BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 361-371
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.586

Analisis Perbandingan Algoritma SVM dan CNN dalam Mendeteksi Website Judi Online Berdasarkan Konten Teks

Nurcahaya Simanjuntak*, Alva Hendi Muhammad

Magister Informatika, Universitas AMIKOM Yogyakarta, Sleman, Indonesia
Email: [1,*]nurcahaya@students.amikom.ac.id, [2]alva@amikom.ac.id
Email Penulis Korespondensi: nurcahaya@students.amikom.ac.id

Abstrak−Penelitian ini bertujuan untuk membandingkan efektivitas algoritma *Support Vector Machine* (SVM) dan *Convolutional Neural Network* (CNN) dalam mendeteksi website perjudian online berbahasa Indonesia. Dengan meningkatnya jumlah pemain judi online di Indonesia, penting untuk mengembangkan metode yang efektif dalam identifikasi konten perjudian. Dataset yang digunakan terdiri dari 24.336 website judi dan 36.529 website non-judi yang dikumpulkan melalui *web scraping*. Model SVM menunjukkan akurasi sebesar 99%, dengan metrik evaluasi presisi 1.00, *recall* 0.99, dan F1-*score* 0.99. Di sisi lain, model CNN mencapai akurasi sempurna 100%, dengan presisi, *recall*, dan F1-*score* masing-masing sebesar 1.00. Namun, penting untuk dicatat bahwa akurasi sempurna ini dicapai dalam kondisi tertentu, termasuk dataset yang relatif bersih dan proses pelatihan yang optimal. Hasil evaluasi model menggunakan teknik *cross-validation* menunjukkan bahwa SVM konsisten dengan akurasi sekitar 99%, sedangkan CNN memiliki akurasi rata-rata 99.61% dengan deviasi standar yang sangat rendah. Penelitian ini menekankan pentingnya proses *pre-processing* data teks dalam meningkatkan akurasi model dan menunjukkan keunggulan CNN dalam menangkap pola kompleks dalam teks. Temuan ini memberikan kontribusi signifikan terhadap pengembangan metode deteksi situs web perjudian online di Indonesia, serta membuka ruang untuk penelitian lebih lanjut dalam bidang ini.

Kata Kunci: Deteksi Website; Judi Online; *Support Vector Machine*; *Convolutional Neural Network*; *Pre-processing* Teks.

Abstract−This study aims to compare the effectiveness of Support Vector Machine (SVM) and Convolutional Neural Network (CNN) algorithms in detecting Indonesian-language online gambling websites. With the increasing number of online gambling players in Indonesia, it is essential to develop effective methods for identifying gambling content. The dataset used consists of 34,336 gambling websites and 36,529 non-gambling websites, collected through web scraping. The SVM model demonstrated an accuracy of 99%, with evaluation metrics including a precision of 1.00, recall of 0.99, and F1-score of 0.99. In contrast, the CNN model achieved perfect accuracy of 100%, with precision, recall, and F1-score all at 1.00. However, it is important to note that this perfect accuracy was achieved under certain conditions, including a relatively clean dataset and optimal training processes. Evaluation results using cross-validation techniques indicated that SVM maintained a consistent accuracy of approximately 99%, while CNN exhibited an average accuracy of 99.61% with a very low standard deviation. This research emphasizes the importance of data pre-processing in enhancing model accuracy and highlights the advantages of CNN in capturing complex patterns within text. These findings contribute significantly to the development of detection methods for online gambling websites in Indonesia and open avenues for further research in this field.

Keywords: Website Detection; Online Gambling; Support Vector Machine; Convolutional Neural Network; Text Pre-processing.

1. PENDAHULUAN

Perkembangan teknologi informasi telah mengubah judi tradisional menjadi judi online yang lebih mudah diakses melalui *smartphone*, komputer, dan *tablet* [1]. Judi online dapat diakses kapan saja dan di mana saja, meningkatkan risiko kecanduan psikologis di kalangan penjudi [2]. Selain dampak psikologis, judi online juga berdampak negatif pada aspek ekonomi dan sosial, bahkan menyebabkan korban jiwa [3]. Pada tahun 2023, PPATK mencatat perputaran uang judi online di Indonesia mencapai Rp 327 triliun [4]. Menurut Drone Emprit, Indonesia memiliki 201.122 pemain judi online terbanyak di dunia [5]. Dari Juli 2023 hingga Oktober 2024, Kominfo juga melaporkan sebanyak 3.796.902 konten judi online telah dihapus, serta banyak sisipan laman judi di situs pendidikan dan pemerintahan [6].

Dalam konteks meningkatnya jumlah situs web judi online di Indonesia, deteksi website judi merupakan tantangan yang krusial. Selain itu, pengembangan penelitian terkait deteksi website judi berbahasa Indonesia juga belum banyak dilakukan. Penelitian sebelumnya menunjukkan bahwa penggunaan algoritma berbasis *machine learning* (ML) dan *deep learning* (DL) memberikan hasil yang menjanjikan dalam mendeteksi situs judi online.

Penelitian [7]–[11] menggunakan algoritma DL untuk mendeteksi situs *web phishing*, judi dan pornografi dengan masing-masing hasil akurasi diatas 97%. Kemudian, penelitian [1], [12] menggunakan algoritma ML dalam mendeteksi website perjudian dan pornografi dengan hasil akurasi yang juga diatas 97%. Selain itu, penelitian [13]–[15] membandingkan kinerja algoritma ML dan DL dalam deteksi berbasis teks untuk menentukan metode yang paling efektif. Oleh karena itu, penelitian ini akan menggunakan algoritma ML dan juga DL untuk mendeteksi situs web perjudian online berbahasa Indonesia dengan melakukan perbandingan hasil algoritma.

Penelitian ini mengimplementasikan algoritma SVM dan CNN untuk mendeteksi website judi online berbahasa Indonesia. Penelitian ini bertujuan memberikan wawasan tentang kinerja masing-masing algoritma. SVM dikenal karena kemampuannya dalam menangani data yang tidak linear dan ukuran dataset yang relatif kecil. SVM bekerja dengan mencari *hyperplane* optimal yang memisahkan data ke dalam kelas yang berbeda. Hal ini membuat SVM sangat efektif dalam situasi di mana data sulit dipisahkan secara linear dan ketika jumlah data terbatas. SVM juga tahan terhadap *overfitting* [16], terutama ketika digunakan dengan kernel trik yang tepat, menjadikannya pilihan yang andal untuk berbagai tugas klasifikasi. Di sisi lain, CNN yang arsitektur awalnya dirancang untuk menangani pemrosesan data gambar, tetapi juga telah berhasil digunakan dalam berbagai tugas pemrosesan teks, termasuk klasifikasi teks [17]. CNN

