Jurnal Sistem Informasi dan Ilmu Komputer
Vol. 2, No. 1 Februari 2024
e-ISSN: 2986-4976; p-ISSN: 2986-5158, Hal 77-86
DOI: https://doi.org/10.59581/jusiik-widyakarya.v2i1.2262

Perbandingan Algoritma Random Forest dan Logistic Regression Untuk Analisis Sentimen Ulasan Aplikasi Tumbuh Kembang Anak Di Play Store

Muhammad Alfyando
Universitas Pembangunan Nasional Veteran Jawa Timur

Fetty Tri Anggraeny
Universitas Pembangunan Nasional Veteran Jawa Timur

Andreas Nugroho Sihananto
Universitas Pembangunan Nasional Veteran Jawa Timur

Alamat: Jl. Rungkut Madya No.1, Gn. Anyar,
Kec. Gn. Anyar, Surabaya, Jawa Timur 60294
Korespondensi penulis: 19081010037@student.upnjatim.ac.id

Abstract. *Early childhood plays an important role in forming the basis of development, which involves stimulation of various aspects such as moral religious values, social emotional, language, cognitive, and physical motor skills. The concept of early childhood learning is focused on play, where every activity is designed to be play, so that learning becomes more effective. Parents also need to understand today's children's education to interact with children positively. This research focuses on sentiment analysis of children's education-based app reviews on the Google Play Store, using Random Forest and Logistic Regression methods. The review data is taken from three apps with the theme of child development, namely "About Kids", "PrimaKu", and "Teman Bumil", with a range of review years between 2018 and 2023. The test results show that Logistic Regression has higher accuracy compared to Random Forest, especially in the "About Kids" and "PrimaKu" applications with accuracy above 90%. The conclusion of this research highlights the importance of sentiment analysis in improving understanding of user responses to children's education applications, with suggestions for future research to increase the number of datasets and variations in testing schemes by tuning hyperparameters to improve prediction accuracy and more optimal results.*

Keywords: *sentiment, child development, random forest, logistic regression.*

Abstrak. Anak usia dini memegang peran penting dalam pembentukan dasar perkembangan, yang melibatkan stimulasi beragam aspek seperti nilai agama moral, sosial emosional, bahasa, kognitif, dan fisik motorik. Konsep belajar anak usia dini terfokus pada bermain, di mana setiap aktivitas dirancang untuk menjadi bermain, sehingga belajar menjadi lebih efektif. Orang tua juga perlu memahami pendidikan anak zaman sekarang untuk berinteraksi dengan anak secara positif. Penelitian ini berfokus pada analisis sentimen terhadap ulasan aplikasi berbasis pendidikan anak di Google Play Store, menggunakan metode Random Forest dan Logistic Regression. Data ulasan diambil dari tiga aplikasi dengan tema tumbuh kembang anak, yaitu "Tentang Anak", "PrimaKu", dan "Teman Bumil", dengan rentang tahun ulasan antara 2018 hingga 2023. Hasil pengujian menunjukkan bahwa Logistic Regression memiliki akurasi lebih tinggi dibandingkan dengan Random Forest, terutama pada aplikasi "Tentang Anak" dan "PrimaKu" dengan akurasi di atas 90%. Kesimpulan penelitian ini menyoroti pentingnya analisis sentimen dalam meningkatkan pemahaman terhadap respons pengguna terhadap aplikasi pendidikan anak, dengan saran untuk penelitian selanjutnya menambah jumlah dataset dan variasi skema pengujian dengan tunning hyperparameter untuk meningkatkan akurasi prediksi dan hasil yang lebih optimal.

Kata kunci: sentimen, tumbuh kembang anak, random forest, logistic regression.

