

Naïve Bayes Classifier Method on Sentiment Analysis of Bibit Application Users in Play Store

Afifa Lufti Inani, Zamahsary Martha*, Yenni Kurniawati, Zilrahmi

Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia

*Corresponding author: zamahsarymartha@fmipa.unp.ac.id

Submitted : 19 September 2023

Revised : 23 Oktober 2023

Accepted : 30 November 2023

ABSTRACT

The Bibit app is one of the most widely used investment apps these days. This application is widely used by novice investors because of its convenience in opening accounts, disbursing funds, purchasing mutual funds and easy-to-understand application design. However, there are still many people who doubt and worry about the quality of the Bibit application due to the lack of understanding of the advantages and disadvantages of the Bibit application. So, review data on the application is used which is available in the play store with the aim of knowing user reviews of the application and being a consideration for prospective users before using the application. Because reviews on the application have a large number and can be positive or negative, so sentiment analysis is needed that can help classify these reviews quickly. Then classification is carried out to obtain a classification model that can be used to predict user sentiment using the Naive Bayes Classifier method. The results obtained by Bibit application users tend to have positive sentiments with an accuracy value of 79.45%.

Keywords: Aplikasi Bibit, Naive Bayes Classifier(NBC), Sentiment Analysis



This is an open access article under the Creative Commons 4.0 Attribution License, which permits unrestricted use, distribution, and reproduction in medium, provided the original work is properly cited. ©2022 by author and Universitas Negeri Padang.

I. PENDAHULUAN

Berdasarkan data Kementerian Investasi/BKPM, investasi di Indonesia mengalami peningkatan selama periode 2017-2022. Mulai meningkatnya minat masyarakat Indonesia akan berinvestasi dan didukung oleh kemajuan teknologi, investasi yang awalnya hanya dapat dilakukan secara konvensional tetapi sekarang dapat dilakukan secara online. Sehingga, banyaknya muncul aplikasi-aplikasi investasi ditengah masyarakat yang bertujuan membantu masyarakat untuk berinvestasi secara online, kapan saja dan dimana saja (Kamal & Apriani, 2022). Salah satu aplikasi investasi yang banyak digunakan saat ini oleh masyarakat Indonesia yaitu aplikasi Bibit. Aplikasi ini memiliki rating 4.5 dan sudah didownload lebih dari 5 juta pengguna di play store. Aplikasi ini banyak digunakan oleh investor pemula karena kemudahan dalam membuka rekening, pencairan dana yang dapat dilakukan secara online, pembelian reksa dana yang mudah, desain aplikasi yang mudah dipahami dan dapat berinvestasi mulai dari Rp10.000 (Kamalia dkk, 2022). Namun, ada sebagian masyarakat yang baru ingin memulai berinvestasi secara online memiliki keraguan dan kekhawatiran akan kualitas dari aplikasi investasi ini. Hal ini karena kurang pemahannya masyarakat terhadap aplikasi Bibit dan kurang mengetahui kelebihan serta kekurangan dari aplikasi investasi Bibit (Saputra dkk, 2021).

Sehingga, untuk mengetahui kelebihan dan kekurangan aplikasi Bibit dibutuhkan kumpulan informasi yang didapat dari pengalaman pengguna aplikasi sebelumnya. Dimana, digunakan data ulasan pada aplikasi tersebut yang tersedia di play store dengan tujuan untuk mengetahui ulasan pengguna terhadap aplikasi dan menjadi bahan pertimbangan bagi calon pengguna sebelum menggunakan aplikasi tersebut. Karena ulasan pada aplikasi memiliki jumlah yang banyak dan dapat bersifat positif maupun negatif, sehingga diperlukan analisis sentimen yang dapat membantu dalam mengklasifikasikan ulasan tersebut agar proses penganganan terhadap ulasan menjadi cepat (Kamalia dkk, 2022). Pada analisis sentimen dilakukan pengelolaan data tesktual berupa opini, penilaian dan ulasan terhadap produk serta layanan apakah cenderung positif atau negatif (Pakpahan & Hilda, 2014). Untuk mengklasifikasi sentimen tersebut digunakan metode *Naive Bayes Classifier* (NBC). NBC salah satu metode klasifikasi untuk data sentimen yang banyak digunakan pada analisis sentimen karena merupakan metode yang sederhana tetapi memiliki tingkat akurasi yang tinggi dalam mengklasifikasian teks (Routtray, 2013).

