**PROPOSAL TA**

A picture containing text, clipart

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

**Disusun Oleh:**

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

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK**

**UNIVERSITAS MUHAMMADIYAH JAKARTA**

**2023**

**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

**DAFTAR ISI**

**DAFTAR TABEL**

**DAFTAR GAMBAR**

**DAFTAR LAMPIRAN**

**BAB I**

**PENDAHULUAN**

* 1. **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).

Pembelian 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).

Ilmu untuk analisa terhadap opini pembelian produk di toko *online* diperlukan untuk memahami mana opini yang bersifat positif atau negatif, maka ilmu untuk hal tersebut adalah *Natural processing, Natural processing language* merupakan cabang ilmu kecerdasan buatan yang dapat berinteraksi antara mesin dengan bahasa manusia (Nofiyanti & Oki Nur Haryanto, 2021). *Natural processing language* didalam penerapannya terdapat analisis sentimen (Vicari & Gaspari, 2021). Maka dari itu sangat tepat jika menggunakan analisis opini dari konsumen yang sudah melakukan pembelian untuk menentukan opini positif atau opini negatif berdasarkan kata.

Penelitian terhadap analisis sentimen di *e-commerce* pada ulasan restoran menggunakan *Naives Bayes* sebagai *machine learning,* peneliti tersebut menggunakan 1000 data, setelah mendapat 1000 data peneliti tersebut melakukan pelabelan, melakukan *preprocessing* dengan menghapus *noise*, selanjutnya melakukan ekstraksi fitur, membagi data menjadi data latih dan data uji, kemudian latih *Naives Bayes* menggunakan data latih, setelah melakukan latih penelitian Reddy melakukan uji dengan data uji menghasilkan akurasi sebesar 77,5% (Reddy & Reddy, 2021). Selain fokus masalah pada ulasan restoran, penelitian dengan *Naives Bayes* dilakukan oleh Apriani, fokus masalah komentar aplikasi Tokopedia di Googleplay, penelitian tersebut menghasilkan hasil akurasi sebesar 97,13% (Apriani et al., 2019).

Selain dari metode *Naives Bayes,* untuk sentimen analisis dapat menggunakan *decision tree* dengan fokus masalah mengenai ulasan hotel, Apriliani menggunakan ulasan dari tahun 2015 sampai 2018, serta menggunakan data bahasa indonesia, kemudian melakukan *preprocessing* data*,* peneliti melakukan model *decision* tree, dan juga menggunakan *cross validation* untuk mencari akurasi tertinggi, maka penelitian Apriliani menggunakan 8 Kfold menghasilkan akurasi 88,54% (Apriliani et al., 2020).

Selain dari kedua metode yang sudah disebutkan, analisis sentimen dapat menggunakan *K - Nearest Neighbor*, fokus masalah pada opini mengenai pilkada DKI (Daerah Khusus Ibukota) tahun 2017 di Twitter, penelitian ini menghasilkan akurasi sebesar 67,2% dengan menggunakan nilai K=5 (Deviyanto & Wahyudi, 2018). Selain fokus masalah mengenai opini pilkada, penelitian dilakukan yang Pajri dengan K*-Nearest Neighbor*, fokus masalah di *e-commerce* Tokopedia, menghasilkan akurasi sebesar 88,11% dengan nilai K=1(Pajri et al., 2020).

Berdasarkan beberapa penelitian yang sudah dipaparkan, maka penelitian ini melakukan analisa sentimen di *e-commerce* Tokopedia dengan berbagai *machine learning supervised*.

* 1. **Identifikasi Masalah**

Berdasarkan permasalahan di latar belakang, banyaknya penelitian terhadap analisis sentimen akan tetapi tidak adanya perbandingan menggunakan beberapa metode *machine learning,* maka diperlukan perbandingan agar dapat mengetahui akurasi dalam mengolah analisis sentimen.

* 1. **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?
   1. **Batasan Masalah**

Proposal tugas akhir ini memiliki batasan agar lebih mengerucut dan tidak

melebar, maka diberikan batasan-batasan sebagai berikut:

1. Melakukan pengambilan data dengan cara teknik *scraping*, 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. Mengolah data teks yang sudah didapatkan untuk dilakukan pelabelan, yaitu penilaian atau ulasan konsumen berdasarkan rating apakah opini tersebut positif, negatif, dan netral.
5. Melakukan pengolahan data dengan *case folding, Removal stopwords, stemming,* dan juga pembobotan *menggunakan Term Frequency Inverse Document*
6. Melakukan komparasi *machine learning decision tree, naives bayes, k - nearest neighbor* untuk sentimen analisis ulasan konsumen Tokopedia di fitur ulasan dan *review*.
7. Melakukan evaluasi dari tiga *machine learning,* menggunakan *confusion matrix, metrics accuracy, metrics recall, metrics precision.*
   1. **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 beberapa metode *machine learning* yang dapat digunakan untuk sentimen analisis ulasan *review* pembelian produk di Tokopedia.
   1. **Metodologi Penelitian**
4. **Data penelitian**

