JURNAL TEKNIK INFORMATIKA
Halaman Jurnal: http://jurnal.stmik-dci.ac.id/index.php/jutekin/
Halaman LPPM STMIK DCI : http://lppm.stmik-dci.ac.id/

SISTEM DETEKSI BERITA HOAX DENGAN ALGORITMA NAIVE BAYES DAN TEKNOLOGI EKSTRAKSI DATA OPTICAL CHARACTER RECOGNITION

Rifky Rahayu [1], Andri Sukmaindrayana [2]
Program Studi Teknik Informatika, STMIK DCI
Email : *rifkyrahayu@gmail.com [1], andrisukmaindrayana@gmail.com [2]*

ABSTRAK

Penyebaran berita hoaks melalui media digital menjadi tantangan serius karena dapat memengaruhi opini publik dan mengganggu ketertiban sosial. Untuk mengatasi masalah tersebut, penelitian ini mengembangkan sistem deteksi otomatis berita hoaks pada konten gambar berbasis teks dengan memanfaatkan teknologi Optical Character Recognition (OCR) untuk mengekstraksi teks dari gambar dan algoritma Naïve Bayes untuk melakukan klasifikasi. Teks hasil ekstraksi terlebih dahulu melalui tahap prapemrosesan, seperti normalisasi, penghapusan karakter non-alfabet, dan stopword removal, sebelum dianalisis. Berdasarkan pengujian terhadap 100 sampel gambar berita yang terdiri dari 50 berita valid dan 50 berita hoaks, sistem berhasil mencapai tingkat akurasi sebesar 80%, sehingga menunjukkan bahwa integrasi OCR dan Naïve Bayes efektif dalam mendeteksi berita palsu serta berpotensi menjadi alat bantu verifikasi informasi digital di media sosial.

Kata kunci: Deteksi hoaks, Optical Character Recognition, Naïve Bayes, klasifikasi teks, akurasi sistem, ekstraksi data.

I. PENDAHULUAN

1.1 Latar Belakang Penelitian

Kemajuan teknologi informasi yang berlangsung pesat telah membawa perubahan signifikan terhadap pola komunikasi serta cara manusia memperoleh informasi dalam kehidupan sehari-hari. Internet yang awalnya hanya bisa digunakan oleh sebagian orang, kini telah menjadi bagian tak terpisahkan dari kehidupan sehari hari. Memasuki awal tahun 2024, sebanyak 221,6 juta penduduk Indonesia setara dengan 79,5% dari total populasi telah tercatat sebagai pengguna internet. (Prisca Rizti 2024). Dengan meningkatnya penggunaan internet, muncul berbagai platform media sosial yang memungkinkan individu untuk berbagi informasi dengan mudah dan cepat.

Kehadiran media sosial telah merevolusi cara masyarakat berkomunikasi. Jika sebelumnya penyebaran informasi bergantung pada media konvensional seperti surat

kabar, televisi, dan radio, kini interaksi dan pertukaran informasi dapat berlangsung secara langsung melalui platform digital. Menurut data dari situs Telkomsel, pada tahun 2024 terdapat sekitar 139 juta pengguna media sosial di Indonesia, yang setara dengan 49,9% dari seluruh penduduk. Sebagian besar pengguna ini berada dalam kelompok usia 16 hingga 64 tahun.

Di era digital, media sosial menjadi sarana utama pertukaran informasi, namun juga mempermudah penyebaran hoaks atau informasi menyesatkan yang dapat menimbulkan kebingungan, kepanikan, dan berbagai dampak merugikan bagi masyarakat. Oleh karena itu, diperlukan upaya yang efektif dan strategis untuk mendeteksi serta mencegah penyebaran hoaks di media sosial guna menjaga kualitas dan keakuratan informasi yang beredar.