Penelitian dengan menggunakan NBC pernah dilakukan oleh Tanggraeni & Melkior (2022) mengenai sentimen pengguna aplikasi *E-Government* memperoleh ketepatan klasifikasi dengan nilai akurasi sebesar 89%.

Kemudian, penelitian yang dilakukan Putra & Safitri (2021) mengenai sentimen pengguna aplikasi Bibit dengan metode K-Nearest Neighbor (KNN) diperoleh nilai akurasi sebesar 85,14%. Dari penjelasan tersebut, maka dilakukan penelitian yang bertujuan untuk mengklasifikasikan sentimen pengguna aplikasi Bibit ke dalam dua kelas yaitu positif dan negatif, dan mengetahui performa klasifikasi metode NBC pada data sentimen.

II. METODE PENELITIAN

A. Jenis Penelitian dan Sumber Data

Penelitian ini merupakan penelitian terapan. Dimana, jenis data yang digunakan yaitu data sekunder yang diperoleh dari kumpulan ulasan pada aplikasi Bibit di Play Store. Pengambilan data dilakukan dengan menggunakan teknik *scraping*. Teknik *scraping* merupakan teknik mengumpulkan data pada sebuah website dengan memasukan *Uniform Resource Locator* (URL)(Arsi & Waluyo, 2021). Variabel penelitian yang digunakan terlihat pada Tabel 1.

Tabel 1. Variabel Penelitian

| Variabel | Keterangan |
|----------|---|
| X | Frekuensi kata ke-i yang muncul pada objek (sentimen) |
| Y | 0 = sentimen negatif 1 = sentimen positif |

B. Tahapan Penelitian

Adapun langkah-langkah yang dilakukan pada penelitian ini sebagai berikut.

1. Mengumpulkan data ulasan aplikasi Bibit di Play Store dengan teknik *scraping*.
2. Melakukan *preprocessing*.

Preprocessing tahapan yang dilakukan sebelum mengolah data dengan tujuan menyiapkan data dengan melakukan pembersihan dan pemilihan data agar memiliki data yang berkualitas sehingga siap digunakan. *Preprocessing* ini memiliki beberapa tahapan sebagai berikut (Chairunnisa dkk, 2022).

- a. *Case folding*, yaitu mengubah semua teks dengan huruf kecil.
 - b. *Cleansing*, yaitu membersihkan karakter seperti simbol, tanda baca, *emoticon*, angka dan lain-lain.
 - c. *Tokenizing*, yaitu memisahkan kata pada suatu kalimat sehingga menjadi sebuah kata individu.
 - d. *Stemming*, yaitu mengubah kata yang memiliki imbuhan menjadi kata dasar.
 - e. *Stopwords*, yaitu menghapus kata yang tidak relevan dengan teks atau tidak memiliki arti.
3. Melakukan pelabelan pada data.

Data akan diberi label positif atau negatif secara otomatis dengan menggunakan metode *lexicon based*. Metode ini bekerja berdasarkan kamus *lexicon* yang terdapat pada *library Python* yang berisi daftar kata-kata positif, netral dan negatif, selanjutnya dilakukan pembobotan untuk menentukan sentimen positif atau negatif dari data (Mahendrajaya dkk, 2019).

