E-ISSN: 2655-0865
DOI: https://doi.org/10.38035/rrj.v6i5
Received: 27 Juni 2024, Revised: 24 Juli 2024, Publish: 25 Juli 2024

Ranah Research :
Journal of Multidisciplinary Research and Development
+62 821-7074-3613 ranahresearch@gmail.com https://jurnal.ranahresearch.com/

Klasifikasi Sentimen pada Aplikasi Shopee Menggunakan Fitur Bag of Word dan Algoritma Random Forest

Ahnaf Ananta Firdaus[1], Asep Id Hadiana[2], Ade Kania Ningsih[3]
[1] Universitas Jenderal Achmad Yani, Cimahi, Indonesia, ahnafanantaf18@if.unjani.ac.id
[2] Universitas Jenderal Achmad Yani, Cimahi, Indonesia, asep.hadiana@lecture.unjani.ac.id
[3] Universitas Jenderal Achmad Yani, Cimahi, Indonesia, Ade.kanianingsih@lecture.unjani.ac.id

Corresponding Author: ahnafanantaf18@if.unjani.ac.id

Abstract: *Sentiment analysis is a classification method used to categorize opinions contained within a text. There are three types of opinions: positive, negative, and neutral. Sentiment analysis is often utilized to understand public opinion about e-commerce applications, such as Shopee, through reviews found in application comments. In this study, the Bag of Words feature was employed to represent text data in a vector format that can be processed by machine learning algorithms. The algorithm used was Random Forest, known for its robust handling of complex data and ability to produce accurate predictions. The results of this study indicate that the model achieved an accuracy of 84.91%. This outcome demonstrates that the approach used is quite effective in analyzing user sentiments towards the Shopee application. This research makes a significant contribution to the application of machine learning in text analysis and opens up opportunities for further development in this field.*

Keyword: *Sentiment Analysis, Shopee, Bag of Word, Random Forest.*

Abstrak: Analisis sentimen merupakan metode klasifikasi yang digunakan untuk mengelompokkan opini yang terkandung dalam sebuah teks. Terdapat tiga jenis opini, yaitu opini positif, negatif, dan netral. Analisis sentimen sering digunakan untuk mengetahui opini masyarakat terhadap aplikasi e-commerce, seperti Shopee, melalui ulasan yang terdapat di komentar aplikasi. Dalam penelitian ini, digunakan fitur Bag of Words untuk merepresentasikan data teks ke dalam bentuk vektor yang dapat diolah oleh algoritma machine learning. Algoritma yang digunakan adalah Random Forest, yang dikenal memiliki kemampuan yang baik dalam menangani data yang kompleks dan menghasilkan prediksi yang akurat. Hasil dari penelitian ini menunjukkan bahwa model yang dibangun memiliki akurasi sebesar 84,91%. Hasil ini menunjukkan bahwa pendekatan yang digunakan cukup efektif dalam menganalisis sentimen pengguna terhadap aplikasi Shopee. Penelitian ini memberikan kontribusi signifikan dalam penerapan machine learning untuk analisis teks dan membuka peluang untuk pengembangan lebih lanjut dalam bidang ini.

Kata Kunci: Analisis Sentiment, Shopee, Bag of Word, Random Forest.

PENDAHULUAN

Analisis sentimen merupakan sebuah metode yang digunakan untuk mengidentifikasi atau mengelompokan opini yang terkandung dalam sebuah teks terdapat tiga buah opini yaitu positif, negatif dan netral. Teknik ini biasanya sering digunakan untuk mengolah data hasil komentar, ulasan produk, atau sosial media dan sering ditemukan juga di berbagai *platform* di mana pengguna dapat bebas mengekspresikan pendapat mereka. Analisis sentimen juga merupakan sebuah proses cabang dari *text mining* yang merupakan proses klasifikasi dalam sebuah teks proses ekstraksi pendapat, emosi, dan evaluasi yang diungkapkan oleh seseorang mengenai topik tertentu, menggunakan teknik pemrosesan bahasa alam [(Cahyaningtyas et al., 2021)]. Beberapa penelitian mengenai analisis sentimen sudah pernah dilakukan pada beberapa sosial media seperti twitter atau x di mana topiknya tentang kenaikan harga BBM menggunakan metode ekstraksi fitur *Bag Of Word* untuk algoritma sendiri menggunakan *Support Vector Machine* mendapatkan hasil akurasi 65% (Darmawan et al., 2023). Lalu ada juga yang menggunakan analisis sentimen yang data set berasal dari komentar pada salah satu toko beras di aplikasi *Shopee* menggunakan metode *bag of word* dan algoritma *naive bayes* mendapatkan akurasi sebesar 76%(Sugiarto et al., 2023).

