

**IMPLEMENTASI KLASIFIKASI KONTEN NEGATIF WEBSITE  
BERBASIS TEKS MENGGUNAKAN MODEL NAIVE BAYES, K-  
NEAREST NEIGHBORS , DAN SUPPORT VECTOR MACHINE**

**TUGAS AKHIR**

*Diajukan sebagai salah satu syarat untuk meraih gelar sarjana terapan (D-IV)  
pada Program Studi Teknologi Rekayasa Komputer Jaringan*

*Fakultas Teknologi Industri*

*Universitas Bung Hatta*



**Oleh:**

**REIHAN HANAFI WIYANATRA**

**2010017514007**

**PROGRAM STUDI TEKNOLOGI REKAYASA KOMPUTER JARINGAN  
FAKULTAS TEKNOLOGI INDUSTRI  
UNIVERSITAS BUNG HATTA  
PADANG  
2024**

LEMBAR PENGESAHAN  
IMPLEMENTASI KLASIFIKASI KONTEN NEGATIF WEBSITE  
BERBASIS TEKS MENGGUNAKAN MODEL NAÏVE BAYES, K-  
NEAREST NEIGHBOR DAN SUPPORT VECTOR MACHINE.

TUGAS AKHIR

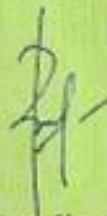
*Diajukan sebagai salah satu syarat untuk meraih gelar sarjana terapan (D-IV)  
pada Program Studi Teknologi Rekayasa Komputer Jaringan  
Fakultas Teknologi Industri  
Universitas Bung Hatta*

Oleh:

REIHAN HANAFI WIYANATRA

NPM: 2010017514007

Disetujui Oleh:  
Pembimbing



Riska Amelia, S.Kom, M.Kom

NIDN/NIP: 1001058601

Diketahui Oleh:

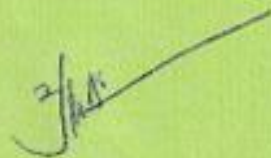
Fakultas Teknologi Industri  
Dekan,

Program Studi  
Teknologi Rekayasa Komputer Jaringan  
Ketua,



Prof. Dr. Eng. Ir. Reni Desmiarti, S.T., M.T.

NIK : 990 500 496



Zulfadli, S.Kom, M.Sc

NIDN: 1002058801



LEMBAR PENGUJI


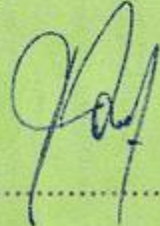
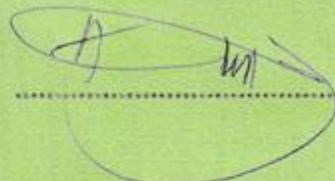
IMPLEMENTASI KLASIFIKASI KONTEN NEGATIF WEBSITE  
BERBASIS TEKS MENGGUNAKAN MODEL NAÏVE BAYES, K-  
NEAREST NEIGHBOR DAN SUPPORT VECTOR MACHINE

TUGAS AKHIR

REIHAN HANAFI WIYANATRA

NPM: 2010017514007

*Dipertahankan Di Depan Penguji Proposal  
Program Sarjana Terapan (D-IV)  
Pada Program Studi Teknologi Rekayasa Komputer Jaringan  
Fakultas Teknologi Industri  
Universitas Bung Hatta  
Hari: Jumat, 20 Agustus 2024*

NO.	Nama	Tanda Tangan
1	Riska Amelia, S.Kom, M.Kom (Ketua dan Penguji)	 .....
2	Dr. Hidayat, S.T, M.T, IPM (Penguji)	 .....
3	Ir. Arnita, MT (Penguji)	 .....