Hoaks merupakan informasi yang tidak dapat dipastikan kebenarannya dan berpotensi menimbulkan dampak negatif bagi masyarakat, sehingga diperlukan sistem yang mampu mendeteksi berita hoaks secara otomatis. Penelitian ini mengembangkan sistem deteksi hoaks dengan memanfaatkan teknologi Optical Character Recognition (OCR) untuk mengekstraksi teks dari gambar berita dan algoritma Naïve Bayes untuk mengklasifikasikan berita sebagai hoaks atau valid berdasarkan probabilitas kemunculan kata. Sistem yang diusulkan diharapkan dapat membantu masyarakat memverifikasi informasi serta mengurangi penyebaran berita palsu di media digital.

II. LANDASAN TEORI

2.1. Sistem Deteksi

Teknologi sistem deteksi berperan penting dalam menjaga keamanan dan pemantauan dengan kemampuan mengidentifikasi ancaman secara cepat dan akurat. Perkembangan kecerdasan buatan juga meningkatkan efektivitas sistem dalam mendeteksi manipulasi digital, seperti deepfake, sehingga membantu menjaga integritas dan keamanan informasi di era digital.

2.2 Berita Hoaxs

Misinformasi merupakan informasi yang tidak benar dan disebarkan tanpa mengetahui kesalahannya, sedangkan disinformasi adalah informasi yang sengaja dibuat untuk menyesatkan atau merugikan pihak lain. Pesatnya perkembangan media sosial telah mempermudah penyebaran informasi, namun juga meningkatkan risiko penyebaran hoaks yang dapat menyesatkan masyarakat dan mengganggu ketertiban publik.

2.3 Algoritma

Algoritma dapat diartikan sebagai serangkaian langkah yang dituliskan secara sistematis dan terurut guna menyelesaikan suatu permasalahan. Sementara itu, algoritma pemrograman merujuk pada langkah-langkah yang ditulis secara berurutan untuk menyelesaikan persoalan yang berkaitan dengan komputer.

2.4 Metode Naïve Bayes

Algoritma klasifikasi Naïve Bayes merupakan metode yang digunakan

untuk mengevaluasi sekumpulan data dengan pendekatan probabilistik. Dasar dari metode ini terletak pada penerapan Teorema Bayes, yakni konsep dalam statistik yang digunakan untuk memperkirakan kemungkinan suatu kejadian berdasarkan informasi sebelumnya. Naïve Bayes bekerja dengan menghitung peluang dari setiap kelas berdasarkan atribut-atribut yang dimiliki oleh data, lalu menentukan kelas mana yang paling mungkin sesuai. Tujuan akhir dari proses ini adalah menemukan nilai probabilitas tertinggi, sehingga data uji dapat dipetakan secara tepat ke dalam kategori yang relevan. (Agustiranti et al. 2024).

2.5 Optical Caracter Recognition (OCR)

Optical Character Recognition (OCR) merupakan teknologi komputer yang dirancang untuk mengenali dan membaca teks, baik yang dicetak menggunakan perangkat seperti printer dan mesin ketik, maupun yang ditulis tangan. Teknologi ini bekerja dengan mengonversi citra yang memuat karakter huruf menjadi teks digital, melalui pencocokan pola huruf per baris dengan pola-pola yang telah tersedia dalam basis data sistem. Output dari proses OCR berupa teks hasil interpretasi citra yang dipindai, di mana tingkat ketepatannya sangat dipengaruhi oleh kualitas visual dari gambar serta metode pengolahan karakter yang digunakan.(Setiawan, Sujaini, and Pn 2020)

III. ANALISIS SISTEM

3.1 Analisis Umum

Sistem deteksi berita hoaks ini dirancang untuk membantu pengguna memverifikasi informasi pada media sosial dengan memanfaatkan teknologi OCR untuk mengekstraksi teks dari gambar dan algoritma Naïve Bayes untuk mengklasifikasikan berita sebagai hoaks atau valid. Hasil analisis ditampilkan secara cepat dalam bentuk status klasifikasi dan nilai probabilitas, sehingga dapat digunakan oleh masyarakat sebagai alat bantu verifikasi informasi digital secara mudah dan semi-otomatis.