4. Melakukan pembobotan kata.

Pembobotan kata merupakan proses perhitungan dan penetapan bobot setiap kata untuk melihat seberapa penting kata tersebut dalam teks. Pembobotan kata dilakukan menggunakan algoritma TF-IDF. Algoritma TF-IDF merupakan algoritma yang berfungsi untuk melihat nilai frekuensi kemunculan kata dalam suatu teks (Chairunnisa dkk, 2022). Berikut rumus TF-IDF yang digunakan ditulis pada persamaan (1).

$$W_{ij} = tf_{ij} \times idf_i \tag{1}$$

dimana W_{ij} adalah bobot dari kata i pada teks ke- j , tf_{ij} adalah jumlah kemunculan kata i pada teks ke- j , dan idf_i adalah jumlah teks yang mengandung kata ke- i .

5. Membagi data menjadi dua bagian yaitu data *training* dan data *testing*.

Data *training* digunakan untuk melatih algoritma untuk menentukan model yang sesuai dan data *testing* digunakan untuk menguji performa model yang telah dilatih sebelumnya.

6. Melakukan klasifikasi menggunakan metode *Naive Bayes Classifier*.

Naive Bayes Classifier (NBC) merupakan metode pengklasifikasian statistik yang didasarkan pada teorema bayes yaitu menghitung probabilitas untuk memprediksi probabilitas masa depan berdasarkan pengalaman sebelumnya (Tanggraeni & Melkior, 2022). Nilai probabilitas NBC secara umum dinyatakan dengan persamaan (2) berikut.

$$P(V_j|a_i) = \frac{P(a_i|V_j) P(V_j)}{P(a_i)} \tag{2}$$

Dimana $P(V_j|a_i)$ probabilitas kelas ke- j ketika kata ke- i muncul, $P(a_i|V_j)$ yaitu probabilitas kata ke- i terdapat dalam kelas ke- j , $P(V_j)$ probabilitas jumlah kejadian kelas ke- j , $P(a_i)$ probabilitas jumlah munculnya kata ke- i . Klasifikasi dilakukan dengan tujuan menentukan kelas terbaik untuk suatu teks dengan cara mencari

probabilitas tertinggi dari semua kelas teks yang diujikan dengan mencari nilai *maximum a posteriori* (MAP) kelas V_{map} dengan persamaan (3) berikut (Nisa, dkk., 2019).

$$V_{map} = \underset{V_j \in V}{\operatorname{argmax}} P(V_j) \times \prod_{i=1} P(a_i|V_j) \quad (3)$$

7. Menghitung ketepatan klasifikasi.

Pengukuran ketepatan klasifikasi berfungsi untuk melihat performa klasifikasi yang telah dilakukan. Untuk menghitung ketepatan klasifikasi digunakan *confusion matrix*. *Confusion matrix* mengetahui tingkat ketepatan klasifikasi dengan menghitung jumlah kuantitas kelas aktual dan kelas prediksi (Saputro & Sari, 2019). Pada Tabel 2 ditunjukkan tabel *confusion matrix*.

Tabel 2. *Confusion Matrix*

| Prediksi | Aktual | |
|----------|--------------------|--------------------|
| | Positif | Negatif |
| Positif | True Positif (TP) | False Positif (FP) |
| Negatif | False Negatif (FN) | True Negatif (TN) |

Dimana, TP adalah jumlah data kelas aktual positif yang diprediksi positif, FN adalah jumlah data kelas aktual positif yang diprediksi negatif, FP adalah jumlah data kelas aktual negatif diprediksi positif dan TN adalah jumlah data kelas aktual negatif yang diprediksi negatif.

Setelah mendapatkan nilai *confusion matrix* dilakukan perhitungan ketepatan klasifikasi untuk mendapatkan nilai akurasi yang berguna mengukur tingkat keberhasilan suatu model klasifikasi (Wahyuni dkk., 2020). Berikut persamaan (4) yang digunakan.

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FN+FP} \times 100\% \quad (4)$$

Nilai akurasi dari *confusion matrix* menggambarkan sejauh mana model klasifikasi berhasil memprediksi dengan benar. Semakin tinggi nilai akurasi maka semakin baik performa model tersebut dalam mengklasifikasikan data.