## ABSTRAK

Perkembangan teknologi informasi yang pesat telah memberikan kemudahan akses informasi, namun juga menimbulkan dampak negatif berupa penyebaran konten negatif seperti pornografi, perjudian, dan phishing. Penelitian ini berfokus pada pengembangan model klasifikasi konten negatif berbasis teks menggunakan algoritma machine learning, yaitu Naive Bayes, Support Vector Machine (SVM), dan K-Nearest Neighbors (KNN). Penelitian ini menggunakan metode kuantitatif eksperimental dengan dataset yang dikumpulkan dari repositori GitHub, yang kemudian diproses melalui tahapan web scraping dan pre-processing untuk membersihkan dan mempersiapkan data. Fitur-fitur penting dari setiap situs web diekstraksi menggunakan teknik TF-IDF, dan model-model machine learning dilatih menggunakan dataset yang telah diproses. Hasil penelitian menunjukkan bahwa model SVM mencapai akurasi tertinggi sebesar 82.24%, menunjukkan kemampuannya yang superior dalam mengklasifikasikan konten negatif dibandingkan dengan model Naive Bayes dan KNN. Sistem yang dikembangkan berhasil mengklasifikasikan konten negatif dengan tingkat keberhasilan 80%. Penelitian ini memberikan kontribusi signifikan dalam meningkatkan keamanan jaringan dan mengurangi risiko paparan konten negatif di internet dengan menyediakan model klasifikasi yang efektif dan akurat. Ke depannya, penelitian ini dapat ditingkatkan dengan memperluas dataset, menguji algoritma lain, dan melakukan pembersihan data yang lebih mendalam untuk mencapai tingkat akurasi yang lebih tinggi.

Kata kunci : Konten Negatif, Machine Learning, Web Scraping, Training, Confusion Matrix.

## **ABSTRACT**

*The rapid advancement of information technology has facilitated access to information but has also led to negative impacts such as the spread of harmful content like pornography, gambling, and phishing. This research focuses on developing a text-based negative content classification model using machine learning algorithms, specifically Naive Bayes, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). The study employs a quantitative experimental method with a dataset collected from GitHub repositories, which is then processed through web scraping and pre-processing stages to clean and prepare the data. Key features from each website are extracted using TF-IDF techniques, and the machine learning models are trained using the processed dataset. The results indicate that the SVM model achieved the highest accuracy of 82.24%, demonstrating its superior capability in classifying negative content compared to Naive Bayes and KNN models. The developed system successfully classified negative content with an 80% success rate. This research makes a significant contribution to enhancing network security and reducing the risk of exposure to harmful content on the internet by providing an effective and accurate classification model. Future work could improve this study by expanding the dataset, testing other algorithms, and performing more thorough data cleaning to achieve higher accuracy levels.*

*Keywords: Negative Content, Positive Content, Website, Machine Learning, Model.*

## KATA PENGANTAR

Alhamdulillah, puji syukur diucapkan kehadiran Allah SWT, yang telah memberikan limpahan rahmat dan hidayah-Nya kepada penulis, sehingga penulis dapat menyelesaikan penulisan proposal ini dengan judul “IMPLEMENTASI KLASIFIKASI KONTEN NEGATIF WEBSITE BERBASIS TEKS MENGGUNAKAN MODEL NAIVE BAYES, K-NEAREST NEIGHBORS, DAN SUPPORT VECTOR MACHINE” Tugas akhir ini merupakan salah satu syarat untuk memperoleh gelar Sarjana Terapan Teknik pada Jurusan Teknologi Rekayasa Komputer Jaringan Fakultas Teknologi Industri Universitas Bung Hatta Padang.