3.2 Analisis Masalah

Penyebaran berita hoaks di media sosial menjadi salah satu tantangan besar di era digital. Informasi yang tidak diverifikasi sering kali dibagikan ulang oleh pengguna tanpa disadari kebenarannya. Beberapa karakteristik utama penyebaran hoaks yang menjadi dasar permasalahan sistem ini antara lain:

1. Format berita tidak selalu berupa teks langsung, melainkan dalam bentuk gambar atau tangkapan layar.
2. Berita hoaks cenderung memprovokasi, bersifat sensasional, dan menggunakan judul-judul yang menarik perhatian.
3. Sebagian besar pengguna tidak memverifikasi kebenaran berita, terutama saat informasi sesuai dengan preferensi atau emosi pribadi.
4. Tidak semua pengguna memiliki akses atau kemampuan untuk mengecek berita secara manual melalui situs-situs verifikasi fakta.

Permasalahan tersebut menggaris bawahi perlunya sistem

otomatis yang dapat mendeteksi hoaks secara cepat dan efisien melalui analisis konten berita.

3.3 Analisis Data Masukan

3.3.1 Pengumpulan Dataset

Langkah pertama dalam membuat suatu model Mechin Learning adalah melakukan pengumpulan data yang di perlukan. Dataset di peroleh pada bulan mei 2025 dari situs webset CNN Indonesia, KOMDIGGI, KompasTV dan Turn back hoax. Setelah itu data di ubah dan di kumpulkan secara manual kedalam bentuk dokumen CSV

3.3.2 Preprocessing Dataset

Langkah selanjutnya yaitu Preprocessing data yang merupakan proses pengolahan data dimana data yang akan di proses mrupakan masih mentah untuk digunakan menjadi training data dan testing data dalam proses klasifikasi. Dimana pada tahap ini ada beberapa proses yaitu labeling data dan cleaning data.

a. Labeling Data

Data yang sudah terkumpul menjadi sebuh dataset akan diberikan labeling data atau pelabelan. Disini setiap feature dari dataset di berikan label atau kelas sesuai dengan kelasifikasi berita yaitu ada berita valid dan berita hoaxs. Pada tahap pra-pemrosesan data, salah satu proses penting yang dilakukan adalah mengubah label data dari bentuk teks menjadi bentuk numerik agar dapat diproses oleh algoritma pembelajaran mesin.

b. Cleaning Data

Prepocessing belum selesai pada labeling data, tahap selanjutnya yaitu cleaning data dimana data yang tidak diperlukan atau kososng/ null akan dihapus guna meningkatkan keaktifan algoritma naïve bayes juga menjadi parameter pengukur Tingkat ketepatan akurasi klasifikasi algoritma.

3.4 Analisis Data Proses

Analisis data proses dilakukan dengan mengolah dataset melalui tahap preprocessing sehingga menghasilkan data yang bersih dan siap digunakan untuk klasifikasi menggunakan algoritma Naïve Bayes. Dataset kemudian dibagi menjadi data latih dan data uji dengan perbandingan 80:20 menggunakan fungsi train_test_split() dari library sklearn Python, serta parameter random_state=42 untuk memastikan hasil pembagian konsisten. Proses ini menghasilkan data latih dan data uji beserta labelnya yang digunakan untuk melatih dan menguji model klasifikasi. Sebelum diklasifikasikan, data teks terlebih dahulu dikonversi menjadi data numerik menggunakan fitur-fitur yang tersedia pada library sklearn agar dapat diproses secara optimal oleh komputer.

3.5 Analisis Optical Caracter Recognition (OCR)

Teknologi Optical Character Recognition (OCR) digunakan sebagai tahap awal dalam sistem deteksi berita hoaks untuk mengekstraksi teks dari berbagai jenis gambar, seperti tangkapan layar artikel, unggahan media sosial, dan pamflet digital. Dengan memanfaatkan pustaka pytesseract, teks pada gambar dikonversi menjadi data digital yang kemudian diproses pada tahap pra-

pemrosesan dan klasifikasi. Penggunaan OCR memungkinkan sistem menganalisis informasi dalam format visual sehingga memperluas kemampuan deteksi hoaks yang banyak beredar melalui media sosial dalam bentuk gambar.