Shopee adalah salah satu *marketplace* yang terkenal di Indonesia. Pada kuartal ketiga tahun 2023, Shopee menempati posisi pertama sebagai e-commerce dengan jumlah pengunjung bulanan terbanyak di Indonesia. Menurut data dari databoks Shopee mencatatkan 273 juta pengunjung pada periode tersebut(Cahyaningtyas et al., 2021). Ulasan positif dari pelanggan dapat meningkatkan persepsi calon pembeli dan mendorong minat mereka untuk membeli produk. Sebaliknya, ulasan negatif dapat menyebabkan calon pembeli membatalkan niat mereka untuk melakukan pembelian(Sugiarto et al., 2023).

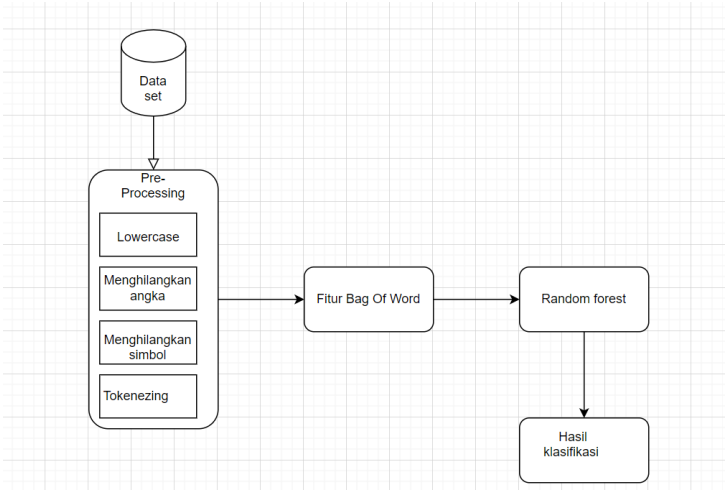
Random forest adalah sebuah metode pembelajaran yang digunakan untuk klasifikasi dan regresi. RF membangun sejumlah Decision Tree selama proses pelatihan. Ketika menghadapi kasus baru, setiap Decision Tree akan diaktifkan dan menghasilkan kelas, label, atau output. Hasil akhir yang dipilih didasarkan pada suara terbanyak atau kelas dengan jumlah output terbanyak(Kusuma Pangkasidhi et al., n.d.). Selama proses pelatihan, RF akan mengambil data pelatihan secara acak. Random forest adalah algoritma yang dapat mengklasifikasikan sejumlah besar data dengan akurat, dengan hasil akhir diperoleh melalui penentuan node akar dan berakhir pada beberapa node daun(Syafia et al., 2023).

Pada penelitian sebelumnya mengenai analisis sentimen menggunakan algoritma *random forest* atau bukan sudah pernah dilakukan dan juga untuk data set sendiri didapatkan bukan hanya dari ulasan *marketplace* tetapi ada juga penelitian yang dilakukan dengan mengambil data dari sosial media seperti twitter seperti pada penelitian yang bertujuan untuk identifikasi ujaran kebencian di media sosial twitter menggunakan metode *backpropagation neural network* berbasis *lexicon base featured* dan *bag of word* mendapatkan akurasi sebesar 78,81% (Munir et al., 2018). Selain dari media sosial ada juga data set yang didapatkan dari portal berita seperti pada penelitian Komparasi Ekstraksi Fitur dalam Klasifikasi Teks Multilabel Menggunakan Algoritma Machine Learning di mana fitur *bag of word* menggunakan algoritma *decision tree* mendapatkan akurasi terbesar pada *split* data 80-20 dengan akurasi sebesar 67%(Efrizoni et al., 2022).