Dalam penyusunan tugas akhir ini ini, penulis banyak mendapatkan bantuan dan bimbingan serta pengarahan dari berbagai pihak, karena itu penulis menyampaikan ucapan terima kasih kepada :

1. Allah SWT yang telah mengizinkan hamba untuk menyelesaikan penulisan proposal ini serta memberikan kesehatan.
2. Kedua orang tua yang selalu mendo'akan saya, memberikan semangat serta memberikan nasehat kepada penulis.
3. Ibu Prof. Dr. Eng. Reni Desmiarti, ST.MT selaku Dekan Fakultas Teknologi Industri Universitas Bung Hatta Padang.
4. Bapak Zufadli, S.Kom, M.Sc, MTCNA selaku Kaprodi Teknologi Rekayasa Komputer Jaringan yang selalu memberi arahan terbaik.
5. Ibu Riska Amelia S.Kom, M.Kom selaku Dosen Pembimbing yang telah membimbing penulis selama ini sehingga dapat menyelesaikan penulisan proposal ini.
6. Bapak Budi Sunaryo, S.T., M.T yang telah memberikan saya kesempatan berharga untuk belajar dan berkontribusi dalam disertasi machine learning beliau.
7. Bapak/Ibu Dosen Program Studi Teknologi Rekayasa Komputer Jaringan Universitas Bung Hatta.
8. Terimakasih kepada adik tercinta yang telah memberikan penulis banyak saran terbaik dan semangat untuk penulis dalam berbagai hal apapun untuk penulisan proposal ini.

9. Terimakasih kepada teman-teman seperjuangan angkatan 20 dan seluruh angkatan TRKJ yang telah memberikan support dan saran dalam penulisan tugas akhir ini.

Penulis menyadari bahwa dalam menyelesaikan laporan ini, masih jauh dari kata sempurna, sehingga penulis memohon maaf.

Padang, Agustus 2024

Penulis,

## DAFTAR ISI

ABSTRAK .....	ii
KATA PENGANTAR.....	iii
DAFTAR ISI .....	v
DAFTAR GAMBAR .....	vii
DAFTAR TABEL.....	ix
BAB I PENDAHULUAN .....	1
1.1 Latar Belakang .....	1
1.2 Rumusan Masalah .....	3
1.3 Tujuan Penelitian.....	3
1.4 Batasan Masalah.....	4
1.5 Manfaat Penelitian.....	4
BAB II TINJAUAN PUSTAKA .....	<b>Error! Bookmark not defined.</b>
2.1 Tinjauan Pustaka.....	<b>Error! Bookmark not defined.</b>
2.2 Landasan Teori .....	<b>Error! Bookmark not defined.</b>
2.2.1 <i>Machine learning</i> .....	<b>Error! Bookmark not defined.</b>
2.2.2 <i>Artificial Intelligence</i> .....	<b>Error! Bookmark not defined.</b>
2.2.3 Klasifikasi .....	<b>Error! Bookmark not defined.</b>
2.2.4 <i>Naive Bayes</i> .....	<b>Error! Bookmark not defined.</b>
2.2.5 <i>K-Nearest Neighbors</i> .....	<b>Error! Bookmark not defined.</b>
2.2.6 <i>Support Vector Machine (SVM)</i> .....	<b>Error! Bookmark not defined.</b>
2.2.7 Hypertext Markup Language (HTML) .....	<b>Error! Bookmark not defined.</b>
2.2.8 <i>Web Scraping</i> .....	<b>Error! Bookmark not defined.</b>
2.2.9 Python .....	<b>Error! Bookmark not defined.</b>
BAB III METODOLOGI PENELITIAN.....	<b>Error! Bookmark not defined.</b>
3.1 Tahapan Penelitian.....	<b>Error! Bookmark not defined.</b>
3.2 Arsitektur Umum.....	<b>Error! Bookmark not defined.</b>
3.3 Perancangan Model .....	<b>Error! Bookmark not defined.</b>
3.4 Dataset .....	<b>Error! Bookmark not defined.</b>
3.5 <i>Web Scraping</i> .....	<b>Error! Bookmark not defined.</b>
3.5.1 Penambahan Protokol HTTP .....	<b>Error! Bookmark not defined.</b>
3.5.2 Request Website .....	<b>Error! Bookmark not defined.</b>