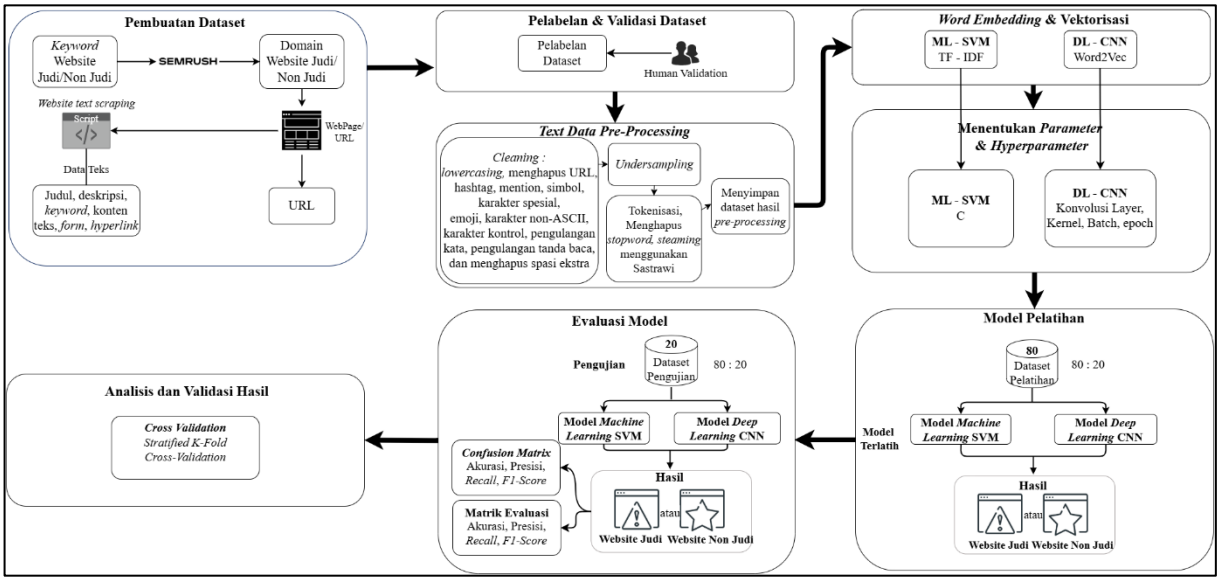
BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 361-371
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.586

memiliki kemampuan untuk menangkap pola kompleks dalam teks. Algoritma ini menggunakan lapisan konvolusi yang dapat mendeteksi fitur penting dalam data teks dengan cara yang mirip dengan cara kerjanya dalam pemrosesan gambar. Dalam konteks pemrosesan teks, CNN mampu mengeksplorasi hubungan antar kata yang tersebar dalam dokumen dan mengenali pola-pola yang lebih kompleks, yang sangat berguna dalam tugas klasifikasi seperti mendeteksi situs web perjudian. Meskipun terdapat perbedaan terhadap arsitektur algoritma SVM dan CNN, keduanya memiliki keunggulan dalam menangani tugas pemrosesan teks dengan efektif. Penelitian [15], [18] membandingkan kinerja algoritma SVM dan CNN terhadap tugas pemrosesan teks yang menunjukkan bahwa keduanya dapat diandalkan dalam berbagai konteks aplikasi.

Penelitian ini bertujuan untuk mengeksplorasi keunggulan masing-masing algoritma dalam mendeteksi situs web perjudian online dan menemukan pendekatan yang paling efektif. Peneliti menekankan pentingnya penggunaan dataset seragam dan langkah-langkah *pre-processing* yang konsisten serta pengaturan parameter yang sesuai untuk memastikan hasil evaluasi algoritma yang adil dan dapat diandalkan. Evaluasi efektivitas kedua algoritma dilakukan dengan penerapan metrik evaluasi serta validasi model yang seragam. Diharapkan, penelitian ini dapat menjadi model deteksi situs web perjudian online berbahasa Indonesia yang akurat dan memberikan panduan bagi pengembangan model serupa di masa mendatang. Dengan membandingkan kinerja kedua algoritma ini secara menyeluruh, peneliti bertujuan menemukan solusi optimal untuk mengatasi tantangan dalam mendeteksi konten berbahaya secara efektif, khususnya situs web perjudian online berbahasa Indonesia.

2. METODOLOGI PENELITIAN

Metode penelitian ini disusun untuk menganalisis perbandingan algoritma SVM dan CNN dalam mendeteksi website judi online berdasarkan konten teks. Proses penelitian melibatkan beberapa tahapan, mulai dari pembuatan dataset, pelabelan dan validasi dataset, *text pre-processing*, vektorisasi, penentuan parameter dan *hyperparameter*, pembuatan model dengan SVM dan CNN, pelatihan dan pengujian model, hingga analisis dan validasi hasil seperti yang ditunjukkan pada ambar 1.



Gambar 1. Diagram Metode Penelitian

2.1 Pembuatan Dataset

Pengumpulan data dilakukan secara independen oleh peneliti dengan membuat sebuah dataset melalui proses *web scraping*. Tool SEMrush juga digunakan untuk mencari kata kunci terkait website judi dan non judi yang relevan dalam bahasa Indonesia.

2.2 Pelabelan dan Validasi Dataset

Pelabelan dataset dilakukan dengan memberikan label "1" untuk website judi, sedangkan lebel "0" untuk website non judi berdasarkan kata kunci yang digunakan. Untuk memastikan bahwa pelabelan dataset dilakukan dengan benar oleh mesin, peneliti melakukan validasi terhadap dataset dengan melibatkan bantuan manusia atau pihak lain untuk memeriksa isi konten teks dalam dataset.

2.3 *Text Pre-Processing*

Text pre-processing adalah langkah untuk memperbaiki dan membersihkan dataset guna meningkatkan kualitas data teks, menghilangkan *noise*, dan membuat data siap untuk analisis atau pemodelan [7]. Dalam penelitian ini, proses *pre-*

ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 361-371
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.586

processing dilakukan dengan pembersihan data (*cleaning*), *undersampling*, penghapusan tokenisasi *stop words*, dan *stemming*.