Data yang digunakan dalam proposal tugas akhir adalah data teks dari hasil *scrape* dari website Tokopedia yang data didalamnya terdapat konten berupa komentar dan rating berupa bintang 1 s/d 5, sedangkan data yang sudah didapat dari *scrape* dilakukan pelabelan berdasarkan rating bintang 1 s/d 5. Bintang 1-2 diberikan label negatif, bintang 3 diberikan label netral, bintang 4-5 diberikan label positif

1. **Pengolahan data**

Pada tahap ini dilakukan proses sebagai berikut:

1. *Case folding*

Merupakan tahap mengolah data teks jika memiliki huruf kapital atau *uppercase* makadiubah menjadi huruf kecil atau *lowercase* (KURNIAWAN & APRILIANI, 2020).

1. *Punctuation Removal*

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

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).

1. *Stemming*

Merupakan tahap untuk mengurangi prefiks sebuah kata menjadi kata dasar (Pradana & Hayaty, 2019).

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).

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*

*Decision* *tree* merupakan algoritma *supervised learning* yang bekerja seperti struktur pohon di setiap *node* atau simpul mewakili dari atribut yang dilatih (Panhalkar & Doye, 2022).

1. *Naïve Bayes*

*Naïve Bayes* merupakan algoritma klasifikasi probabilitas berdasarkan label data untuk memprediksi peluang masa depan dengan data sebelumnya (Watrianthos et al., 2019).

1. *K-Nearest-Neighbor*

Merupakan algoritma klasifikasi dengan menggunakan *input* fitur dan *output* fitur dengan melihat dari kelas atau fitur *Neighbor* terdekat (Cunningham & Delany, 2021).

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:

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).

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).

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)*.*

**BAB II**

**TINJAUAN PUSTAKA**

* 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., 2019).

* 1. **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)*.*

* 1. ***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).

* 1. ***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. *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)*.*

* 1. **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).

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).

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). 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).

* 1. **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**

* 1. **Data penelitian**

Data yang diambil dari ulasan pelanggan Tokopedia menggunakan teknik *scraping* menghasilkan 893 data kategori elektronik (laptop *handphone*), berikut data kategori elektronik dapat dilihat pada gambar 3.1, gambar 3.2.

A picture containing text, screenshot, plot, diagram

Description automatically generated

Gambar 3.1 grafik bar data laptop

Pada gambar 3.1, dapat diketahui bahwa data laptop menghasilkan masing-masing rating, yatu:

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

A picture containing text, screenshot, plot, diagram

Description automatically generated

Gambar 3.2 grafik bar data handphone

Pada gambar 3.2, dapat diketahui bahwa data *handphone* menghasilkan masing-masing rating, yatu:

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

Data kategori pakaian terdiri dari kemeja, kaos menghasilkan 1575

Data, berikut data kategori pakaian yang tediri dari kemeja, dan kaos pada gambar 3.3, dan gambar 3.4:

A picture containing text, screenshot, diagram, plot

Description automatically generated

Gambar 3.3 grafik bar data kaos

Pada gambar 3.3, dapat diketahui bahwa data kaos menghasilkan masing-masing rating, yatu:

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

A picture containing text, screenshot, plot, diagram

Description automatically generated

Gambar 3.4 grafik bar data kemeja

Pada gambar 3.4, dapat diketahui bahwa data kemeja menghasilkan masing-masing rating, yatu:

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

Setelah dilakukan pengambilan 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 gambar.