Received November 19, 2023; Accepted Desember 21, 2023; Published Februari 29, 2024
* Muhammad Alfyando, 19081010037@student.upnjatim.ac.id

LATAR BELAKANG

Anak usia dini merupakan tonggak peletakan dasar perkembangan, dengan cara memberikan berbagai stimulasi guna merangsang aspek-aspek perkembangan pada anak, seperti Nilai agama moral, sosial emosional, bahasa, kognitif, dan fisik motorik. Dunia anak adalah dunia bermain, apapun aktivitas yang dilakukan hendaknya dikemas dengan bermain, karena anak usia dini konsep belajaranya bermain seraya belajar, dan belajar seraya bermain (Juniarti, 2021). Orang tua juga wajib untuk mempelajari Pendidikan anak zaman sekarang agar dapat memahami keadaan dan dapat mengikuti perkembangan anak pada zaman sekarang. Dengan begitu anak bisa merasa nyaman dengan orang tuanya dan tidak merasa tertekan (Sudarsana, 2017). Hingga saat ini terdapat banyak sekali penelitan-penelitian yang membahas mengenai pendidikan anak dan teknologi yang digunakan. Khususnya untuk pendidikan berbasis teknologi, terdapat beberapa cara dan alternative yang telah dikembangkan oleh para peneliti diantaranya adalah pendidikan berbasis aplikasi.

Menurut informasi dari databoks.katadata.co.id, jumlah unduhan aplikasi global pada kuartal pertama tahun 2021 meningkat sebesar 8,7% dibandingkan dengan periode yang sama tahun sebelumnya. Dalam periode tersebut, Google Play Store menjadi platform unduhan terbesar dengan total unduhan sebesar 3,8 miliar (Hendriyanto, 2022).

Pemberian rating aplikasi di Google Play Store diikuti dengan ulasan dari para pengguna terhadap aplikasi tersebut. Ulasan tersebut mengandung opini dari para pengguna mengenai aplikasi tersebut dan calon pengguna melihat ulasan dari sebuah aplikasi sebagai pertimbangan sebelum memutuskan untuk menggunakan aplikasi tersebut (Saputra, 2019). Oleh karena itu, diperlukan analisis sentimen pada data ulasan untuk mengolah data tekstual untuk memperoleh informasi pada teks (Wahyudi, 2021). Ada banyak metode atau studi komputasi untuk mengolah sentimen, salah satunya yakni Random Forest dan Logistic Regression.

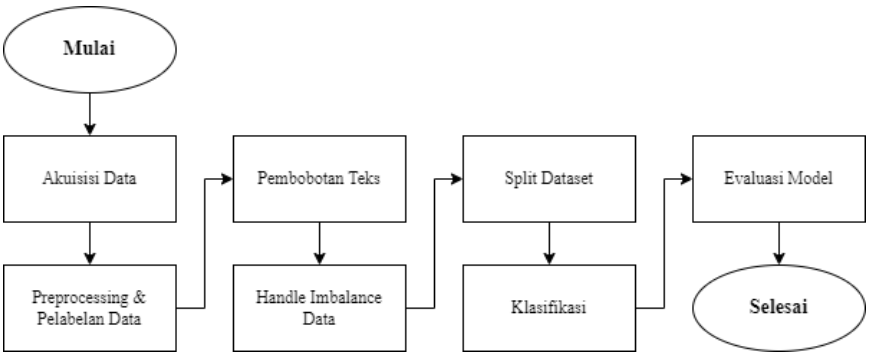
Random Forest adalah salah satu algoritma pembelajaran berkinerja terbaik, lebih akurat memperkirakan tingkat kesalahan dibandingkan dengan decision tree. Karena tingkat kesalahan telah terbukti secara matematis (Schonlau & Zou, 2020). Model Random Forest adalah kumpulan decision tree, digunakan

untuk klasifikasi atau regresi. Dalam kasus klasifikasi digunakan untuk prediksi, didasarkan pada suara terbanyak dari nilai prediksi menggunakan decision tree, dan dalam kasus regresi, hasilnya adalah rata-rata dari hasil tree.

Logistic Regression merupakan model statistik yang digunakan untuk menentukan apakah sebuah independent variable memiliki pengaruh terhadap sebuah binary dependent variable. Dengan menggunakan sigmoid function, Logistic Regression menghasilkan output sebuah probabilitas antara angka 0 dan 1 (Pangaribuan, 2021).