III. HASIL DAN PEMBAHASAN

A. Preprocessing

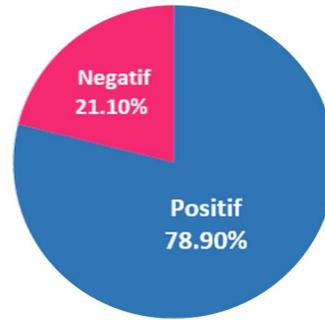
Pada tahap ini akan dilakukan proses *case folding*, *cleansing*, *tokenizing*, *stemming*, dan *stopwords* pada data yang telah di *scraping*. Berikut hasil dari *preprocessing* data sentimen aplikasi Bibit ditunjukkan pada Tabel 3.

Tabel 3. Hasil *Preprocessing*

| Ulasan | Keterangan |
|--|--------------|
| Sangat mudah untuk aplikasinya dan pemilihan investasinya, diberikan rekomendasi sesuai profil resiko. Hanya saja kalau bisa untuk pembayarannya cukup dengan 1 akun saja jadi tidak perlu banyak akun. 🙌👍😊 | Data mentah |
| sangat mudah untuk aplikasinya dan pemilihan investasinya, diberikan rekomendasi sesuai profil resiko. hanya saja kalau bisa untuk pembayarannya cukup dengan 1 akun saja jadi tidak perlu banyak akun. 🙌👍😊 | Case Folding |
| sangat mudah untuk aplikasinya dan pemilihan investasinya diberikan rekomendasi sesuai profil resiko hanya saja kalau bisa untuk pembayarannya cukup dengan akun saja jadi tidak perlu banyak akun | Cleansing |
| sangat, mudah, untuk, aplikasinya, dan, pemilihan, investasinya, diberikan, rekomendasi, sesuai, profil, resiko, hanya, saja, kalau, bisa, untuk, pembayarannya, cukup, dengan, akun, saja, jadi, tidak, perlu, banyak, akun | Tokenizing |
| sangat, mudah, untuk, aplikasi, dan, milih, investasi, beri, rekomendasi, sesuai, profil, resiko, hanya, saja, kalau, bisa, untuk, bayar, cukup, dengan, akun, saja, jadi, tidak, perlu, banyak, akun | Stemming |
| mudah, aplikasi, investasi, rekomendasi, profil, resiko, bayar, akun | Stopwords |

B. Pelabelan Data

Pelabelan akan dilakukan secara otomatis dengan metode *lexicon based*. Sehingga, didapatkan hasil perbandingan jumlah sentimen pada data terlihat pada Gambar 1.



Gambar 1. Persentase Kategori Sentimen Aplikasi Bibit

Pada Gambar 1 menunjukkan persentase sentimen pengguna aplikasi Bibit memperoleh persentase sebesar 78.9% untuk sentimen positif dan 21.1% untuk sentimen negatif. Artinya, sentimen pengguna aplikasi Bibit cenderung bersifat positif.

C. Klasifikasi Naive Bayes Classifier

Sebelum melakukan pengklasifikasian maka data set dibagi menjadi data *training* dan data *testing* terlebih dahulu dengan rasio 80%:20% yang dapat dilihat pada Tabel 4 berikut.

Tabel 4. Hasil Pembagian Data Training dan Testing

| Komposisi | Jumlah Data | |
|-----------|-------------|---------|
| | Training | Testing |
| 80% : 20% | 584 | 146 |

Setelah itu dilakukan proses menghitung masing-masing frekuensi kata yang terdapat pada data *training* untuk setiap kelas. Kemudian dihitung peluang kemunculan kata untuk setiap kelas menggunakan persamaan (2). Hasil perhitungan frekuensi dan peluang kata terlihat pada Tabel 5 berikut.