Oleh karena itu penelitian ini akan dikerjakan menggunakan metode *bag of word* sebagai fitur ekstrasi kata menggunakan metode *random forest* untuk melakukan proses klasifikasi.

METODE

Metode penelitian akan dimulai dengan proses pre-processing atau pembersihan data set sebelum di proses ada 4 tahapan dalam proses pre-processing yaitu lowercase yaitu mengubah huruf besar menjadi kecil, menghilangkan angka, menghilangkan simbol dan tokenezing lalu setelah selesai di pre-processing memasuki mode ekstraksi kata menggunakan *bag of word* setelah selesai lanjut ke metode random forest untuk proses klasifikasi seperti pada gambar 1.



Gambar.1 Metode Penelitian

Pre-processing

Tahap awal adalah pre-processing tahap ini diperlukan untuk memproses data agar lebih efektif pada tahap pre-processing ada empat tahapan yaitu hilangkan angka di mana semua angka dari isi atribut ulasan akan dihilangkan, lalu ada tahap lowercasting di mana mengubah semua huruf kapital menjadi huruf kecil, setelah itu masuk ke tahap pembersihan dari tanda baca atau simbol karena tidak digunakan juga dalam klasifikasi lalu untuk proses terakhir ada tokenezing yaitu perubahan kalimat menjadi perkata untuk diolah fitur *bags of word*. Seperti pada tabel 1 berikut

Tabel 1 hasil sebelum dan sesudah pre processing

Teks sebelum pre processing	Teks sesudah pre processing
puas belanja di shopee banyak diskon dan gratis ongkir. penjualnya baik dan kurir sopan.	['puas', 'belanja', 'di', 'shopee', 'banyak', 'diskon', 'dan', 'gratis', 'ongkir', 'penjualnya', 'baik', 'dan', 'kurir', 'sopan']
Aplikasi yang bagus dan sangat membantu.	['aplikasi', 'yang', 'bagus', 'dan', 'sangat', 'membantu']
saya kecewa kenapa sekarang Shopee sangat eror. Mau checkout aja susah	['saya', 'kecewa', 'kenapa', 'sekarang', 'shopee', 'sangat', 'eror', 'mau', 'checkout', 'aja', 'susah']
Shope sangat bagus bagi saya saya suka banget sama shopee	['shope', 'sangat', 'bagus', 'bagi', 'saya', 'saya', 'suka', 'banget', 'sama', 'shopee']

Bag of Word

Metode ini merupakan salah satu cara yang cukup sederhana dalam memproses data teks dengan mengubahnya menjadi vektor angka agar dapat diproses oleh komputer. Pada dasarnya, metode ini hanya menghitung frekuensi kemunculan kata dalam keseluruhan dokumen yang diproses(Tri Putra et al., n.d.).

Dalam notasi matematis, Bag of Words (BOW) dapat direpresentasikan dengan cara berikut: jika d adalah sebuah dokumen dan v adalah kumpulan kata atau kosakata dari keseluruhan dokumen, maka representasinya adalah sebagai berikut: BoW(d) = [count(w1,d), count(w2,d… count(wn,d)]

Random Forest

Random forest classifier adalah metode klasifikasi yang diibaratkan sebagai sekumpulan pohon keputusan dengan data latih dan fitur acak independen yang memiliki berbagai fitur berbeda (Syafia et al., 2023). Algoritma random forest mampu mengklasifikasikan sejumlah besar data dengan akurat, di mana hasil akhirnya diperoleh melalui penentuan node akar dan berakhir pada beberapa node daun..

HASIL DAN PEMBAHASAN

Penelitian ini dimulai dengan langkah awal yaitu tahap *pre-processing* di mana pada tahap ini dilakukan pembersihan dari data set pembersihan ini memiliki tujuan agar data yang di proses lebih efektif dan efisien setelah melalui tahap pre processing selanjutnya dilakukan tahap pembobotan kata menggunakan *bag of word* mengubah tokenisasi data menjadi vektor yang akan diolah oleh algoritma random forest mendapatkan hasil akurasi 84,91%.