Berdasarkan paparan diatas, maka peneliti akan melakukan analisis sentimen dengan topik ulasan aplikasi Tentang Anak di Play Store, dari data ulasan yang diperoleh akan dilakukan pembobotan kemudian diolah menggunakan metode Random Forest dan Logistic Regression yang akan menghasilkan nilai akurasi sentiment positif dan negatif dari masing-masing metode tersebut.

METODE PENELITIAN



Gambar 1. Metode Penelitian

Akuisisi Data

Akuisisi Data merupakan tahapan pertama dalam sistem, dimana akuisisi data dilakukan menggunakan library Python yakni google_play_scraper. Pada penelitian ini peneliti menggunakan 3 ulasan aplikasi yang bertema tumbuh kembang anak yang terdapat pada *playstore*, dengan rentang tahun ulasan sebagai berikut.

Tabel 1. Rentang Tahun Ulasan

Aplikasi	Rentang Tahun Ulasan
Tentang Anak	2021 - 2023
PrimaKu	2018 - 2023
Teman Bumil	2018 - 2023

Preprocessing & Pelabelan

Teks atau tulisan yang diperoleh dari suatu sumber disebut data mentah. Hal ini dikarenakan dataset merupakan data yang tidak terstruktur sehingga menyulitkan komputer untuk mengolah data tersebut. Oleh karena itu, diperlukan langkah text preprocessing sebelum melakukan penentuan fitur (Anggraeny, 2019).

Tahapan pada proses ini meliputi pembersihan text (*cleaning*), mengubah huruf menjadi huruf kecil (*lowercasing*), memperbaiki kesalahan penulisan pada kata (*normalization*), mengubah kalimat dalam teks menjadi potongan kata (*tokenizing*), membersihkan kata-kata yang sekiranya tidak diperlukan seperti kata ganti, depan dan sambung (*stopword removal*), mengubah kata-kata yang ada menjadi bentuk kata dasar (*stemming*).

Langkah Selanjutnya melakukan pelabelan data, data yang digunakan untuk pelabelan adalah data hasil dari *normalization* yang kemudian diubah menjadi Bahasa inggris, karena proses pelabelan ini menggunakan *library* VADER yang hanya support Bahasa inggris.

Pembobotan Teks (TF-IDF)

Pada tahap ini terdapat dua bagian proses yaitu TF (Term Frequency) dan IDF (Inverse Document Frequency), TF adalah jumlah kemunculan kata dalam sebuah dokumen, semakin banyak kata yang muncul pada setiap dokumen maka semakin besar pula nilai TF-nya. IDF adalah jumlah nilai dokumen untuk setiap kata yang berbanding terbalik, yaitu jika sebuah kata jarang muncul dalam sebuah dokumen, maka nilai IDF lebih besar dari kata yang sering muncul (Septian, 2019). Secara matematis, TF-IDF dapat dijabarkan sebagai berikut.

$$W = TF\ x\ IDF \qquad (2.3)$$

$$W = f_{t,d}\ x \log\left(\frac{\mathrm{D}}{\mathrm{df(t)}}\right) \qquad (2.4)$$

Dalam konteks ini, kita dapat menggunakan beberapa parameter untuk mengukur pentingnya suatu kata kunci dalam sebuah dokumen. Bobot (W) dapat dihitung dengan menggabungkan nilai Term Frequency (TF/f_(t,d)), Inverse Document Frequency (IDF), dan jumlah dokumen secara keseluruhan (D). Term Frequency (TF/f_(t,d)) mengukur seberapa sering kata kunci t muncul dalam dokumen d. Inverse Document Frequency (IDF) memberikan bobot tambahan berdasarkan seberapa umum kata kunci tersebut di seluruh dokumen. Selanjutnya, jumlah dokumen yang mengandung kata kunci t (df(t)) juga memainkan peran dalam menghitung bobot total. Dengan menggunakan rumus-rumus ini, kita dapat memperoleh representasi bobot yang akurat untuk mengevaluasi signifikansi suatu kata kunci dalam konteks keseluruhan dokumen

Handle Imbalance Data

Pada proses ini menggunakan metode SMOTE, proses SMOTE dimulai dengan mengukur jarak antara datapoints dalam dataset kelas minoritas. Selanjutnya, kita menentukan persentase SMOTE yang akan digunakan, dan menentukan jumlah tetangga terdekat (k) yang akan digunakan dalam proses pembuatan data sintetis. Langkah terakhir adalah menciptakan data sintetis dengan menggunakan persamaan berikut (Kasanah, 2019).