3.6 Evaluasi Sistem

Evaluasi menggunakan confusion matrix menunjukkan bahwa model berhasil mengklasifikasikan 8 berita valid dan 8 berita hoaks dengan benar, serta tidak menghasilkan kesalahan dalam mengidentifikasi berita hoaks sebagai valid. Namun, terdapat 4 berita valid yang salah diklasifikasikan sebagai hoaks. Hasil ini menunjukkan bahwa model memiliki kemampuan yang baik dalam mendeteksi hoaks, tetapi cenderung bersifat konservatif sehingga masih menghasilkan kesalahan pada klasifikasi berita valid yang perlu diperbaiki untuk meningkatkan keakuratan sistem.

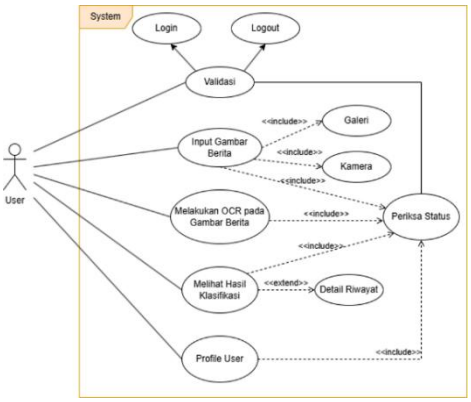
IV PERANCANGAN SISTEM

4.1 Perancangan Umum

Permasalahan utama dalam penelitian ini adalah mengklasifikasikan berita digital ke dalam kategori hoaks atau valid secara otomatis menggunakan algoritma Naïve Bayes. Sistem memanfaatkan dataset berita yang diproses melalui tahap preprocessing, seperti pembersihan teks, stemming, dan stopword removal, kemudian dilakukan ekstraksi fitur menggunakan metode TF-IDF untuk membangun model klasifikasi. Model yang dihasilkan diimplementasikan dalam aplikasi berbasis Android yang mampu mendeteksi status keabsahan berita, menampilkan label prediksi hoaks atau valid beserta tingkat keyakinan hasil klasifikasi, data OCR, dan teks yang telah diproses. Pengembangan sistem didukung dengan perancangan Use Case Diagram, Activity Diagram, dan Sequence Diagram sebagai dasar perancangan aplikasi.

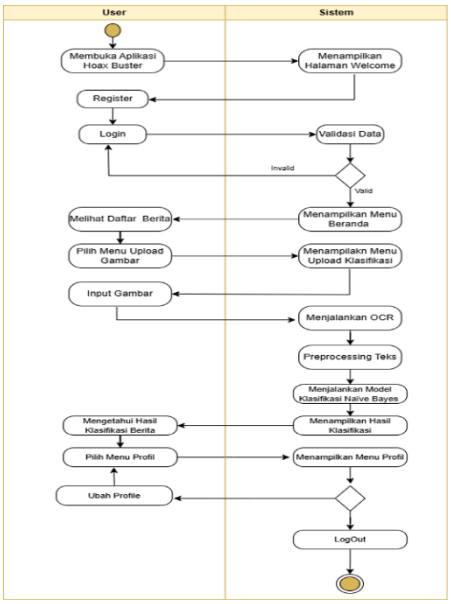
4.2 Rancangan Use Case Diagram

4.2.1 Use Case Diagram



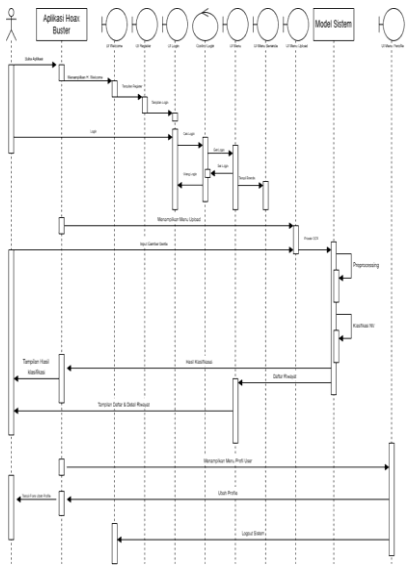
Gambar 4.1 Use Case Diagram Klasifikasi Berita Hoaxs