3.5.3	Get Raw Data .....	<b>Error! Bookmark not defined.</b>
3.6	Pre-processing Dataset .....	<b>Error! Bookmark not defined.</b>
3.6.1	Data Cleaning 1 .....	<b>Error! Bookmark not defined.</b>
3.6.2	Data Cleaning 2.....	<b>Error! Bookmark not defined.</b>
3.6.3	<i>Feature Extraction</i> .....	<b>Error! Bookmark not defined.</b>
3.7	<i>Training</i> Dataset .....	<b>Error! Bookmark not defined.</b>
3.7.1	Naïve Bayes Classifier .....	<b>Error! Bookmark not defined.</b>
3.7.2	<i>K-Nearest Neighbors</i> .....	<b>Error! Bookmark not defined.</b>
3.7.3	<i>Support Vector Machine</i> .....	<b>Error! Bookmark not defined.</b>
3.7.4	Klasifikasi dan Implementasi.....	<b>Error! Bookmark not defined.</b>
3.8	Alat dan Bahan Penelitian .....	<b>Error! Bookmark not defined.</b>
3.9	Rencana Anggaran Biaya .....	<b>Error! Bookmark not defined.</b>
BAB IV HASIL DAN PEMBAHASAN.....		<b>Error! Bookmark not defined.</b>
4.1	Penentuan Jumlah Dataset Sample.....	<b>Error! Bookmark not defined.</b>
4.2	Implementasi <i>Web Scraping</i> .....	<b>Error! Bookmark not defined.</b>
4.2.1	Google Colaboratory .....	<b>Error! Bookmark not defined.</b>
4.2.2	Proses <i>Scraping</i> .....	<b>Error! Bookmark not defined.</b>
4.3	Implementasi Website.....	<b>Error! Bookmark not defined.</b>
4.3.1	Halaman Utama.....	<b>Error! Bookmark not defined.</b>
4.3.2	Halaman Hasil Klasifikasi .....	<b>Error! Bookmark not defined.</b>
4.4	Pengujian Sistem.....	<b>Error! Bookmark not defined.</b>
4.4.1	Pengujian <i>Web Scraping</i> dan Pre-Processing pada Situs.....	<b>Error! Bookmark not defined.</b>
4.4.2	Pengujian <i>Feature Extraction</i> (TF-IDF).....	<b>Error! Bookmark not defined.</b>
4.4.3	Pengujian <i>Training</i> dan <i>Testing</i> .....	<b>Error! Bookmark not defined.</b>
4.4.4	Pengujian <i>Web Scraping</i> dan Pre-Processing pada Situs.....	<b>Error! Bookmark not defined.</b>
BAB V KESIMPULAN DAN SARAN.....		<b>Error! Bookmark not defined.</b>
5.1	Kesimpulan.....	<b>Error! Bookmark not defined.</b>
5.2	Saran.....	<b>Error! Bookmark not defined.</b>
DAFTAR PUSTAKA .....		<b>Error! Bookmark not defined.</b>

