 70% |
| *precision* | 63% | 68% |

Pada tabel 4.18 sampai 4.19 merupakan hasil evaluasi terbaik menggunakan pemodelan *naives bayes, decision tree, k - nearest neighbor* dengan data pakaian dan data elektronik. Nilai akurasi, *recall, precision* paling baik menggunakan pemodelan *decision tree* dengan data pakaian adalah melakukan pembagian data uji 10%, data latih 90% menghasilkan nilai *accuracy* sebesar 68%, nilai *recall* 68%, nilai *precision* 67%,untuk data elektronik menggunakan pembagian data uji 40%, data latih 60% menghasilkan nilai *accuracy* sebesar 79%, nilai *recall* 79%, nilai *precision* 79%.

Nilai akurasi, *recall, precision* menggunakan *naives bayes* dengan data pakaian adalah melakukan pembagian data uji 30%, data latih 70% menghasilkan nilai *accuracy* sebesar 37%, nilai *recall* 37%, nilai *precision* 53%, untuk data elektronik menggunakan pembagian data uji 30%, data latih 70% menghasilkan nilai *accuracy* sebesar 41%, nilai *recall* 41%, nilai *precision* 54%.

Pemodelan *k-nearest neighbor* (knn)menggunakan nilai K= 5 dengan data pakaian kemudian melakukan pembagian data uji 10% data latih 90% menghasilkan nilai akurasi 65%, *recall* 65%*, precision* 63%, pada pemodelan knn menggunakan nilai k = 4 dengann data data elektronik menghasilkan nilai paling baik dengan melakukan pembagian data uji sebanyak 10% data latih sebenyak 90% menghasilkan nilai akurasi 70%, *recall* 70%, *precision* 68%. Dari tabel 4.3 sampai 4.17 bahwa pembagain data uji dan data latih dan nilai K (menggunakan pemodelan *k-nearest neighbor*) berpengaruh pada hasil nilai evaluasi.

# BAB V

# KESIMPULAN DAN SARAN

## Kesimpulan

Dari hasil peneltitian yang telah dilakukan dengan melakukan komparasi algoritma pembelajaran mesin *supervised learning* menggunakan data sentimen ulasan pelanggan di Tokopedia dari ulasan pembelian elektronik, pembelian pakaian, dari data ulasan pembelian elektronik sebanyak 895 data, data ulasan pembelian pakaian sebanyak 1575, kemudian di implementasikan menggunakan pembelajaran mesin *naives bayes,decision tree, k-nearest neighbor*, menghasilkan hasil terbaik adalah menggunakan pembelajaran mesin *decision* dengan data uji sebanyak 40%, data latih sebanyak 60% menghasilkan nilai akurasi sebesar 79%, nilai *recall* 79%, nilai *precision* 79%.

## Saran-saran

Pada penelitian yang sudah dilakukan, peneliti memberikan saran bahwa penelitian ini berpengaruh dari banyaknya data untuk menghasilkan nilai evaluasi, pada tahap selanjutnya untuk mendapat hasil evaluasi yang lebih baik lagi, maka harus memperbanyak data, dan juga pada penelitian ini tidak dilakukan transformasi kata-kata yang mengandung bahasa inggris ke bahasa indonesia oleh karena itu diharapkan untuk memperhatikan data sentimen untuk diubah ke bahasa indonesia terlebih dahulu.

# DAFTAR PUSTAKA

Ainur Rahman, & Suroyo, H. (2021). Analisis Data Produk Elektronik Di E-Commerce Dengan Metode Algoritma K-Means Menggunakan Python. *Journal of Advances in Information and Industrial Technology*, *3*(2), 11–18. https://doi.org/10.52435/jaiit.v3i2.158

Apriani, R., Gustian, D., Program, S., Sistem, I., Putra, U. N., Indonesia, S., Raya, J., Kaler, C., 21, N., & Sukabumi, K. (2019a). 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

Apriani, R., Gustian, D., Program, S., Sistem, I., Putra, U. N., Indonesia, S., Raya, J., Kaler, C., 21, N., & Sukabumi, K. (2019b). 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

Cahyaningtyas, C., Nataliani, Y., & Widiasari, I. R. (2021). Analisis Sentimen Pada Rating Aplikasi Shopee Menggunakan Metode Decision Tree Berbasis SMOTE. *AITI*, *18*(2), 173–184. https://doi.org/10.24246/AITI.V18I2.173-184

Chamorro-Atalaya, O., Alvarado-Bravo, N., Aldana-Trejo, F., Poma-Garcia, C., Aliaga-Valdez, C., Peralta-Eugenio, G., & Tasayco-Jala, A. (2022). Supervised learning through k-nearest neighbor, used in the prediction of university teaching performance. *Indonesian Journal of Electrical Engineering and Computer Science*, *27*(3), 1625–1634. https://doi.org/10.11591/ijeecs.v27.i3.pp1625-1634

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

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

Deviyanto, A., & Wahyudi, M. D. R. (2018). PENERAPAN ANALISIS SENTIMEN PADA PENGGUNA TWITTER MENGGUNAKAN METODE K-NEAREST NEIGHBOR. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *3*(1), 1. https://doi.org/10.14421/jiska.2018.31-01