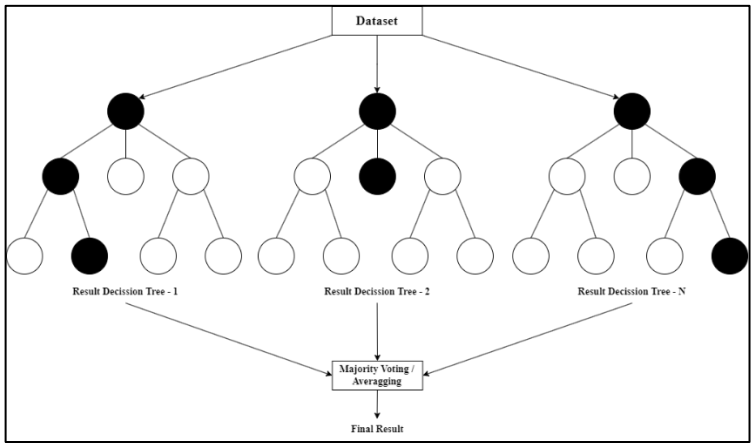
Split Dataset

Dalam tahap split dataset, dilakukan proses pemisahan data menggunakan rasio 70% untuk data latih dan 30% untuk data uji.

Klasifikasi

1. Random Forest

Random Forest adalah salah satu algoritma klasifikasi yang paling terkenal dan paling populer, algoritma ini telah digunakan dalam banyak studi pembelajaran mesin karena keakuratan dan kematangannya (Sihananto, 2023).



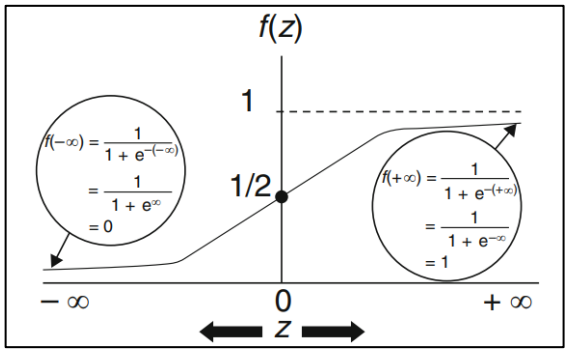
Gambar 2. Ilustrasi Random Forest

Langkah-langkah dalam penerapan metode Random Forest antara lain:

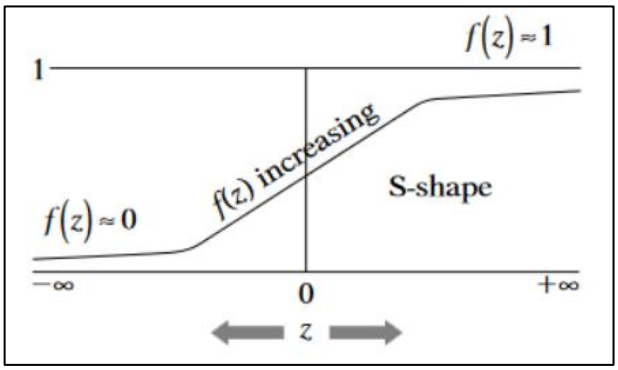
1. Membuat data sampel dengan cara pengambilan acak dengan pengembalian dari dataset.
2. Gunakan sampel data untuk membangun pohon ke i (i=1, 2, 3, ..., k)
3. Ulangi langkah 1 dan 2 sebanyak k kali