## DAFTAR GAMBAR

Gambar 2. 1 Algoritma <i>machine learning</i> .....	<b>Error! Bookmark not defined.</b>
Gambar 2. 2 Konsep kecerdasan buatan .....	<b>Error! Bookmark not defined.</b>
Gambar 2. 3 Klasifikasi sebagai pemetaan sebuah himpunan atribut input x ke dalam label kelasnya. ....	<b>Error! Bookmark not defined.</b>
Gambar 2. 4 Ilustrasi Naïve Bayes.....	<b>Error! Bookmark not defined.</b>
Gambar 2. 5 Ilustrasi model <i>K-Nearest Neighbors</i> .....	<b>Error! Bookmark not defined.</b>
Gambar 2. 6 Penentuan <i>Hyperplane</i> Terbaik .....	<b>Error! Bookmark not defined.</b>
Gambar 3. 1 Tahapan Penelitian .....	<b>Error! Bookmark not defined.</b>
Gambar 3. 2 Arsitektur Umum Sistem.....	<b>Error! Bookmark not defined.</b>
Gambar 3. 3 Model Pendeteksi Konten Website....	<b>Error! Bookmark not defined.</b>
Gambar 3. 4 Contoh Data Mentah situs <a href="http://www.example.com">www.example.com</a>	<b>Error! Bookmark not defined.</b>
Gambar 3. 5 Hasil data cleaning situs <a href="http://www.example.com">www.example.com</a> ..	<b>Error! Bookmark not defined.</b>
Gambar 3. 6 Contoh Ilustrasi Implementasi KNN.	<b>Error! Bookmark not defined.</b>
Gambar 3. 7 Contoh Ilustrasi Implementasi SVM.	<b>Error! Bookmark not defined.</b>
Gambar 3. 8 Skema proses implementasi sistem pendeteksi situs yang mengandung konten negatif .....	<b>Error! Bookmark not defined.</b>
Gambar 4. 1 Tampilan pengunduhan dataset pada repositori GitHub .....	<b>Error! Bookmark not defined.</b>
Gambar 4. 2 Skema Cara Kerja Protokol HTTP ....	<b>Error! Bookmark not defined.</b>
Gambar 4. 3 URL yang di-load dan telah diberi protokol HTTP melalui Google Colaboratory.....	<b>Error! Bookmark not defined.</b>
Gambar 4. 4 Hasil proses <i>Scraping</i> konten website dan ekspor data ke file CSV .....	<b>Error! Bookmark not defined.</b>
Gambar 4. 5 Dataset hasil <i>Scraping</i> dalam bentuk .csv .....	<b>Error! Bookmark not defined.</b>
Gambar 4. 6 Halaman dashboard .....	<b>Error! Bookmark not defined.</b>
Gambar 4. 7 Halaman situs blok jika terdeteksi konten negatif .	<b>Error! Bookmark not defined.</b>
Gambar 4. 8 Konten situs yang berhasil discrease ..	<b>Error! Bookmark not defined.</b>
Gambar 4. 9 Case Folding dan String Conversion.	<b>Error! Bookmark not defined.</b>
Gambar 4. 10 HTML Tag Removal .....	<b>Error! Bookmark not defined.</b>
Gambar 4. 11 URL Removal.....	<b>Error! Bookmark not defined.</b>
Gambar 4. 12 Non-Alphanumeric Symbol Removal	<b>Error! Bookmark not defined.</b>
Gambar 4. 13 Penghapusan Angka .....	<b>Error! Bookmark not defined.</b>
Gambar 4. 14 Non-ASCII Character Removal .....	<b>Error! Bookmark not defined.</b>
Gambar 4. 15 Stopwords removal.....	<b>Error! Bookmark not defined.</b>
Gambar 4. 16 Lemmatizer.....	<b>Error! Bookmark not defined.</b>

Gambar 4. 17 Grafik persebaran data per kategori dalam bentuk bar ..... **Error! Bookmark not defined.**

Gambar 4. 18 Unigram *Porn*    Gambar 4. 19 Unigram *Gambling* ..... **Error! Bookmark not defined.**

Gambar 4. 20 Unigram *Phishing*    Gambar 4. 21 Unigram *Whitelist*..... **Error! Bookmark not defined.**

Gambar 4. 22 Bigram *Porn*    Gambar 4. 23 Bigram *Gambling*. **Error! Bookmark not defined.**

Gambar 4. 24 Bigram *Phishing*    Gambar 4. 25 Bigram *Whitelist* ..... **Error! Bookmark not defined.**

Gambar 4. 26 Plot bar untuk 20 kata yang paling muncul paling banyak ..... **Error! Bookmark not defined.**

Gambar 4. 27 Plot bar untuk 20 kata yang paling populer digunakan pada situs *pornografi* ..... **Error! Bookmark not defined.**