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

El Mohadab, M., Bouikhalene, B., & Safi, S. (2019). Predicting rank for scientific research papers using supervised learning. *Applied Computing and Informatics*, *15*(2), 182–190. https://doi.org/10.1016/j.aci.2018.02.002

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

Farisi, A. A., Sibaroni, Y., & Faraby, S. Al. (2019). Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier. *Journal of Physics: Conference Series*, *1192*(1), 012024. https://doi.org/10.1088/1742-6596/1192/1/012024

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

Handayani, R. N. (2021). Optimasi Algoritma Support Vector Machine untuk Analisis Sentimen pada Ulasan Produk Tokopedia Menggunakan PSO. *Media Informatika*, *20*(2), 97–108. https://doi.org/10.37595/MEDIAINFO.V20I2.59

Hardanto, L. T., & Wulandhari, L. A. (2021). Lithofacies Classification Using Supervised and Semi-Supervised Machine Learning Approach. *International Journal on Advanced Science, Engineering and Information Technology*, *11*(2), 542–548. https://doi.org/10.18517/ijaseit.11.2.11764

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

Ikhsanti, A. Y., Fauziah, Y., & Perwira, R. I. (2021). Implementation of the c4.5 decision tree learning algorithm for sentiment analysis in e-commerce application reviews on google play store. *Computing and Information Processing Letters*, *1*(1), 25. https://doi.org/10.31315/cip.v1i1.6128

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).

KURNIAWAN, R., & APRILIANI, A. (2020). ANALISIS SENTIMEN MASYARAKAT TERHADAP VIRUS CORONA BERDASARKAN OPINI DARI TWITTER BERBASIS WEB SCRAPER. *Jurnal INSTEK (Informatika Sains Dan Teknologi)*, *5*(1), 67. https://doi.org/10.24252/instek.v5i1.13686

Lee, C. S., Cheang, P. Y. S., & Moslehpour, M. (2022). Predictive Analytics in Business Analytics: Decision Tree. *Advances in Decision Sciences*, *26*(1), 1–29. https://doi.org/10.47654/V26Y2022I1P1-30

Limbong, J. J. A., Sembiring, I., & Hartomo, K. D. (2022). Analisis Klasifikasi Sentimen Ulasan pada E-Commerce Shopee Berbasis Word Cloud dengan Metode Naive Bayes dan K-Nearest Neighbor. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *9*(2), 347. https://doi.org/10.25126/jtiik.2022924960

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

Merinda Lestandy, Abdurrahim Abdurrahim, & Lailis Syafa’ah. (2021). Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(4), 802–808. https://doi.org/10.29207/resti.v5i4.3308

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

Pintoko, B. M., & Lhaksmana, K. M. (2018). Analisis Sentimen Jasa Transportasi Online Pada Twitter Menggunakan Metode NaÃ¯ve Bayes Classifier. *EProceedings of Engineering*, *5*(3). https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/7447

Pradana, A. W., & Hayaty, M. (2019). The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 375–380. https://doi.org/10.22219/kinetik.v4i4.912

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

Rahman, N. A., Idrus, S. D., & Adam, N. L. (2022). Classification of customer feedbacks using sentiment analysis towards mobile banking applications. *IAES International Journal of Artificial Intelligence*, *11*(4), 1579–1587. https://doi.org/10.11591/ijai.v11.i4.pp1579-1587

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.v7i1.11204

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).

Romli, I., Prameswari R, S., & Kamalia, A. Z. (2021). Sentiment Analysis about Large-Scale Social Restrictions in Social Media Twitter Using Algoritm K-Nearest Neighbor. *Jurnal Online Informatika*, *6*(1), 96. https://doi.org/10.15575/join.v6i1.670

Rozi, F. N., & Sulistyawati, D. H. (2019). KLASIFIKASI BERITA HOAX PILPRES MENGGUNAKAN METODE MODIFIED K-NEAREST NEIGHBOR DAN PEMBOBOTAN MENGGUNAKAN TF-IDF. *KONVERGENSI*, *15*(1). https://doi.org/10.30996/KONV.V15I1.2828

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

Samsir, Irmayani, D., Edi, F., Harahap, J. M., Jupriaman, Rangkuti, R. K., Ulya, B., & Watrianthos, R. (2021). Naives Bayes Algorithm for Twitter Sentiment Analysis. *Journal of Physics: Conference Series*, *1933*(1). https://doi.org/10.1088/1742-6596/1933/1/012019

Septian, J. A., Fachrudin, T. M., & Nugroho, A. (2019). Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor. *INSYST: Journal of Intelligent System and Computation*, *1*(1), 43–49. https://doi.org/10.52985/INSYST.V1I1.36

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

Stone, A. S., & Fathoni, F. (2022). Analisis Sentiment Pelanggan Terhadap Penilaian Produk Pada Toko Online Shop Amreta Menggunakan Metode Naïve Bayes Classification. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *6*(3), 1590. https://doi.org/10.30865/mib.v6i3.4436

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

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

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

Zamzami, F. N., Adiwijaya, A., & P, M. D. (2021). Analisis Sentimen Terhadap Review Film Menggunakan Metode Modified Balanced Random Forest dan Mutual Information. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(2), 415. https://doi.org/10.30865/mib.v5i2.2844

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

# LAMPIRAN