a. *Cleaning* adalah proses penghapusan elemen-elemen yang tidak diperlukan untuk mengurangi *noise* dalam data yang dapat mengganggu analisis [19]. Menghilangkan informasi yang tidak relevan dan tidak diperlukan dapat membuat kualitas data meningkat, sehingga model dapat lebih fokus pada informasi signifikan. Selain itu, proses ini membantu meningkatkan akurasi hasil analisis dan pemodelan dengan meminimalkan potensi kesalahan yang disebabkan oleh data yang tidak bersih atau tidak terstandarisasi.
b. *Undersampling* adalah teknik untuk menyeimbangkan dataset yang tidak seimbang dengan mengurangi jumlah sampel dari kelas mayoritas agar setara dengan kelas minoritas. Teknik ini penting untuk mencegah bias model terhadap kelas yang dominan [20]. Dengan menjaga keseimbangan distribusi data, model dapat lebih fokus pada deteksi kelas minoritas yang signifikan. Selain itu, *undersampling* meningkatkan kinerja model dengan memastikan bahwa model tidak mengabaikan informasi penting dari kelas yang lebih kecil, sehingga menghasilkan deteksi yang lebih akurat dan adil. Pada penelitian ini, dataset non-judi lebih besar daripada dataset judi sehingga memerlukan teknik *undersampling* dalam penerapannya agar distribusi dataset seimbang.
c. Tokenisasi adalah proses memecah teks menjadi unit-unit yang lebih kecil, yang disebut token. Token ini biasanya berupa kata, frasa, atau bahkan karakter, tergantung pada tujuan analisis teks yang dilakukan [21].
d. Penghapusan *stop words* adalah teknik penghapusan kata-kata umum yang tidak memberikan nilai analitis signifikan, seperti "dan", "yang", dan "di", sehingga hasil analisis lebih fokus pada kata-kata penting secara kontekstual [19].
e. *Stemming* adalah proses linguistik untuk mengurangi kata-kata ke bentuk dasarnya atau akarnya; misalnya, kata "berjalan" diubah menjadi "jalan".Tujuan utama stemming adalah menghilangkan imbuhan seperti awalan, akhiran, atau sisipan dari kata-kata sehingga berbagai bentuk kata yang memiliki makna dasar yang sama dapat disederhanakan menjadi satu bentuk dasar [21].

2.4 Vektorisasi

Vektorisasi teks adalah sebuah langkah untuk mengubah teks menjadi representasi numerik agar dapat diproses oleh model [22]. Dalam penelitian ini, penulis membedakan teknik vektorisasi untuk masing-masing algoritma berdasarkan kemampuannya dalam memanfaatkan informasi dari data teks secara efektif. Pemilihan teknik vektorisasi sangat penting karena mempengaruhi keselarasan dengan algoritma yang dipilih, sehingga diperlukan untuk mencapai hasil deteksi yang optimal.

2.5 Parameter dan *Hyperparameter*

Penentuan parameter dan *hyperparameter* sangat penting dalam pengembangan model untuk mengoptimalkan kinerjanya, mencegah *overfitting* dan *underfitting*, serta meningkatkan generalisasi. Parameter adalah nilai yang dipelajari oleh model selama pelatihan, menentukan fungsi dan perilaku model terhadap data. Sementara itu, *hyperparameter* adalah nilai yang ditentukan sebelum pelatihan dan tidak dipelajari dari data. Dalam penelitian ini, SVM menggunakan *hyperparameter* C, sementara CNN mengatur konvolusi *layer*, *batch size*, dan *epochs*. Nilai-nilai ini dipilih untuk memastikan model bekerja secara akurat tanpa mengorbankan efisiensi penggunaan sumber daya.

2.6 Penerapan Algoritma SVM dan CNN

Algoritma deteksi merupakan metode atau teknik yang digunakan dalam mengidentifikasi website judi online berdasarkan analisis konten teks. Pada tahap ini, algoritma ML dan DL diterapkan pada data teks yang telah diolah untuk melatih model deteksi yang dapat mengidentifikasi website judi atau non judi. Algoritma yang akan digunakan untuk perbandingan kinerja pada penelitian ini adalah algoritma ML yaitu *Support Vector Machine* (SVM) dan algoritma DL yaitu *Convolutional Neural Network* (CNN).

Dalam proses kerjanya, algoritma ML dan DL memiliki kinerja yang berbeda. Oleh karena itu, perlu dilakukan beberapa pendekatan untuk memastikan bahwa perbandingan antara kedua algoritma yang berbeda tersebut dapat dilakukan dengan adil dan setara antara satu dan lainnya. Untuk memastikan bahwa perbandingan antara SVM dan CNN dilakukan secara adil, maka penting untuk menjaga keseragaman dalam berbagai aspek penelitian, termasuk dataset, *pre-processing*, pengaturan parameter dan *hyperparameter*, metrik evaluasi, arsitektur model, serta perlu dilakukan perhitungan waktu pelatihan dan inferensi. Dengan memperhatikan semua faktor ini, maka dapat membuat perbandingan yang lebih valid dan memberikan wawasan yang lebih berarti tentang efektivitas masing-masing algoritma dalam konteks deteksi website judi online.

2.5 Evaluasi dan Validasi Model

Setelah model dilatih, maka dilakukan evaluasi menggunakan data pengujian untuk mengukur kinerja model deteksi. Adapun model evaluasi pada penelitian ini akan menggunakan metrik evaluasi dan juga *confusion matrix* yang digunakan untuk mengevaluasi keefektifan dan performa model. *Confusion matrix* adalah alat penting untuk mengevaluasi performa model klasifikasi. Matriks ini memberikan rincian mendetail tentang prediksi yang benar dan salah untuk setiap kelas. Ini akan membantu dalam memahami area tertentu di mana model unggul atau mengalami kesulitan. *Confusion matrix* menghasilkan empat nilai yang menunjukkan hasil perhitungan evaluasi model [19]:

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 361-371
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.586

a. *True Positive* (TP) : jumlah data yang sebenarnya positif dan diprediksi sebagai positif, yang menunjukkan bahwa model berhasil mengidentifikasi data positif dengan tepat.
b. *False Positive* (FP) : jumlah data yang sebenarnya negatif tetapi salah diprediksi sebagai positif, yang menunjukkan bahwa model melakukan kesalahan dalam mengklasifikasikan data negatif sebagai positif.
c. *True Negative* (TN) : jumlah data yang benar-benar negatif yang juga diprediksi sebagai negatif, yang berarti model berhasil mengidentifikasi data negatif dengan akurat
d. *False Negative* (FN) : jumlah data yang sebenarnya positif tetapi diprediksi sebagai negatif, yang menunjukkan bahwa model keliru dalam mengklasifikasikan data positif sebagai negatif.