Gambar 3.5 grafik bar data laptop

Pada gambar 3.5, dapat diketahui bahwa data laptop menghasilkan 3 kelas atau label diantaranya sebagai berikut:

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

Gambar 3.6 grafik bar data handphone

Pada gambar 3.6, dapat diketahui bahwa data tersebut menghasilkan 3 kelas atau label diantaranya sebagai berikut:

1. Total data *handphone* label positif 262 data
2. Total data *handphone* label netral 89 data
3. Total data *handphone* label negatif 142 data

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

Gambar 3.7 grafik bar data kaos

Pada gambar 3.7, dapat diketahui bahwa data kaos menghasilkan 3 kelas atau label diantaranya sebagai berikut:

1. Total data kaos label positif 495 data
2. Total data kaoslabel netral 190 data
3. Total data kaos label negatif 245 data

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

Gambar 3.8 grafik bar data kemeja

Pada gambar 3.8, dapat diketahui bahwa data kemeja menghasilkan 3 kelas atau label diantaranya sebagai berikut:

1. Total data kaos label positif 345 data
2. Total data kaoslabel netral 145 data
3. Total data kaos label negatif 146 data

Setelah dilakukan pelabelan, maka tahap selanjutnya persiapan data teks adalah *text preprocessing* sebelum menuju ke tahap pemodelan.

* 1. **Text preprocessinng**

Tahap *preprocessing text* adalah tahap untuk menyiapkan data teks sebelum dilakukan pelatihan ke pemodelan *machine learning*

* + 1. ***Casefolding***

*Casefolding* merupakan tahap untuk transformasi data teks menjadi huruf kecil. Program yang digunakan untuk yaitu:

Gambar 3.2.2 kode caseFolding

* + 1. ***Punctuation removal***

*Punctuation removal* merupakan tahapan untuk menghapus tanda baca dan nomor karena agar tidak memperbanyak bobot kata pada tahap pembobotan kata, pada tahap ini digunakan *regular expression.* Kode yang digunakan untuk tahapan ini sebagai berikut:

A picture containing text, screenshot, font, multimedia

Description automatically generated

Gambar 3.2.3 kode punctuation removal

* + 1. ***Stopwords removal***

*Stopwords removal* merupakan tahapan untuk menghilangkan kata hubung, seperti “nya”, “dan”, dihapuskan karena untuk mengurangi bobot kata dalam proses pembobotan kata, untuk tahap ini menggunakan *library* nltk bahasa indonesia, berikut pemograman yang digunakan:

A screen shot of a computer code

Description automatically generated with low confidence

Gambar 3.2.4 kode stopwords removal

* + 1. ***Stemming***

*Stemming* merupakan tahapan transformasi teks data kata menjadi ke bentuk dasar, seperti “menyapu” -> “sapu”, pada tahapan ini menggunakan *library* sastrawi, berikut pemograman sebagai berikut:

A picture containing text, screenshot, font

Description automatically generated

Gambar 3.2.5 kode stemming.

* + 1. **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 = kemunculan frekuensi kata di setiap dokumen

Tahap ini menggunakan *library scikit-learn* *tfidfvectorizer* berikut program yang digunakan berikut ini:

A screen shot of a computer code

Description automatically generated with low confidence

Gambar 3.2.6 kode pembobotan kata.

Pada gambar 3.2.6 data yang sudah bersih dari *preprocessing* dilakukan *term frequency inverse document*, lalu sesudah dilakukan *term frequency inverse document* ubah datanya menjadi bentuk tipe data *array.*

* 1. **Pemodelan**

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

* + 1. ***Naives bayes***

Pada tahap ini menggunakan *machine learninng* *naives bayes* dari *libray scikit-learn*, dua *naives bayes* yang digunakan untuk klasifikasi, antara lain *Multinomial Naive Bayes, Gaussian Naive Bayes*. Berikut pemograman yang digunakan:

A picture containing text, screenshot, font

Description automatically generated

Gambar 3.3.1 kode naives bayes.

Pada gambar 3.3.1 sebelum dilakukan pelatihan terlebih dahulu pembagian antara data latih, data testing, agar *machine learning* yang dilatih menggunakan data latih, hasil dari latih dilakukan prediksi dengan data testing.

* + 1. ***Decision tree***

Pada tahap pembelajaran mesin *decision tree* atau pohon keputusan, metode pohon keputusan ini bekerja berdasarkan *node* atau cabang sampai N tingkatan yang disetiap *node* mempunyai *leaf,* metode pembelajaran mesin *decision tree* menggunakan dari *scikit-learn,* berikut program yang digunakan:

A screen shot of a computer code

Description automatically generated with low confidence

Gambar 3.3.2 kode naives decision tree.

Pada gambar 3.3.2 merupakan program yang digunakan, pada penelitian ini akan melakukan perubahan parameter, antara lain sebagai berikut:

1. *Criterion* menggunakan dua yaitu, *gini, entropy.*
2. *Max depth* atau kedalaman pohon, dari 1 – 10
3. *Max feature* menggunakan *auto, sqrt, log2*
4. Max leaf node dari 1-10

Dari parameter yang sudah disebutkan akan dilakukan akan dikomparasi mana yang menghasilkan hasil akurasi tertinggi.