Tahap Pre-Processing

Penelitian ini dimulai dengan tahap pre-processing yang bertujuan untuk membersihkan data dari unsur-unsur yang tidak relevan dan mengganggu proses analisis. Pre-processing merupakan langkah krusial karena data yang bersih dan terstruktur akan meningkatkan efektivitas dan efisiensi dalam pemrosesan selanjutnya. Tahap ini meliputi beberapa langkah, antara lain:

1. Penghapusan Karakter Khusus Menghilangkan karakter seperti tanda baca, angka, dan simbol yang tidak memberikan informasi berarti dalam konteks analisis teks.
2. Stopword Removal Menghapus kata-kata umum yang tidak memberikan nilai informasi signifikan, seperti "dan", "yang", "di", dan sebagainya.
3. Stemming dan LemmatizationMengubah kata-kata menjadi bentuk dasar atau akar katanya untuk menyederhanakan variasi kata yang akan diproses.

Tahap Pembobotan Kata

Setelah tahap pre-processing, dilakukan tahap pembobotan kata menggunakan metode Bag of Words. Metode ini mengubah tokenisasi data teks menjadi representasi vektor yang dapat diolah oleh algoritma machine learning. Proses ini meliputi:

1. Tokenisasi Memecah teks menjadi unit-unit kata (token).
2. Pembentukan Vektor Setiap dokumen direpresentasikan sebagai vektor yang panjangnya sama dengan jumlah seluruh kata unik dalam korpus, di mana setiap elemen vektor menyatakan frekuensi atau keberadaan kata tersebut dalam dokumen.

Algoritma Random Forest

Data yang telah dibobotkan kemudian diolah menggunakan algoritma Random Forest. Random Forest adalah algoritma ensemble learning yang terdiri dari sejumlah pohon keputusan yang dilatih secara independen. Keputusan akhir ditentukan melalui voting dari hasil prediksi masing-masing pohon. Algoritma ini dipilih karena keandalannya dalam menangani data yang kompleks dan kemampuannya untuk mengurangi risiko overfitting.

Hasil

Pengujian dilakukan untuk mengevaluasi kinerja model yang telah dibangun. Dari hasil pengujian, model yang dibentuk menunjukkan akurasi sebesar 84,91%. Akurasi ini menunjukkan bahwa model memiliki kemampuan yang baik dalam mengklasifikasikan data sesuai dengan kategori yang benar.

KESIMPULAN

Hasil akurasi sebesar 84,91% menunjukkan bahwa pendekatan yang dilakukan cukup efektif dalam pemrosesan dan klasifikasi data teks. Beberapa faktor yang mempengaruhi hasil ini antara lain kualitas data awal, efektivitas pre-processing, dan pemilihan algoritma yang tepat. Kualitas data awal yang bersih dan representatif sangat mempengaruhi performa model, sehingga tahap pre-processing menjadi sangat penting. Proses pembersihan dan transformasi data yang tepat membantu meningkatkan kualitas fitur yang digunakan dalam model.

Algoritma Random Forest terbukti efektif dalam menangani data teks yang kompleks dengan memberikan hasil yang konsisten dan akurat. Namun, terdapat beberapa aspek yang bisa ditingkatkan untuk penelitian selanjutnya, seperti penggunaan metode pembobotan kata lain seperti TF-IDF atau Word Embeddings yang bisa jadi memberikan hasil yang lebih baik, optimasi hyperparameter pada algoritma Random Forest untuk mencari kombinasi yang optimal sehingga bisa meningkatkan akurasi model, serta eksplorasi algoritma lain seperti SVM, Naive Bayes, atau Neural Networks untuk melihat apakah ada peningkatan performa. Dengan demikian, penelitian ini memberikan kontribusi signifikan dalam penerapan machine learning untuk analisis teks dan membuka peluang untuk pengembangan lebih lanjut dalam bidang ini.