Tabel 5. Frekuensi dan Peluang Kemunculan Kata Untuk setiap Kategori

| No | Kata | Frekuensi | | Peluang | |
|----|-------------|-----------|---------|---------|---------|
| | | Positif | Negatif | Positif | Negatif |
| 1 | Mudah | 114 | 5 | 0.063 | 0.008 |
| 2 | Aplikasi | 274 | 86 | 0.151 | 0.109 |
| 3 | Investasi | 129 | 33 | 0.071 | 0.043 |
| 4 | Rekomendasi | 11 | 2 | 0.007 | 0.004 |
| 5 | Profil | 1 | 0 | 0.001 | 0.001 |
| 6 | Resiko | 5 | 0 | 0.003 | 0.001 |
| 7 | Bayar | 30 | 14 | 0.017 | 0.019 |
| 8 | Akun | 28 | 1 | 0.016 | 0.003 |

Selanjutnya dilakukan perhitungan peluang dengan nilai tertinggi dari setiap kategori untuk memprediksi kelas sentimen menggunakan persamaan (3) sebagai berikut.

$$\begin{aligned}
 P(\text{ulasan}|V_{\text{positif}}) &= P(V_{\text{positif}}) \times P(\text{Mudah}|V_{\text{positif}}) \times P(\text{Aplikasi}|V_{\text{positif}}) \times P(\text{Investasi}|V_{\text{positif}}) \times \\
 &\quad P(\text{Rekomendasi}|V_{\text{positif}}) \times P(\text{Profil}|V_{\text{positif}}) \times P(\text{Resiko}|V_{\text{positif}}) \times P(\text{Bayar}|V_{\text{positif}}) \times \\
 &\quad P(\text{Akun}|V_{\text{positif}}) \\
 &= 0.789 \times 0.063 \times 0.151 \times 0.071 \times 0.007 \times 0.001 \times 0.003 \times 0.017 \times 0.016 \\
 &= 3.38 \times 10^{-1} \\
 P(\text{ulasan}|V_{\text{negatif}}) &= P(V_{\text{negatif}}) \times P(\text{Mudah}|V_{\text{negatif}}) \times P(\text{Aplikasi}|V_{\text{negatif}}) \times P(\text{Investasi}|V_{\text{negatif}}) \times \\
 &\quad P(\text{Rekomendasi}|V_{\text{negatif}}) \times P(\text{Profil}|V_{\text{negatif}}) \times P(\text{Resiko}|V_{\text{negatif}}) \times P(\text{Bayar}|V_{\text{negatif}}) \times \\
 &\quad P(\text{Akun}|V_{\text{negatif}}) \\
 &= 0.211 \times 0.008 \times 0.109 \times 0.043 \times 0.004 \times 0.001 \times 0.001 \times 0.019 \times 0.003 \\
 &= 2.07 \times 10^{-18}
 \end{aligned}$$

Dari hasil perhitungan peluang setiap kelas didapatkan nilai peluang positif lebih besar dibandingkan peluang negatif, sehingga dapat dikatakan bahwa data *testing* kelas positif.

D. Ketepatan Klasifikasi

Untuk menghitung ketepatan klasifikasi digunakan *confusion matrix* dengan menghitung jumlah kuantitas kelas aktual dan kelas prediksi pada data *training*. Hasil *confusion matrix* dapat dilihat pada Tabel 6 berikut.

Tabel 6. Confusion Matrix

| Kelas Prediksi | Kelas Aktual | |
|----------------|--------------|---------|
| | Negatif | Positif |
| Negatif | 0 | 0 |
| Positif | 30 | 116 |

Kemudian dari nilai pada tabel *confusion matrix* diatas akan digunakan untuk menghitung akurasi menggunakan persamaan (4) sebagai berikut.

$$\text{Akurasi} = \frac{116 + 0}{116 + 0 + 0 + 30} \times 100\% = 79.45\%$$

Diperoleh nilai akurasi dalam mengklasifikasikan sentimen pengguna aplikasi Bibit yaitu sebesar 79.45% yang artinya model klasifikasi NBC memiliki kemampuan yang baik dalam mengklasifikasikan sentimen kelas positif dan negatif.