* + 1. ***K-nearest neighbor***

Pada tahap pembelajaran mesin *k-nearest neighbor. K-nearest neighbor* bekerja berdasarkan label dari nilai K tetangga terdekat. Pada tahap ini menggunakan *library* dari *scikit-learn* untuk membuat *machine learning k-nearest neighbor*, dan juga melakukan analisa dari perubahan parameter sebagai berikut:

1. Nilai K menggunakan dari rentang 1 - 10
2. Algoritma menghitung tetangga terdekat menggunakn 3, yaitu *ball\_tree, kd\_tree, brute*

Dari parameter yang sudah disebutkan akan dilakukan analisa komparasi mana yang menghasilkan hasil akurasi tertinggi.

* 1. **Evaluasi**

Pada tahap ini melakukan evaluasi dari tahap pemodelan dengan data teks, kemudian dilakukan prediksi dari data testing, dari hasil prediksi dibutuhkan evaluasi untuk melihat hasil akurasi, penelitian ini menggunakan evaluasi dengan empat metode, yaitu:

1. *Recall*

Melihat hasil dari hasil total prediksi benar (*true positifve)*, dengan total prediksi salah (*false negative),* kemudian dikalkukasi dengan cara sebagai berikut:

1. *Precision*

Melihat dari hasil total *true positive* dan hasil total *false positive,* kemudian dikalkukasi dengan cara sebagai berikut:

1. *Accuracy*

Melihat dari hasil total *true positive,* total *false negative,* total *true negative*, dan *false negative,* kemudian dilakukan kalkulasi dengan cara sebagai berikut:

1. *Confusion matrix*

Merupakan matriks yang mempunyai kolom dan baris sesuai dengan label kelas klasifikasi pada penelitian ini, yaitu 5 kolom dan 5 baris, untuk melihat berapa total *true positive,* total *false negative,* total *true negative*, dan *false negative,* berikut *confusion matrix* dari tabel 3.4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Kelas | Prediksi  Sangat baik | Prediksi  Baik | Prediksi  Biasa | Prediksi  Buruk | Prediksi  Sangat buruk |
| Kelas Asli:  Sangat baik |  |  |  |  |  |
| Kelas Asli:  Baik |  |  |  |  |  |
| Kelas Asli:  Biasa |  |  |  |  |  |
| Kelas Asli:  Buruk |  |  |  |  |  |
| Kelas Asli:  Sangat buruk |  |  |  |  |  |

Tabel 3.4 Tabel *confusion matrix*

**DAFTAR PUSTAKA**

Apriani, R., Gustian, D., Program, S., Sistem, I., Putra, U. N., Indonesia, S., Raya, J., Kaler, C., 21, N., & Sukabumi, K. (2019). ANALISIS SENTIMEN DENGAN NAÏVE BAYES TERHADAP KOMENTAR APLIKASI TOKOPEDIA. *Jurnal Rekayasa Teknologi Nusa Putra*, *6*(1), 54–62. https://doi.org/10.52005/REKAYASA.V6I1.86

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

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

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

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

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

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

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

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

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).

Kang, S. (2021). K-nearest neighbor learning with graph neural networks. *Mathematics*, *9*(8). https://doi.org/10.3390/math9080830

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

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

Nofiyanti, E., & Oki Nur Haryanto, E. M. (2021). Analisis Sentimen terhadap Penanggulangan Bencana di Indonesia. *Jurnal Ilmiah SINUS*, *19*(2), 17. https://doi.org/10.30646/sinus.v19i2.563

Pajri, D., Umaidah, Y., & Padilah, T. N. (2020). K-Nearest Neighbor Berbasis Particle Swarm Optimization untuk Analisis Sentimen Terhadap Tokopedia. *Jurnal Teknik Informatika Dan Sistem Informasi*, *6*(2). https://doi.org/10.28932/jutisi.v6i2.2658

Panhalkar, A. R., & Doye, D. D. (2022). Optimization of decision trees using modified African buffalo algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*(8), 4763–4772. https://doi.org/10.1016/j.jksuci.2021.01.011

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

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

Reddy, K. N., & Reddy, D. B. I. (2021). Restaurant Review Classification Using Naives Bayes Model. *Journal of University of Shanghai for Science and Technology*, *23*(08), 646–656. https://doi.org/10.51201/JUSST/21/08443

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

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

Vicari, M., & Gaspari, M. (2021). Analysis of news sentiments using natural language processing and deep learning. *AI and Society*, *36*(3), 931–937. https://doi.org/10.1007/s00146-020-01111-x

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

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

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