Confusion matrix juga dapat digunakan untuk menghitung berbagai metrik evaluasi seperti akurasi, presisi, recall, dan F1 Score [19] :

a. Akurasi mengukur proporsi prediksi yang benar dari keseluruhan prediksi yang dilakukan oleh model. Rumus akurasi sebagaimana ditampilkan pada rumus (1).

$$\text{Akurasi} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

b. Presisi dihitung sebagai rasio prediksi positif yang benar terhadap total prediksi positif yang dibuat oleh model. Rumus presisi sebagimana ditampilkan pada rumus (2).

$$\text{Presisi} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

c. *Recall* yang juga dikenal sebagai sensitivitas atau tingkat positif yang benar, adalah rasio prediksi positif yang benar terhadap semua prediksi positif yang sebenarnya. Rumus presisi sebagimana ditampilkan pada rumus (3).

$$Recall = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

d. F1-*score* adalah rata-rata harmonik dari presisi dan recall, memberikan keseimbangan antara keduanya. Rumus presisi sebagimana ditampilkan pada rumus (4).

$$F1 - Score = 2 \times \frac{\text{Presisi*Recall}}{\text{Presisi+Recall}} \quad (4)$$

Selain menggunakan evaluasi dengan matrik evaluasi dan *confusion matrix*, penelitian ini juga akan menggunakan model validasi yaitu *cross validation*. *Cross validation* yang dilakukan adalah teknik *k-fold* untuk membantu memastikan bahwa model tidak *overfit* pada set pelatihan tertentu dan memberikan estimasi kinerja yang lebih konsisten dan terpercaya.

3. HASIL DAN PEMBAHASAN

3.1 Dataset

Dataset yang digunakan pada penelitian ini memiliki 12 atribut, yaitu: URL, *domain*, *subdomain*, *domain_ext*, *text_content*, *keyword_count*, *meta_title*, *meta_description*, *meta_keyword*, *num_form*, *external_link*, dan *is_gambling,* seperti yang ditunjukkan pada gambar 2.

```
Atribut/Kolom dalam dataset:
Index(['url', 'domain', 'subdomain', 'domain_ext', 'text_content',
       'keyword_count', 'meta_title', 'meta_description', 'meta_keywords',
       'num_forms', 'external_links', 'is_gambling'],
      dtype='object')
```

Gambar 2. Atribut Dataset

Langkah awal dalam pembuatan dataset adalah pencarian domain website judi menggunakan *tool* SEMrush. Peneliti menentukan beberapa kata kunci seperti: judi, slot, *casino*, taruhan, poker, deposit, bonus, *bet*, gacor, dan *sportsbook*, yang menghasilkan total 4.525 kata kunci turunan terkait website judi. Dari 4.525 kata kunci yang dicari melalui Bing, diperoleh 40.636 URL website judi. Setelah dilakukan proses *scraping* dari 40.636 URL, didapati 34.647 website yang berhasil di-*scrape*.

Sementara itu, untuk kata kunci website non-judi berjumlah 17.794 yang diambil dari situs berita Detik, Tribun, dan Kumparan, yang merupakan kata kunci yang sedang tren. Dari 17.794 kata kunci yang dicari melalui Bing, diperoleh 54.475 URL website non-judi. Setelah dilakukan proses *scraping*, didapati 41.503 website non-judi yang berhasil di-*scrape*..

3.2 Pelabelan dan Validasi Dataset

Setelah dilakukan pelabelan dataset pada penelitian ini, dilakukan validasi ulang dengan bantuan manusia terhadap dataset. Setelah dilakukan validasi, dari 34.647 website judi yang berhasil di-*scrape* sebelumnya, terdapat 24.336 website yang terkonfirmasi sebagai website judi. Sedangkan untuk website non judi, dari 41.503 website terdapat 36.529 website

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 361-371
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.586

yang terkonfirmasi sebagai website non-judi. Adapun website yang tidak terkonfirmasi sebagai website judi dan non judi merupakan website yang sudah tidak dapat di akses (*can't be reached*) atau pun website yang sudah ditangguhkan (*suspended*) dan juga website dengan kesalahan pelabelan.

3.3 *Text Pre-Processing*

3.3.1 *Cleaning*

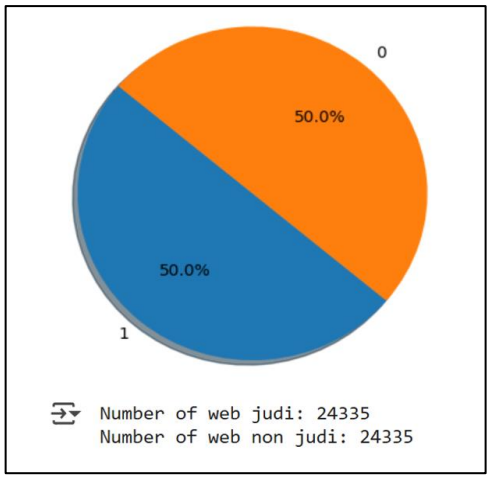
Dalam penelitian ini, proses *pre-processing* dimulai dengan pembersihan data (*cleaning*), yang mencakup beberapa langkah sekaligus, yaitu: standarisasi teks dengan mengonversi semua huruf menjadi huruf kecil (*lowercasing*), penghapusan URL, *hashtag*, *mention*, simbol, karakter spesial termasuk emoji, karakter non-ASCII, karakter kontrol, pengulangan kata, pengulangan tanda baca, dan menghapus spasi ekstra. Perbandingan dari proses *cleaning* dataset dapat dilihat pada tabel 1.

Tabel 1. Tabel Perbandingan *Cleaning* Dataset

Sebelum *Cleaning*	Setelah *Cleaning*
? Demo Spaceman ? Bandar-nya Manjain Pemain Baru ??Banjir Bonus Akhir Tahun Skip to content Slot Gacor Hari Ini Anti Rungkad [KLIK DISINI] SLOT GACOR HARI INI Search Cart Item added to your cart View cart Check out Continue shopping Skip to product information Open media 1 in modal 1 / of 1 Slot Gacor Demo Spaceman ? Bandar-nya Manjain Pemain Baru ??Banjir Bonus Akhir Tahun Demo Spaceman ?	demo spaceman bandarnya manjain pemain baru banjir bonus akhir tahun skip to content slot gacor hari ini anti rungkad klik disini slot gacor hari ini search cart item added to your cart view cart check out continue shopping skip to product information open media 1 in modal 1 of 1 slot gacor demo spaceman bandarnya manjain pemain baru banjir bonus akhir tahun demo spaceman

3.3.2 *Undersampling*

Tahap selanjutnya adalah *undersampling* terhadap kelas non-judi. Pada penelitian ini, dataset non-judi lebih banyak dibandingkan dataset judi, sehingga sampel dari kelas non-judi dikurangi secara acak hingga keduanya memiliki jumlah yang sama, yaitu 24.335 sampel masing-masing, seperti yang terlihat pada gambar 3. Teknik ini bertujuan untuk mencegah bias model terhadap kelas dominan dan meningkatkan kinerja model.