4.2.2 Rancangan Diagram Activity Pemrosesan Model Klasifikasi



Gambar 4.2 Diagram Activity Pemerosesan Model Klasifikasi

4.2.3 Rancangan Squence Diagram



Gambar 4.3 Sequance Diagram Klasifikasi Berita Hoax

V IMPLEMENTASI SISTEM

5.1 Konfigurasi Perangkat Lunak (Pengembang)

Berikut spesifikasi perangkat lunak yang menyususn pakai:

1. Bahasa Pemerograman Python
2. Basa Pemerograman Kotlin
3. Juyter Notebook
4. Android Studio
5. Postman

5.2 Konfigurasi Perangkat Keras (Pengembang)

Berikut spesifikasi perangkat keras yang menyusun pakai :

1. Procesor AMD Raizen5-5000 series
2. RAM 8 GB
3. Kapasitas Hardisk 500 GB
4. System Oprasi Windows 11 64 bit

5.3 Pedoman Pengoprasian Sistem

Sistem ini dirancang dengan antarmuka yang sederhana agar mudah digunakan oleh pengguna umum. Pengguna hanya perlu membuka aplikasi, login, lalu mengunggah gambar berita yang ingin diuji. Aplikasi akan memproses gambar tersebut melalui tahapan OCR untuk mengekstrak teks, kemudian dilanjutkan dengan klasifikasi menggunakan model machine learning Naive Bayes yang telah dilatih dengan dataset berita hoax dan valid.

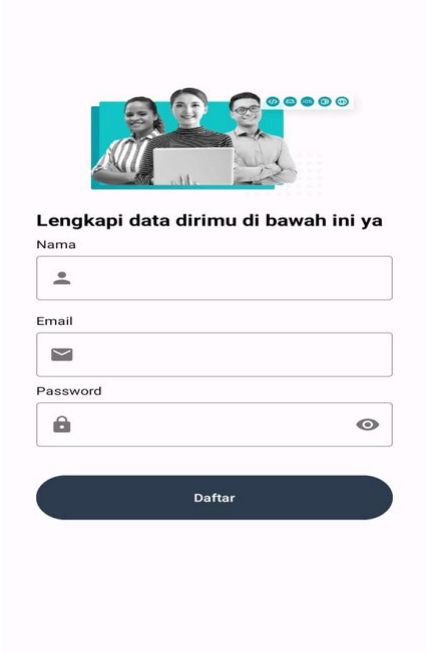
Hasil klasifikasi ditampilkan secara real-time kepada pengguna dengan penjelasan apakah berita tersebut terindikasi hoax atau valid. Selain itu, aplikasi juga menyimpan hasil klasifikasi ke dalam riwayat pengguna yang bisa diakses kembali melalui menu home. Sistem juga memprioritaskan keamanan data pengguna melalui Fitur Login.