2. Logistic Regression

Logistic regression adalah model komputasi yang digunakan untuk menentukan apakah sebuah variable independent memiliki pengaruh terhadap sebuah binary variable dependent. Logistic Regression popular untuk digunakan disebabkan dari hasil logistic function f(z) yang dimana hasil yang diberikan diantara 0 dan 1. Model tersebut di desain untuk menjelaskan tentang probabilitas yang dimana selalu diantara nilai 0 dan 1. Salah satu contohnya yaitu probabilitas dari suatu individual terjangkit suatu penyakit (David G. Kleinbaum & Mitchel Klein, 1994, p.5-6). Pada “Gambar 2.1. Rumus Logistic Regression” dibawah, dapat dilihat bahwa hasil yang akan diberikan yaitu antara nilai 0 dan 1, walaupun nilai masukkan untuk variabel z itu sendiri memiliki masukkan nilai yang berbeda-beda (David G. Kleinbaum & Mitchel Klein, 1994, p.5-6).



Gambar 3. Rumus Logistic Regression (David G. Kleinbaum and Mitchel Klein, *Logistic Regression: A Self-Learning Text, 2nd Edition*, Halaman 6)



Gambar 4. Shape Logistic Regression (David G. Kleinbaum and Mitchel Klein, *Logistic Regression: A Self-Learning Text, 2nd Edition*, Halaman 6)

Pada "Gambar 3. dan Gambar 4. Shape Logistic Function", Terlihat bahwa hasilnya mendekati 1 jika semakin besar atau besar nilai masukan (z), sedangkan hasilnya mendekati 0 bila nilai masukan (z) kecil. Dari sini, model logistik kemudian dapat memastikan bahwa semua perkiraan hasil yang diperoleh selalu berupa angka antara 0 dan 1.

Evaluasi Model

Evaluasi model merupakan proses analisis sentimen yang dilakukan dalam penelitian ini. Pada fase ini, hasil evaluasi disajikan dalam bentuk nilai-nilai tertentu yang dapat mewakili kinerja model dalam kaitannya dengan atribut-atribut yang telah disiapkan sebelumnya.

HASIL DAN PEMBAHASAN

Hasil dan pembahasan memuat hasil penelitian dan pembahasana terkait hasil penelitian tersebut, hasil yang pertama ialah akuisisi data:

Tabel 2. Total Hasil Akuisis Data

Nama Aplikasi	Total
Tentang Anak	623
PrimaKu	2045
Teman Bumil	2115

Setelah berhasil mendapatkan data, kemudian akan dilakukan proses preprocessing dan pelabelan data, berikut merupakan hasil dari proses *preprocessing* yang nantinya akan digunakan untuk pelabelan (dari data *normalization*) dan proses pembobotan, agar dapat diolah secara komputasi.

Tabel 3. Preprocessing

Sebelum preprocessing	Setelah preprocessing
selalu minta update padahal kemaren baru diperbarui,selalu minta update padahal kemaren baru diperbarui	['selalu', 'minta', 'update', 'padahal', 'kemaren', 'baru', 'baru']

Pada tahap hasil pelabelan data Tabel , dapat dilihat bahwa kelas negatif dari masing-masing data ulasan aplikasi memiliki jumlah perbandingan yang cukup jauh dengan kelas positif, maka dari itu akan diakukan proses penanganan *imbalance data* dengan SMOTE, berikut hasilnya.

Tabel 4. Hasil SMOTE

Nama Aplikasi	Jumlah Positif	Jumlah Negatif	Total
Tentang Anak	441	441	882
PrimaKu	1306	1306	2612
Teman Bumil	1147	1147	2294

Selanjutnya, data latih hasil SMOTE dan label data dilakukan pelatihan data. Pelatihan data ini melibatkan pengklasifikasi Random Forest dan Logistic Regression. Data hasil pengujian dari data masing-masing data latih yang terbentuk ditunjukkan Tabel 5.