Gambar 3. Hasil Proses *Undersampling*

3.3.3 Tokenisasi, *Stop Words,* dan *Stemming*

Proses *pre-processing* dengan tokenisasi juga diterapkan dalam penelitian ini. Proses ini berfungsi untuk memecah teks menjadi token atau unit kata individu. Hasil proses tokenisasi dataset dapat dilihat pada tabel 2. Setelah tokenisasi dilakukan, kemudian dilakukan penghapusan *stop words* pada dataset dengan menggunakan pustaka NLTK. Menggunakan NLTK untuk menghapus *stop words* pada penlitian ini dinilai sangat bermanfaat, karena menyediakan daftar *stop words* untuk bahasa Indonesia. Pustaka ini mudah diintegrasikan dengan langkah-langkah *pre-processing* lain seperti tokenisasi dan *stemming*, sehingga mendukung efisiensi *pipeline pre-processing*. Contoh perbandingan token sebelum dan setelah proses penghapusan *stop words* dapat dilihat pada tabel 3.

Tabel 2. Tabel Hasil Proses Tokenisasi

Sebelum Tokenisasi	Setelah Tokenisasi
ligaciputra link daftar slot gacor link alternatif login masuk daftar selamat datang di situs ligaciputra situs slot	['ligaciputra', 'link', 'daftar', 'slot', 'gacor', 'link', 'alternatif', 'login', 'masuk', 'daftar', 'selamat', 'datang', 'di', 'situs',

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 361-371
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.586

Sebelum Tokenisasi	Setelah Tokenisasi
populer gampang menang 2023 daftar game slot online paling gacor hari ini gates of olympus slot dengan tema kerajaan dewa yunani kuno dapatkan petir dengan nilai perkalian terbesar serta dapatkan jackpot besarnya rtp 9579 mahjong ways 2 slot tema catur nasional tiongkok beli fitur free spins untuk bonus maksimum akan meningkat hingga 10 kali lipat rtp 9615	'ligaciputra', 'situs', 'slot', 'populer', 'gampang', 'menang', '2023', 'daftar', 'game', 'slot', 'online', 'paling', 'gacor', 'hari', 'ini', 'gates', 'of', 'olympus', 'slot', 'dengan', 'tema', 'kerajaan', 'dewa', 'yunani', 'kuno', 'dapatkan', 'petir', 'dengan', 'nilai', 'perkalian', 'terbesar', 'serta', 'dapatkan', 'jackpot', 'besarnya', 'rtp', '9579', 'mahjong', 'ways', '2', 'slot', 'tema', 'catur', 'nasional', 'tiongkok', 'beli', 'fitur', 'free', 'spins', 'untuk', 'bonus', 'maksimum', 'akan', 'meningkat', 'hingga', '10', 'kali', 'lipat', 'rtp', '9615']

Tabel 3. Tabel Hasil Proses Penghapusan *Stop Words*

Sebelum *Stop Words*	Setelah *Stop Words*
['ligaciputra', 'link', 'daftar', 'slot', 'gacor', 'link', 'alternatif', 'login', 'masuk', 'daftar', 'selamat', 'datang', 'di', 'situs', 'ligaciputra', 'situs', 'slot', 'populer', 'gampang', 'menang', '2023', 'daftar', 'game', 'slot', 'online', 'paling', 'gacor', 'hari', 'ini', 'gates', 'of', 'olympus', 'slot', 'dengan', 'tema', 'kerajaan', 'dewa', 'yunani', 'kuno', 'dapatkan', 'petir', 'dengan', 'nilai', 'perkalian', 'terbesar', 'serta', 'dapatkan', 'jackpot', 'besarnya', 'rtp', '9579', 'mahjong', 'ways', '2', 'slot', 'tema', 'catur', 'nasional', 'tiongkok', 'beli', 'fitur', 'free', 'spins', 'untuk', 'bonus', 'maksimum', 'akan', 'meningkat', 'hingga', '10', 'kali', 'lipat', 'rtp', '9615']	['ligaciputra', 'link', 'daftar', 'slot', 'gacor', 'link', 'alternatif', 'login', 'masuk', 'daftar', 'selamat', 'situs', 'ligaciputra', 'situs', 'slot', 'populer', 'gampang', 'menang', '2023', 'daftar', 'game', 'slot', 'online', 'gacor', 'gates', 'of', 'olympus', 'slot', 'tema', 'kerajaan', 'dewa', 'yunani', 'kuno', 'dapatkan', 'petir', 'nilai', 'perkalian', 'terbesar', 'dapatkan', 'jackpot', 'besarnya', 'rtp', '9579', 'mahjong', 'ways', '2', 'slot', 'tema', 'catur', 'nasional', 'tiongkok', 'beli', 'fitur', 'free', 'spins', 'bonus', 'maksimum', 'meningkat', '10', 'kali', 'lipat', 'rtp', '9615']

Selanjutnya, diterapkan proses *stemming* menggunakan pustaka Sastrawi untuk mengonversi kata-kata ke bentuk dasarnya. Pustaka Sastrawi dipilih karena kesesuaiannya dengan penggunaan bahasa Indonesia dalam dataset. Contoh perbandingan token sebelum dan setelah proses penghapusan *stemming* dapat dilihat pada tabel 4. Setelah menghapus *stop words* dan melakukan *stemming*, token-token tersebut digabungkan kembali menjadi teks utuh. Proses ini memastikan bahwa data teks yang dihasilkan lebih bersih dan siap dianalisis lebih lanjut, serta memberikan dasar yang kuat bagi model untuk bekerja lebih efektif. Setelah seluruh langkah *pre-processing* diterapkan, dataset hasil *pre-processing* disimpan dalam bentuk CSV dan akan digunakan pada langkah selanjutnya dalam pembuatan model. Perbandingan data sebelum dan sesudah proses tokenisasi, penghapusan *stop words* dan *stemming* dapat dilihat pada tabel 5.