REFERENSI

Rusdi Rahman, M., & Febri Diansyah, A. (2024). Sentiment Analysis on the Shopee Application on Playstore Using the Random Forest Classification Method. *Inform : Jurnal Ilmiah Bidang Teknologi Informasi Dan Komunikasi*, *9*(1). https://doi.org/10.25139/inform.v9i1.5465

Muara Sains, J., Ilmu Kesehatan, dan, Trisari Harsanti Putri, W., & Hendrowati, R. (2018). *PENGGALIAN TEKS DENGAN MODEL BAG OF WORDS TERHADAP DATA TWITTER*. *2*(1), 129–138.

Utama, H., & Masruro, A. (2022). *Analisis Sentimen pada Twitter menggunakan Word Embedding dengan Pendekatan Word2Vec*.

Rusdi Rahman, M., & Febri Diansyah, A. (2024). Analisis Sentimen pada Aplikasi Shopee di Playstore Menggunakan Metode Klasifikasi Random Forest. *Jurnal Ilmiah Bidang Teknologi Informasi Dan Komunikasi*, *9*(1). https://doi.org/10.25139/inform.v9i1.5465

Suswadi, S., & Erkamim, Moh. (2023). Sentiment Analysis of Shopee App Reviews Using Random Forest and Support Vector Machine. *ILKOM Jurnal Ilmiah*, *15*(3), 427–435. https://doi.org/10.33096/ilkom.v15i3.1610.427-435

Saepudin, A., Faqih, A., & Dwilestari, G. (n.d.). *Perbandingan Algoritma Klasifikasi Support Vector Machine, Random Forest dan Logistic Regression Pada Ulasan Shopee* (Vol. 18, Issue 1).

Rozy, F., Rangkuti, S., Fauzi, M. A., Sari, Y. A., Dewi, E., & Sari, L. (2018). *Analisis Sentimen Opini Film Menggunakan Metode Naïve Bayes dengan Ensemble Feature dan Seleksi Fitur Pearson Correlation Coefficient* (Vol. 2, Issue 12). http://j-ptiik.ub.ac.id

Sugiarto, D., Utami, E., & Yaqin, A. (n.d.). *Perbandingan Kinerja Model TF-IDF dan BOW untuk Klasifikasi Opini Publik Tentang Kebijakan BLT Minyak Goreng*.

Khomsah, S., & Agus Sasmito Aribowo. (2020). Text-Preprocessing Model Youtube Comments in Indonesian. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *4*(4), 648–654. https://doi.org/10.29207/resti.v4i4.2035

Munir, M. M., Fauzi, M. A., & Perdana, R. S. (2018). *Implementasi Metode Backpropagation Neural Network berbasis Lexicon Based Features dan Bag of Words Untuk Identifikasi Ujaran Kebencian Pada Twitter* (Vol. 2, Issue 10). http://j-ptiik.ub.ac.id

Kusuma Pangkasidhi, M., Novianus Palit, H., & Gunawan, A. (n.d.). *Analisis Sentimen Mahasiswa di Surabaya Terhadap Pelayanan Vaksinasi COVID-19 Menggunakan Beberapa Classifier*.

Sugiarto, D., Sari, S., Ariwibowo, A. B., Nabilah Putri, F., Mulya, D., Aulia, T., & Zaki, A. N. (2023). *PERBANDINGAN KINERJA KLASIFIKASI SENTIMEN ULASAN PRODUK PEMBELIAN BERAS DI MARKETPLACE SHOPEE*. *17*(1). https://doi.org/10.47111/JTI

Cahyaningtyas, C., Nataliani, Y., & Widiasari, I. R. (2021). Analisis sentimen pada rating aplikasi Shopee menggunakan metode Decision Tree berbasis SMOTE. *AITI: Jurnal Teknologi Informasi*, *18*(Agustus), 173–184.

Muara Sains, J., Ilmu Kesehatan, dan, Trisari Harsanti Putri, W., & Hendrowati, R. (2018). *PENGGALIAN TEKS DENGAN MODEL BAG OF WORDS TERHADAP DATA TWITTER*. *2*(1), 129–138.