5.4 Cara Menjalankan Program

1. Halaman Welcome

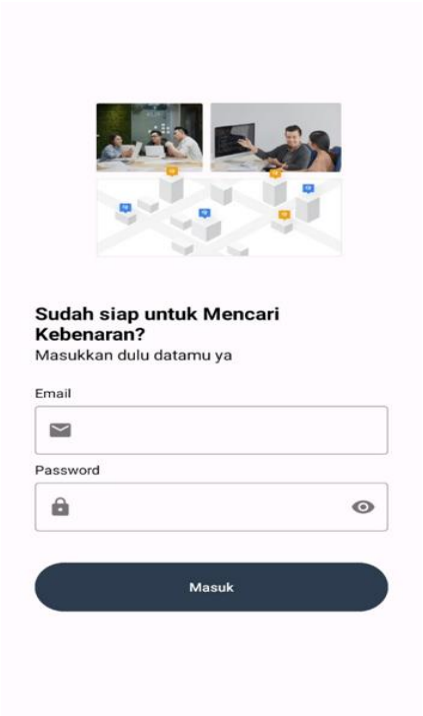
Gambar 5. 1Halaman Welcome

2. Halaman Register



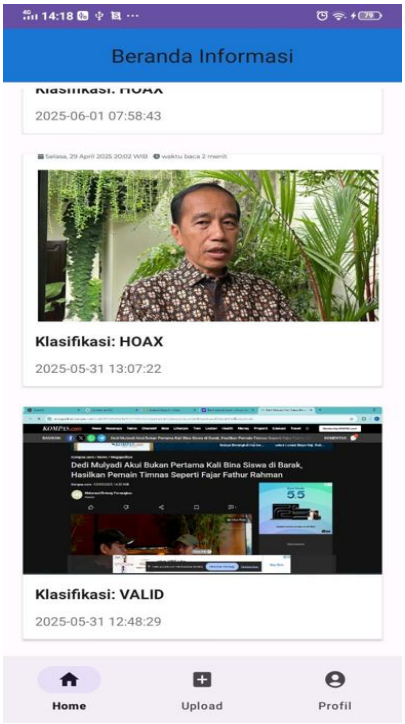
Gambar 5.2 Halaman Daftar

3. Halaman Login



Gambar 5.3 Halaman Login

4. Masuk Menu Beranda

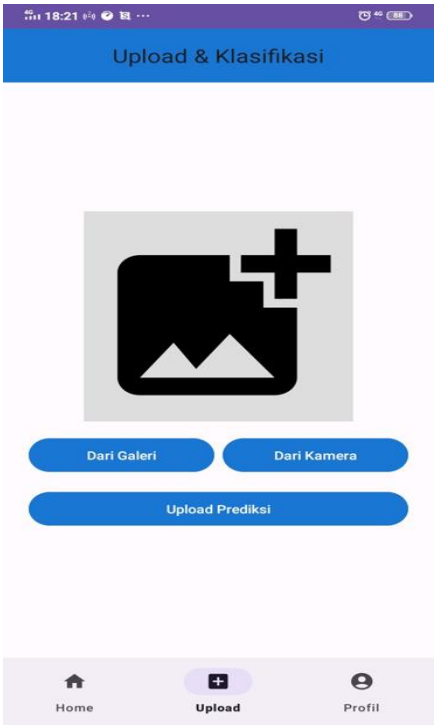


Gambar 5.4 Halaman Beranda

5. Melihat Detail Informasi

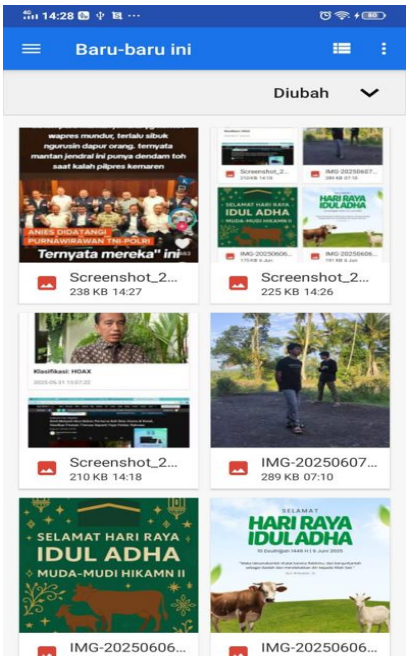
Gambar 5. 5Halaman Detail Informasi

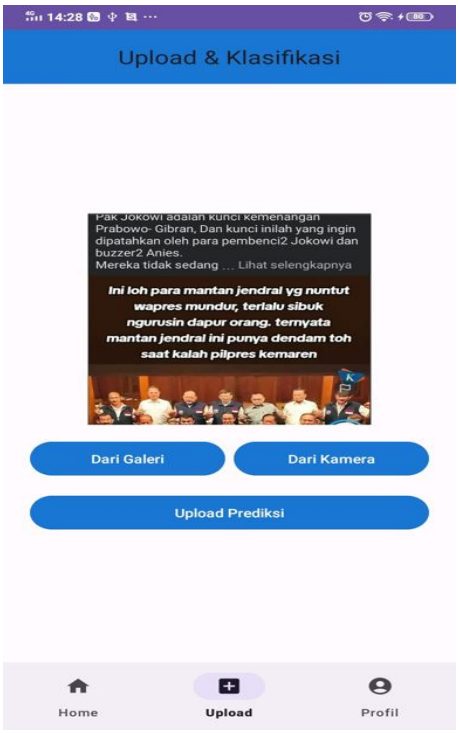
6. Upload Gambar Berita



Gambar 5.6 Halaman Detail Informasi

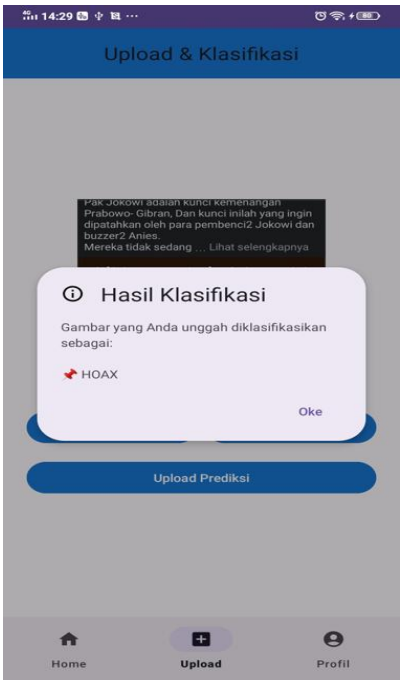
Berikut adalah Contoh gambar implementasi dari upload gambar dari galeri.





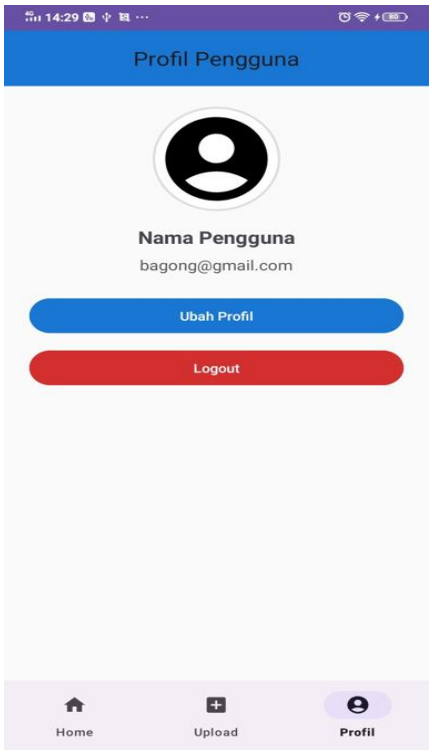
Gambar 5.7 Upload Gambar Dari Galeri

7. Menampilkan Hasil



Gambar 5.8 Hasil Klasifikasi Gambar

8. Akses Profil Pengguna



Gambar 5.9 Hasil Klasifikasi Gambar

VI KESIMPULAN

Berdsarkan uraian dari laporan sekripsi ini di ambil Kesimpulan sebagai berikut:

1. Penggunaan Model Machin Learning saat proses training dan testing menghasilkan akurasi rata-rata tertinggi dengan 80 % dengan model naïve bayes.