IV. KESIMPULAN

Berdasarkan pembahasan diatas mengenai sentimen aplikasi Bibit dengan metode *Naive Bayes Classifier* (NBC) untuk mengklasifikasikan sentimen ke kelas positif dan negatif, didapatkan sentimen pengguna dengan kelas positif sebesar 78.9% dan kelas negatif sebesar 21.1% yang menunjukkan bahwa pengguna aplikasi Bibit cenderung memberikan respon positif dan nilai ketepatan klasifikasi diperoleh sebesar 79.45% yang menunjukkan bahwa NBC memberikan performa kinerja yang baik dalam mengklasifikasikan teks. Diharapkan untuk penelitian selanjutnya lebih memperhatikan saat melakukan teks preprocessing dengan menambahkan tahap normalisasi sehingga data menjadi lebih baik dan meningkatkan nilai akurasi.

DAFTAR PUSTAKA

- Arsi, P., & Waluyo, R. 2021. Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine. *Jurnal Ilmu Teknologi dan Ilmu Komputer*. Vol. 8, No.1, pp. 147-152.
- Chairunnisa, C., Iin, E., & Mayanda, M. M. 2022. Klasifikasi Sentimen Ulasan Pengguna Aplikasi Peduli Lindungi di Google Play Menggunakan Algoritma SVM dengan Seleksi Fitur Chi Square. *Jurnal Informatika*. Edisi 18, No 1, April 2022.
- Kamal, M. F., & Apriani, R. 2022. Pengaruh Perkembangan Teknologi di Era Digital Terhadap Investasi dan Pasar Modal. *Jurnal Ilmu Hukum dan Humaniora*. Vol. 9, No. 1, pp. 488-496.
- Kamalia, A. Z., Andi, A. Z & Miftah, D. 2022. Analisis Sentimen Pada Ulasan Aplikasi Bibit di Playstore Dengan Metode Naïve Bayes, Support Vector Machine, C4.5 Dan KNN. *Jurnal Teknologi Pelita Bangsa*. Vol. 13, No. 1, Maret 2022, hml 2407-3903.
- Mahendrajaya, R., Ghulam, A. B., & Moh., B. S. Analisis Sentimen Pengguna GOPAY Menggunakan Metode Lexicon Based dan Support Vector Machine. *Jurnal Teknik Universitas Muhammadiyah Ponorogo*, 3(2), 2019:52-63.
- Pakpahan, D., & Hilda, W. 2014. Aplikasi Opinion Mining dengan Algoritma Naïve Bayes untuk Menilai Berita Online. *Jurnal Integrasi*. Vol. 6, No. 1, Maret 2014, hml. 1-10.
- Putra, A. D. A & Safitri, J. 2021. Analisis Sentimen Pada Ulasan Pengguna Aplikasi Bibit dan Bareksa dengan Algoritma KNN. *Jurnal Teknik Informatika dan Sistem Informasi*. Vol. 8, No. 2, Hal. 636-646.
- Routray, P., Swain, C. K. & Mishra, S. P. 2013. A Survey on Sentiment Analysis. *International Journal of Computer Applications*, Agustus, 70(10), pp. 1-8.

- Saputra, A. A., Fuji, A., & Mia, K. 2021. Analisis Sentimen Aplikasi Investasi Online di Google Play Store Menggunakan Metode Random Forest. *Skripsi*. Jurusan Sistem Informasi. Universitas Nusa Putra: Sukabumi.
- Saputro, I. W., & Sari, B. W. 2019. Uji Perfoma Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa. *Citec Journal*. Vol. 6, No.1.
- Tanggaeni, A. I., & Melkior, N. N. 2022. Analisis Sentimen Pengguna E-Government Pada Play Store Menggunakan NBC. *Jurnal Teknik Informatika dan Sistem Informasi*. Vol. 9, No. 2, Juni 2022, hml. 785-795.
- Wahyuni, E. D., Arifiyanti, A. A., & Afandi, I. M. 2020. *Klasifikasi Teks dengan Python*. Sidoarjo: Indomedika Pustaka.