Tabel 4. Tabel Hasil Proses *Stemming*

Sebelum *Stemming*	Setelah *Stemming*
['ligaciputra', 'link', 'daftar', 'slot', 'gacor', 'link', 'alternatif', 'login', 'masuk', 'daftar', 'selamat', 'situs', 'ligaciputra', 'situs', 'slot', 'populer', 'gampang', 'menang', '2023', 'daftar', 'game', 'slot', 'online', 'gacor', 'gates', 'of', 'olympus', 'slot', 'tema', 'kerajaan', 'dewa', 'yunani', 'kuno', 'dapatkan', 'petir', 'nilai', 'perkalian', 'terbesar', 'dapatkan', 'jackpot', 'besarnya', 'rtp', '9579', 'mahjong', 'ways', '2', 'slot', 'tema', 'catur', 'nasional', 'tiongkok', 'beli', 'fitur', 'free', 'spins', 'bonus', 'maksimum', 'meningkat', '10', 'kali', 'lipat', 'rtp', '9615']	['ligaciputra', 'link', 'daftar', 'slot', 'gacor', 'link', 'alternatif', 'login', 'masuk', 'daftar', 'selamat', 'situs', 'ligaciputra', 'situs', 'slot', 'populer', 'gampang', 'menang', '2023', 'daftar', 'game', 'slot', 'online', 'gacor', 'gates', 'of', 'olympus', 'slot', 'tema', 'raja', 'dewa', 'yunani', 'kuno', 'dapat', 'petir', 'nilai', 'kali', 'besar', 'dapat', 'jackpot', 'besar', 'rtp', '9579', 'mahjong', 'ways', '2', 'slot', 'tema', 'catur', 'nasional', 'tiongkok', 'beli', 'fitur', 'free', 'spins', 'bonus', 'maksimum', 'tingkat', '10', 'kali', 'lipat', 'rtp', '9615']

Tabel 5. Tabel Perbandingan Hasil *Pre-Processing* Dataset

Sebelum Tokenisasi, *Stop Words* dan *Stemming*	Setelah Tokenisasi, *Stop Words* dan *Stemming*
ligaciputra link daftar slot gacor link alternatif login masuk daftar selamat datang di situs ligaciputra situs slot populer gampang menang 2023 daftar game slot online paling gacor hari ini gates of olympus slot dengan tema kerajaan dewa yunani kuno dapatkan petir dengan nilai perkalian terbesar serta dapatkan jackpot besarnya rtp 9579 mahjong ways 2 slot tema catur nasional tiongkok beli fitur free spins untuk bonus maksimum akan meningkat hingga 10 kali lipat rtp 9615	ligaciputra link daftar slot gacor link alternatif login masuk daftar selamat situs ligaciputra situs slot populer gampang menang 2023 daftar game slot online gacor gates of olympus slot tema raja dewa yunani kuno dapat petir nilai kali besar dapat jackpot besar rtp 9579 mahjong ways 2 slot tema catur nasional tiongkok beli fitur free spins bonus maksimum tingkat 10 kali lipat rtp 9615

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 361-371
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.586

Visualisasi data menggunakan *WordCloud* juga diterapkan untuk situs judi dan non-judi, guna memberikan gambaran tentang distribusi kata dalam dataset ini. Gambar 4 adalah visualisasi data untuk dataset judi dan non judi setelah dilakukan proses *pre-processing* pada penelitian ini. Visualisasi data judi menunjukkan dominasi kata-kata seperti "judi online", "slot online", dan "situs judi", yang menandakan fokus pada jenis permainan perjudian dan *platform* tempat bermain. Hal ini menunjukkan bahwa dataset ini berkaitan erat dengan industri perjudian online. Di sisi lain, visualisasi data non-judi menyoroti kata-kata seperti "berita iklan", "baca berita", "iklan klik" dan "timnas indonesia", yang mengindikasikan fokus pada informasi, khususnya berita, olahraga dan iklan. Perbedaan ini mencerminkan tujuan yang berbeda antara data judi dan non judi, data dengan konten judi berorientasi pada analisis pasar dan promosi perjudian, sementara yang data konten non judi lebih menekankan penyebaran informasi terkini di bidang berita dan olahraga.

Gambar 4. Hasil Proses *Undersampling*

3.4 Penentuan Teknik Vektorisasi

Pemilihan teknik vektorisasi untuk SVM dan CNN sangat penting karena masing-masing algoritma memiliki karakteristik yang berbeda. Untuk SVM, teknik vektorisasi yang akan digunakan adalah TF-IDF, karena metode ini sederhana, efisien, dan memberikan bobot lebih pada kata-kata relevan, dan membantu pemisahan kelas dengan lebih baik.

Sementara itu, CNN lebih diuntungkan dengan penggunaan *word embeddings* seperti Word2Vec, yang dapat menangkap hubungan semantik antar kata dan menghasilkan representasi data yang efisien. Dengan *word embeddings*, CNN dapat belajar dari pola yang lebih kompleks dalam data teks dan mengatasi variasi bahasa. Secara keseluruhan, pemilihan teknik vektorisasi yang tepat sangat berpengaruh pada efektivitas kedua algoritma dalam memanfaatkan informasi dari data teks

3.5 Penentuan Parameter dan *Hyperparamenter*

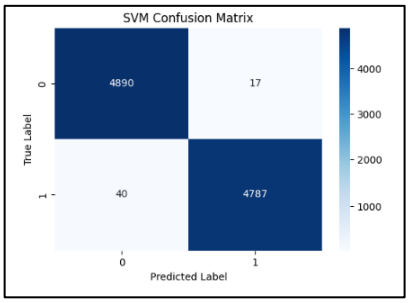
Pada penelitian ini, algoritma SVM menggunakan *hyperparameter* C. C adalah parameter regulasi yang mengontrol *trade-off* antara memaksimalkan margin dan meminimalkan kesalahan klasifikasi. Nilai C yang lebih kecil membuat model lebih toleran terhadap kesalahan, yang dapat membantu dalam menghindari *overfitting*. Sebaliknya, nilai C yang lebih besar memberikan penalti yang lebih besar untuk kesalahan, yang dapat meningkatkan akurasi pada set pelatihan tetapi berisiko *overfitting* pada data baru. Dalam eksperimen ini, peneliti memilih nilai C=1, berdasarkan pengaturan *default* yang sering digunakan sebagai titik awal, tanpa eksplorasi sistematis seperti *grid search* atau metode lain, karena fokus utama adalah pada efisiensi pelatihan model. Selain itu, nilai C=1 juga sering digunakan dalam penelitian lainnya yang tidak menerapkan metode *grid search*, sehingga dianggap memadai untuk konteks penelitian ini.

Algoritma CNN menggunakan parameter kernel dan *hyperparameter* : konvolusi *layer*, *batch* dan *epoch*. Konvolusi *layer* (Conv1D) pada model CNN menggunakan jumlah filter yaitu 128 dan ukuran kernel yaitu 5. Ini adalah pengaturan yang dianggap umum untuk model CNN dan dapat memberikan hasil yang baik, terutama untuk analisis teks. *Batch size* yang digunakan adalah 32 dan *epochs* 10. Peneliti memilih *batch size* 32 agar memungkinkan model untuk lebih sering diperbarui tanpa menghabiskan terlalu banyak memori. Hal ini seringkali memberikan keseimbangan yang baik untuk pelatihan. Jumlah *epochs* 10 juga dinilai cukup untuk penelitian ini.