Gambar 4. 28 Plot bar untuk 20 kata yang paling populer digunakan pada situs *gambling*..... **Error! Bookmark not defined.**

Gambar 4. 29 Plot bar untuk 20 kata yang paling populer digunakan pada situs *phishing* ..... **Error! Bookmark not defined.**

Gambar 4. 30 Plot bar untuk 20 kata yang paling populer digunakan pada situs *whitelist* ..... **Error! Bookmark not defined.**

Gambar 4. 31 Persebaran fitur pada setiap kategori ..... **Error! Bookmark not defined.**

Gambar 4. 32 Perbandingan Hasil pada Setiap Model *Machine learning*..... **Error! Bookmark not defined.**

Gambar 4. 33 *Confusion Matrix* Gaussian NB ..... **Error! Bookmark not defined.**

Gambar 4. 34 *Confusion Matrix* Multinomial NB ..... **Error! Bookmark not defined.**

Gambar 4. 35 *Confusion Matrix K-Nearest Neighbors* ..... **Error! Bookmark not defined.**

Gambar 4. 36 *Confusion Matrix OneVsRestClassifier* ..... **Error! Bookmark not defined.**

Gambar 4. 37 *Confusion Matrix SVC*..... **Error! Bookmark not defined.**

## DAFTAR TABEL

Tabel 3. 1 Contoh Dataset URL dari Github .....	<b>Error! Bookmark not defined.</b>
Tabel 3. 2 <i>Labeling</i> Data .....	<b>Error! Bookmark not defined.</b>
Tabel 3. 3 Rencana Anggaran Biaya .....	<b>Error! Bookmark not defined.</b>
Tabel 4. 1 Tabel unigram dan bigram dari masing-masing kategori .....	<b>Error! Bookmark not defined.</b>
Tabel 4. 2 Tabel skor tertinggi unigram dan bigram per kategori	<b>Error! Bookmark not defined.</b>
Tabel 4. 3 Hasil Ekstraksi Fitur TF IDF <a href="https://altomonlinecasino.dk/">https://altomonlinecasino.dk/</a> .....	<b>Error! Bookmark not defined.</b>
Tabel 4. 4 Hasil Persentase Akurasi <i>Training</i> Pada Setiap Model .....	<b>Error! Bookmark not defined.</b>
Tabel 4. 5 Classification Report GaussianNB .....	<b>Error! Bookmark not defined.</b>
Tabel 4. 6 Classification Report MultinomialNB	<b>Error! Bookmark not defined.</b>
Tabel 4. 7 Classification Report <i>K-Nearest Neighbors</i> .....	<b>Error! Bookmark not defined.</b>
Tabel 4. 8 Hasil Pengujian Sistem.....	<b>Error! Bookmark not defined.</b>





# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang

Saat ini teknologi informasi semakin berkembang, sangat cepat dan terutama di Indonesia, dengan kehadiran teknologi pada dasarnya adalah mempermudah pekerjaan orang lakukan sesuatu [1]. Perkembangan teknologi telah menjadi pendorong utama perubahan dalam berbagai aspek kehidupan manusia sepanjang sejarah. Internet dan jaringan komputer telah membuka pintu bagi pertukaran informasi yang cepat dan luas di seluruh dunia. Kemajuan teknologi memiliki dampak yang signifikan pada berbagai bidang kehidupan, memungkinkan manusia untuk melakukan berbagai aktivitas dengan lebih mudah [2]. Salah satu contohnya adalah kemudahan dalam berkomunikasi antar individu, baik dekat maupun jauh, meskipun berada pada jarak yang jauh, dengan adanya teknologi memungkinkan interaksi tanpa terpengaruh oleh batasan geografis.