Tabel 5 Hasil Skenario Pengujian

No.	Aplikasi	Metode	Accuracy	Recall	Precission	F1-Score
1.	Tentang Anak	Random Forest	90.94 %	85.96 %	92.45 %	89.09 %
2.		Logistic Regression	92.83 %	85.96 %	97.03 %	91.16 %
3.	PrimaKu	Random Forest	88.01 %	80.64 %	93.54 %	86.61 %
4.		Logistic Regression	88.52 %	83.82 %	91.59 %	87.53 %
5.	Teman Bumil	Random Forest	85.78 %	82.57 %	88.65 %	85.5 %
6.		Logistic Regression	84.33 %	79.71 %	88.29 %	83.78 %

KESIMPULAN DAN SARAN

Berdasarkan evaluasi enam skenario menggunakan metode Random Forest dan Logistic Regression, dapat disimpulkan bahwa secara umum Logistic Regression menunjukkan kinerja yang lebih baik dengan akurasi lebih tinggi, terutama pada aplikasi "Tentang Anak" dan "PrimaKu" dengan akurasi di atas 90%. Meskipun demikian, perlu diperhatikan bahwa recall pada Logistic Regression cenderung lebih rendah dibandingkan precision, menunjukkan kecenderungan untuk mengklasifikasikan instance positif secara tidak benar. Sebaliknya, Random Forest, meskipun memiliki akurasi yang sedikit lebih rendah, menunjukkan recall yang lebih baik pada beberapa aplikasi. Pemilihan antara kedua metode sebaiknya bergantung pada kebutuhan spesifik aplikasi dan pentingnya masing-masing metrik evaluasi, dengan pertimbangan untuk keseimbangan antara recall dan precision. Perlu diingat bahwa evaluasi model tidak hanya bergantung pada akurasi tetapi juga mempertimbangkan metrik lainnya seperti recall, precision, dan F1-Score untuk mendapatkan pemahaman yang lebih komprehensif tentang kinerja model.

Adapun saran yang dapat peneliti sampaikan ialah Penelitian selanjutnya diharapkan dapat menambah jumlah dataset. Jumlah dataset yang makin banyak membuat model algoritma menjadi lebih akurat dalam melakukan prediksi dan penambahan skema pengujian dengan *tunning hyperparameter* yang bervariatif untuk mendapatkan hasil yang lebih optimal.

DAFTAR REFERENSI

Anggraeny, F. T., Purbasari, I. Y., & Wulandari, E. F. (2019). Undergraduate Thesis Supervisor Recommendation Based On Text Similarity. INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY AND BUSINESS (ICITB) 5, Icitb 2019, 86–94. https://jurnal.darmajaya.ac.id/index.php/icitb/article/view/2077.

Hendriyanto, Muhammad Diki. 2022. ANALISIS SENTIMEN ULASAN APLIKASI MOLA PADA GOOGLE PLAY STORE MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE. Journal of Information Technology and Computer Science (INTECOMS). Volume 5 Nomor 1, Juni 2022.

Juniarti, Y. Utoyo, Setiyo. & Ramadan, Gilang. 2021. Pengembangan Aplikasi Game Edukasi dalam Membentuk Karakter anak. WIDYA WACANA: JURNAL ILMIAH.

Kasanah, A. N., Muladi, M., & Pujianto, U. (2019). Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 3(2), 196–201. https://doi.org/10.29207/resti.v3i2.945.

Pangaribuan, Jefri Junifer. 2021.Mendeteksi Penyakit Jantung Menggunakan Machine Learning Dengan Algoritma Logistic Regression. Information System Development, VOLUME 6 NO.2 JULI 2021.

Saputra. 2021. ANALISIS SENTIMEN APLIKASI INVESTASI ONLINE DI GOOGLE PLAY STORE MENGGUNAKAN METODE ALGORITMA RANDOM FOREST.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. Stata Journal, 20(1), 3–29. https://doi.org/10.1177/1536867X20909688.

Sihananto, A. N., Safitri, E. M., Subagio, A. W., Ardiansyah, M. D., Primayudha, A. (2023). Classification of Covid-19 RT-PCR Test Results Using Auto-encoder And Random Forest. 7stInternational Seminar of Research Month 2022. NST Proceedings. halaman 237-243. https://nstproceeding.com/index.php/nuscientech/article/view/944/898.

Wahyudi, R., & Kusumawardana, G. (2021). Analisis Sentimen pada Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine. Jurnal Informatika, 8(2), 200–207. https://doi.org/10.31294/ji.v8i2.96 81.