3.6 Evaluasi Model Deteksi

Model SVM dan CNN dilatih menggunakan data pelatihan untuk mendeteksi konten perjudian dengan rasio 80 : 20. Ini berarti perbandingan sebanyak 80% data pelatihan dan 20% data pengujian. Dalam pengujian ini, model SVM menunjukkan akurasi sekitar 99%, yang menunjukkan kemampuannya yang sangat baik dalam mengidentifikasi situs judi. Hasil evaluasi model SVM diperoleh melalui *confusion matrix*, yang memberikan gambaran mengenai banyaknya prediksi benar dan salah yang dibuat oleh model.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 361-371
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.586



Gambar 5. *Confusion Matrix* Model SVM

Seperti yang terlihat pada gambar 5, *True Positive* (TP) berjumlah 4890, menunjukkan bahwa model memprediksi situs sebagai judi dan situs tersebut memang benar judi. Ini menandakan bahwa model memiliki kemampuan yang sangat baik dalam mengidentifikasi konten judi, dengan prediksi benar yang sangat tinggi. *False Positive* (FP) berjumlah 17, yang berarti model salah mengklasifikasikan situs non-judi sebagai judi. Meskipun terdapat kesalahan, jumlah ini relatif kecil dan menunjukkan bahwa model cukup akurat.

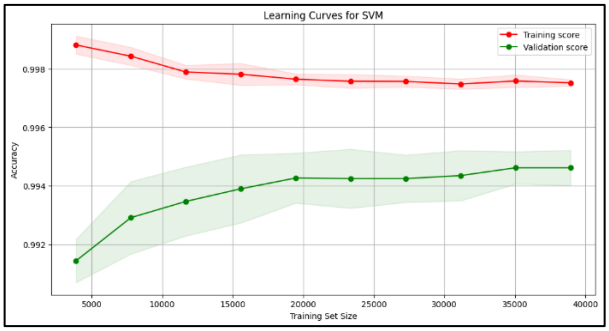
True Negative (TN) berjumlah 4787, yang berarti model memprediksi situs sebagai non-judi dan situs tersebut memang benar non-judi. Ini menandakan bahwa model memiliki kemampuan baik dalam mengidentifikasi situs non-judi, meskipun jumlahnya sedikit lebih rendah dibandingkan dengan jumlah prediksi benar untuk judi. *False Negative* (FN) berjumlah 40, yang berarti model salah mengklasifikasikan situs judi sebagai non-judi. Ini menunjukkan bahwa model masih memiliki beberapa kesalahan dalam mendeteksi konten judi, tetapi jumlah kesalahan ini dapat dianggap minimal.

```
SVM Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      4907
           1       1.00      0.99      0.99      4827

    accuracy                           0.99      9734
   macro avg       0.99      0.99      0.99      9734
weighted avg       0.99      0.99      0.99      9734

SVM Confusion Matrix:
 [[4890   17]
 [  40 4787]]
SVM Accuracy: 0.99
SVM Precision: 1.00
SVM Recall: 0.99
SVM F1 Score: 0.99
```

Gambar 6. Matrik Evaluasi Model SVM

Evaluasi model juga dilakukan dengan menggunakan metrik evaluasi seperti : akurasi, presisi, *recall*, dan F1-*score*, seperti yang terlihat pada gambar 6. Presisi untuk kelas judi adalah 1.00, yang berarti 100% dari situs yang diprediksi sebagai judi oleh model memang benar judi. *Recall* untuk kelas judi adalah 0.99, menunjukkan bahwa 99% dari semua situs judi berhasil diidentifikasi oleh model. F1-*score* untuk model ini adalah 0.99, menandakan keseimbangan yang sangat baik dengan *recall*. Secara keseluruhan, akurasi model SVM mencapai 99%, yang berarti model mampu memprediksi dengan benar sebanyak 99% dari seluruh data yang diuji. Model SVM ini menunjukkan performa yang sangat baik dalam mengidentifikasi konten judi.

Kurva pembelajaran untuk model SVM yang diilustrasikan pada gambar 7, juga dianalisis untuk mengevaluasi performa model seiring dengan peningkatan ukuran set pelatihan. Grafik menunjukkan bahwa akurasi pelatihan hampir mencapai 100%, yang mengindikasikan kemampuan model yang sangat baik dalam memahami dan belajar dari data pelatihan. Namun, akurasi validasi meningkat secara bertahap hingga mencapai nilai stabil, tetapi tetap berada di bawah akurasi pelatihan. Ini mengisyaratkan bahwa meskipun model dapat melakukan generalisasi dengan baik, terdapat sedikit jarak antara kinerja pada data pelatihan dan validasi, yang mungkin menandakan adanya potensi *overfitting*. Hasil ini menunjukkan bahwa model SVM sudah cukup efektif, tetapi perlu diperhatikan potensi *overfitting* untuk memastikan generalisasi yang lebih baik pada data yang tidak terlihat.

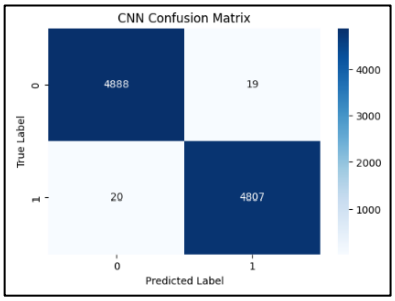


Gambar 7. Kurva Pembelajaran Model SVM

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 361-371
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.586

Model CNN pada penelitian ini mencapai akurasi sempurna, yaitu 100%. Hasil evaluasi model CNN juga diperoleh melalui *confusion matri*x yang terlihat pada gambar 8. *True Positive* (TP) berjumlah 4888, menunjukkan bahwa model memprediksi situs sebagai judi dan situs tersebut memang benar judi. Ini menandakan bahwa model memiliki kemampuan yang sangat baik dalam mengidentifikasi konten judi, dengan jumlah prediksi benar yang sangat tinggi. *False Positive* (FP) berjumlah 19, yang berarti model salah mengklasifikasikan situs non-judi sebagai judi. Meskipun ada beberapa kesalahan, jumlah ini tergolong kecil dan menunjukkan bahwa model cukup akurat dalam klasifikasi.

True Negative (TN) berjumlah 4807, yang berarti model memprediksi situs sebagai non-judi dan situs tersebut memang benar non-judi. Ini menunjukkan bahwa model memiliki kemampuan yang baik dalam mengidentifikasi situs non-judi, dengan sedikit kesalahan. *False Negative* (FN) berjumlah 20, yang berarti model salah mengklasifikasikan situs judi sebagai non-judi. Meskipun terdapat beberapa kesalahan, jumlah ini tetap tergolong rendah.