Perkembangan teknologi dalam bidang komunikasi telah memberikan dampak signifikan dalam kehidupan manusia. Berbagai perangkat seperti laptop, smartphone, tablet, dan PC, dilengkapi dengan akses internet, telah memudahkan manusia untuk mengakses informasi di mana pun dan kapan pun. Namun, dampaknya tidak hanya positif, melainkan juga negatif. Contohnya, kemudahan akses informasi dan komunikasi bisa meningkatkan efisiensi, tetapi juga dapat menyebabkan ketergantungan yang berlebihan dan perilaku malas. Selain itu, internet juga dapat disalahgunakan untuk aktivitas konten negatif seperti mengakses konten *porn* dan lainnya.

Konten negatif, seperti ujaran kebencian, pornografi, intimidasi, judi online, dan penyebaran informasi palsu, semakin menjadi perhatian karena dapat dengan mudah menyebar di berbagai platform daring. Fenomena ini menciptakan lingkungan yang tidak aman dan berpotensi merugikan bagi pengguna. Terpaparnya pengguna oleh konten-konten negatif tersebut dapat menyebabkan gangguan emosional, kehilangan kepercayaan diri, serta meningkatkan risiko menjadi korban penipuan atau tindakan kriminal. Selain itu, konten negatif juga dapat memicu konflik antar individu atau kelompok, dan merusak kepercayaan masyarakat pada

informasi yang disajikan secara daring. Oleh karena itu, penting untuk mencari solusi yang cerdas dan efektif dalam mendeteksi serta mengelola konten semacam itu untuk menjaga keamanan dan kenyamanan pengguna di ruang daring.

Berdasarkan data dari Kementerian Komunikasi dan Informatika, Indonesia menerima pengaduan dari masyarakat sebanyak 16.370 aduan terkait konten negatif per Maret 2022. Tiga klasifikasi pengaduan tertinggi yaitu judi sebanyak 10.921 aduan. Kemudian diikuti oleh 5071 aduan konten *porn* dan 203 aduan penipuan.

Oleh karena itu penelitian ini akan menggunakan beberapa metode *machine learning* yaitu *Naive Bayes*, *Support Vector Machine* (SVM), dan *K-Nearest Neighbors* (KNN), dengan itu perlu adanya pendekatan proaktif dalam bentuk penelitian mengenai pendeteksian konten negatif website berdasarkan tiga klasifikasi yaitu *gambling*, *porn*, dan *phishing*, serta konten positifnya adalah whitelist. Dataset untuk kebutuhan *training* model sistem diambil dari <https://github.com/olbat/ut1-blacklists> (kontributor: obalt dan dependabot). Urgensi penelitian ini sangat penting karena konten negatif dapat merusak pola pikir pengguna layanan internet, juga dapat disisipi *malicious software* (*malware*). Tentu hal ini dapat mengganggu keamanan dan kenyamanan pengguna lain dari jaringan yang sama.

Penerapan dengan *machine learning* terbukti efektif untuk identifikasi permasalahan dalam berbagai bidang seperti pada penelitian yang dilakukan oleh Saputri dkk [3] tentang penggunaan *machine learning* untuk mendeteksi situs web saat proses *training* dan *testing*, dapat memberikan akurasi rata-rata tertinggi dengan nilai akurasi rata-rata mencapai lebih dari 96% menggunakan algoritma *Support Vector Machine*. Kemudian penelitian yang dilakukan oleh Sandag dkk [4], menemukan bahwa algoritma *K-Nearest Neighbors* (K-NN) menghasilkan performa yang sangat baik dalam mendeteksi *malicious* dan *benign website*, dengan akurasi sebesar 95.51%. Lalu penelitian yang dilakukan oleh Lee dkk [5] tentang menggunakan *machine learning* untuk mengklasifikasi *malicious* URL yang menggunakan *Naive Bayes* dan *Support Vector Machine* untuk metode klasifikasi dalam, penelitian ini mendapatkan nilai akurasi 99%.