2. Proses Prapocessing pada dataset dan gambar input berita memberikan hasil prediksi klasifikasi yang tepat.
3. Menggunakan Fitur stopword untuk mengkasifikasi kata dalam gambar sehingga kata yang tidak memiliki makna secara langsung akan tereliminasi dan tidak akan di klasifikasi.
4. Implementasi model Machin Learning menggunakan sistem yang telah di bangun menjadi aplikasi mobile android dengan menerima inputan dari user untuk dilakukan proses klasifikasi dan koreksi terhadap berita yang ada pada suatu gambar yang termasuk ke dalam berita politik.
5. Sistem ini dapat digunakan sebagai alat bantu untuk mengetahui berita yang valid dan hoax, sehingga membantu mencegah terjadinya penyebaran disinformasi di public.

DAFTAR PUSTAKA

Agustiranti, Tifani, Aulia Khalfani Izzati Kurdiana, Bilal Al Ghiffari, Elza Dwi Juniar, and Diki Gita Purnama. 2024. “Penerapan Naive Bayes Terhadap Sentimen Analisis Media Sosial Twitter Pengguna Kereta Cepat Jakarta-Bandung (Whoosh).” Jurnal Ilmu Komputer Dan Sistem Informasi (JIKOMSI) 7 (1): 297–305. https://doi.org/10.55338/jikomsi.v7i1.2946.