Mengingat begitu pentingnya penerapan *machine learning* pada sistem informasi, maka peneliti mengangkat penelitian ini dengan judul ” IMPLEMENTASI KLASIFIKASI KONTEN NEGATIF WEBSITE BERBASIS TEKS MENGGUNAKAN *NAIVE BAYES*, *K-NEAREST NEIGHBORS* , DAN *SUPPORT VECTOR MACHINE* ” sebagai pendekatan utama. Penelitian ini akan difokuskan pada empat klasifikasi utama, yaitu *gambling*, *porn*, *phishing*, dan *whitelist* dengan menggunakan dataset yang tersedia untuk melatih model sistem. Urgensi penelitian ini tidak dapat diabaikan karena konten negatif dapat mengganggu keamanan dan kenyamanan pengguna internet, serta berpotensi menyebarkan *malware*. Dengan demikian untuk menjawab penelitian ini dapat memberikan kontribusi dalam menciptakan lingkungan internet yang lebih aman dan positif bagi semua pengguna.

## **1.2 Rumusan Masalah**

1. Bagaimana mengidentifikasi dan mengklasifikasikan konten negatif pada website dengan menggunakan pendekatan *machine learning*?
2. Apa saja karakteristik dan pola yang dapat dijadikan acuan untuk pengembangan model pendeteksi konten negatif?
3. Sejauh mana efektivitas model klasifikasi dalam meningkatkan keamanan jaringan dan mengurangi risiko paparan konten negatif?

## **1.3 Tujuan Penelitian**

1. Mengembangkan model yang mampu mendeteksi dan mengklasifikasikan konten website secara *real time*, sehingga dapat mengidentifikasi dan mengelompokkan konten berdasarkan kategori yang telah ditentukan
2. Mengevaluasi dan membandingkan efektivitas model yang digunakan, yaitu *Naive Bayes*, *K-Nearest Neighbors* , dan *Support Vector Machine*
3. Meningkatkan keamanan jaringan komputer dengan cara mengidentifikasi dan memblokir konten website yang tidak pantas atau berbahaya, guna melindungi pengguna dari potensi risiko.

#### 1.4 Batasan Masalah

1. Data yang digunakan untuk melatih model sistem klasifikasi konten negatif dibatasi hanya pada dataset yang tersedia dari URL <https://github.com/olbat/ut1-blacklists>. Dataset ini dikontribusikan oleh olbat dan dependabot, yang menyediakan basis data untuk pelatihan yang spesifik dan terkontrol.
2. Sistem klasifikasi konten negatif ini dibatasi hanya untuk analisis konten berbasis teks, sehingga tidak mencakup analisis konten multimedia seperti gambar atau video.
3. Total data yang akan digunakan adalah lebih dari 2500 data
4. Bahasa pemrograman yang digunakan untuk membuat model *Machine learning* penelitian adalah menggunakan bahasa *Python* versi 3.10

#### 1.5 Manfaat Penelitian

1. Penelitian ini akan menyediakan solusi yang dapat meningkatkan keamanan online dengan mendeteksi konten negatif secara otomatis, membantu mengurangi risiko paparan terhadap konten yang merugikan atau berbahaya.
2. Dengan menggunakan model pendeteksi konten negatif yang efektif, pengguna internet akan lebih dilindungi dari konten yang tidak pantas, mengurangi kemungkinan mereka terpapar kepada informasi yang tidak diinginkan atau potensial merugikan.
3. Penelitian ini akan mendorong pengembangan teknologi keamanan yang inovatif, seperti model *machine learning* untuk klasifikasi konten negatif.
4. Dengan adanya pendekatan otomatis untuk mengklasifikasi konten negatif, penelitian ini dapat mengurangi beban kerja manual dalam pengawasan konten online.
5. Hasil penelitian ini dapat memberikan kontribusi pada pembentukan kebijakan publik terkait dengan keamanan internet dan juga meningkatkan kesadaran pengguna internet akan pentingnya perlindungan terhadap konten negatif online.