Fardianto, Fariz. 2025. “Pakar Keamanan Digital Udinus Nova Rijati Pakai AI Deteksi Validasi Gambar.” IDN TIMES JATENG. 2025. https://jateng.idntimes.com/tech/gadget/fariz-fardianto/pakar-keamanan-digital-udinus-nova-rijati-pakai-ai-deteksi-validasi-gambar?utm_source=chatgpt.com.

I Gede Wilantara Jaya, I Ketut Gede Suhartana, and I Wayan Supriana. 2022. “Implementasi Rest Api Pada Aplikasi Admin Penjualan Banten.” Jurnal Pengabdian

Informatika 1 (1): 167–72. https://doi.org/10.24843/jupita.2022.v01.i01.p24.

Idris, Muhammad, and Amalia Rahmah. 2023. "Jurnal Informatika Terpadu." Jurnal Informatika Terpadu 9 (1): 34–39. https://journal.nurulfikri.ac.id/index.php/JIT.

Muttaqien, Faisal Azis, and Anang Dony Irawan. 2021. "Penerapan Hukum Pidana Penyebaran Berita Hoax Melalui Media Sosial Era Pandemi Covid-19." Media of Law and Sharia 2 (4): 305–15. https://doi.org/10.18196/mls.v2i4.12016.

Naufal Al Ghazali, Muhammad, Adnan Azizi, Ok Syahdan Khair, and Ziyan Tsabit Saifullah Kusnandar. 2024. "Pengembangan Dashboard Admin Bukupedia." Jurnal Teknik Indonesia 3 (1): 11–21. https://doi.org/10.58860/jti.v3i1.320.

Neighbor, K-nearest. 2025. "Klasifikasi Soal Menggunakan Multi-Label Problem Transformation Dengan Metode Random Forest Dan k-Nearest Neighbor" 10 (1): 367–80.

Nistrina, Khilda, and Anisa Rahmania. 2021. "Sistem Informasi Point of Sale Berbasis Website Studi Kasus: Pt Barokah Kreasi Solusindo (Artpedia)." Jurnal Sistem Informasi, J-SIKA 03 (02): 1–12. https://ejournal.unibba.ac.id/index.php/j-sika/article/view/687.

Agustiranti, Tifani, Aulia Khalfani Izzati Kurdiana, Bilal Al Ghiffari, Elza Dwi Juniar, and Diki Gita Purnama. 2024. "Penerapan Naive Bayes Terhadap Sentimen Analisis Media Sosial Twitter Pengguna Kereta Cepat Jakarta-Bandung (Whoosh)." Jurnal Ilmu Komputer Dan Sistem Informasi (JIKOMSI) 7 (1): 297–305. https://doi.org/10.55338/jikomsi.v7i1.2946.

Fardianto, Fariz. 2025. "Pakar Keamanan Digital Udinus Nova Rijati Pakai AI Deteksi Validasi Gambar." IDN TIMES JATENG. 2025. https://jateng.idntimes.com/tech/gadget/fariz-fardianto/pakar-keamanan-digital-udinus-nova-rijati-pakai-ai-deteksi-validasi-gambar?utm_source=chatgpt.com.

I Gede Wilantara Jaya, I Ketut Gede Suhartana, and I Wayan Supriana. 2022. "Implementasi Rest Api Pada Aplikasi Admin Penjualan Banten." Jurnal Pengabdian Informatika 1 (1): 167–72. https://doi.org/10.24843/jupita.2022.v01.i01.p24.

Idris, Muhammad, and Amalia Rahmah. 2023. "Jurnal Informatika Terpadu." Jurnal Informatika Terpadu 9 (1): 34–39. https://journal.nurulfikri.ac.id/index.php/JIT.

Muttaqien, Faisal Azis, and Anang Dony Irawan. 2021. "Penerapan Hukum Pidana Penyebaran Berita Hoax Melalui Media Sosial Era Pandemi Covid-19." Media of Law and Sharia 2 (4): 305–15. https://doi.org/10.18196/mls.v2i4.12016.