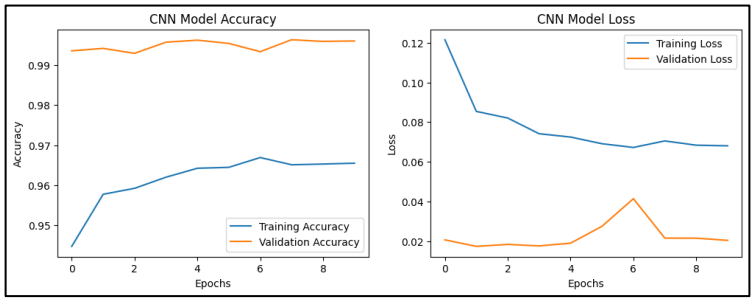
Gambar 8. *Confusion Matrix* Model CNN

Metrik evaluasi pada model CNN juga melihat nilai akurasi, presisi, *recall*, dan F1-*score* yang ditunjukkan pada gambar 9, seperti hal nya SVM. Presisi untuk kelas judi adalah 1.00, yang berarti 100% dari situs yang diprediksi sebagai judi oleh model memang benar judi. *Recall* untuk kelas judi juga mencapai 1.00, menunjukkan bahwa 100% dari semua situs judi berhasil diidentifikasi oleh model. F1-*score* untuk kelas ini adalah 1.00, menandakan keseimbangan yang sangat baik antara presisi dan *recall*. Secara keseluruhan, akurasi model CNN mencapai 100%, yang berarti model mampu memprediksi dengan benar seluruh data yang diuji. Model ini menunjukkan performa yang sempurna dalam mengidentifikasi konten judi.

```
305/305 ━━━━━━━━━━━━━━━━━━━━ 2s 7ms/step
CNN Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      4907
           1       1.00      1.00      1.00      4827

    accuracy                           1.00      9734
   macro avg       1.00      1.00      1.00      9734
weighted avg       1.00      1.00      1.00      9734

CNN Confusion Matrix:
 [[4888   19]
 [  20 4807]]
CNN Accuracy: 1.00
CNN Precision: 1.00
CNN Recall: 1.00
CNN F1 Score: 1.00
```

Gambar 9. Matrik Evaluasi Model CNN

Pada kurva pembelajaran model CNN yang diilustrasikan pada gambar 10, terlihat juga grafik akurasi menunjukkan bahwa akurasi pelatihan meningkat secara konsisten, mendekati 100%, sementara akurasi validasi juga meningkat namun sedikit lebih rendah. Ini mengindikasikan bahwa model telah belajar dengan baik dari data pelatihan, tetapi perbedaan antara keduanya dapat menandakan potensi *overfitting*. Pada grafik *loss*, terlihat bahwa *loss* pelatihan turun drastis di awal *epoch* dan kemudian stabil di nilai rendah, sementara *loss* validasi juga menurun tetapi tidak secepat loss pelatihan. Hal ini menunjukkan bahwa meskipun model efektif dalam mempelajari data pelatihan, performanya pada data validasi kurang optimal. Secara keseluruhan, kedua grafik memberikan gambaran positif tentang performa model, tetapi perlu dipastikan jika perbedaan antara akurasi pelatihan dan validasi terus melebar, yang dapat menunjukkan *overfitting*.



Gambar 10. Kurva Pembelajaran Model CNN

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 5, No 4, June 2025 | Hal 361-371
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.586

3.7 Validasi Model

Adanya indikasi terhadap *overfitting* pada kedua model deteksi perlu dipastikan pada penelitian ini dengan melakukan validasi terhadap model yang dibangun. Validasi model dilakukan menggunakan teknik *cross-validation* yang bertujuan untuk mengevaluasi kinerja algoritma SVM dan CNN dalam mendeteksi konten perjudian. Metode *Stratified K-Fold Cross-Validation* diterapkan dengan membagi dataset menjadi lima *fold*, memastikan bahwa setiap *fold* mempertahankan proporsi kelas yang seimbang. Proses ini tidak hanya mengurangi risiko *overfitting*, tetapi juga memberikan estimasi kinerja yang lebih akurat dengan melatih dan menguji model pada berbagai subset data.

```
Cross-Validation Scores: [0.99455517 0.99537703 0.9952743  0.9949661  0.99414424]
Mean Accuracy: 0.99
Standard Deviation: 0.00
```

Gambar 11. *Cross-Validation* Model SVM

Model SVM dilatih dan diuji pada setiap bagian secara bergantian. Hasilnya seperti yang terlihat pada gambar 11, menunjukkan akurasi yang tinggi sekitar 99.41% hingga 99.53% dengan variasi performa yang sangat rendah, menandakan konsistensi dan stabilitas model. Rata-rata akurasi keseluruhan adalah 99%, menunjukkan bahwa model berkinerja sangat baik dalam memprediksi data yang tidak terlihat. *Standard deviation* dari akurasi adalah 0.00%, yang menandakan bahwa performa model sangat konsisten dan stabil di seluruh iterasi.

```
Accuracy for fold: 99.58%
Accuracy for fold: 99.67%
Accuracy for fold: 99.63%
Accuracy for fold: 99.58%
Accuracy for fold: 99.61%
Mean Accuracy: 99.61%
Standard Deviation: 0.03%
```

Gambar 12. *Cross-Validation* Model CNN

Hasil *cross-validation* untuk model CNN dengan metode *Stratified K-Fold Cross-Validation* yang sama dengan model SVM, menunjukkan akurasi yang sangat tinggi pada setiap *fold*, dengan nilai sekitar 99.58% hingga 99.67%, seperti yang terlihat pada gambar 12. Rata-rata akurasi keseluruhan adalah 99.61%, menunjukkan bahwa model berkinerja sangat baik dalam memprediksi data yang tidak terlihat. *Standard deviation dari* akurasi antar fold adalah 0.03%, yang menandakan bahwa performa model sangat konsisten dan stabil di seluruh iterasi.

Dari hasil validasi model, menunjukkan bahwa baik model SVM maupun CNN memiliki kinerja yang sangat baik dalam mendeteksi konten perjudian, dengan akurasi masing-masing rata-rata 99% dan 99.61%. Teknik *Stratified K-Fold Cross-Validation* yang diterapkan pada kedua model berhasil mempertahankan proporsi kelas yang seimbang dan mengurangi risiko *overfitting*, sehingga memberikan estimasi kinerja yang lebih akurat. Model SVM menunjukkan stabilitas yang tinggi dengan *standard deviation* 0.00%, sedangkan model CNN juga menunjukkan konsistensi dengan *standard deviation* 0.03%.

Meskipun kedua model menunjukkan akurasi yang tinggi, CNN sedikit unggul dalam hal akurasi rata-rata. Perbedaan ini menunjukkan bahwa model CNN mungkin lebih adaptif terhadap kompleksitas data dalam konteks deteksi konten perjudian dibandingkan dengan model SVM.

4. KESIMPULAN

Hasil penelitian ini menunjukkan bahwa kedua algoritma, *Support Vector Machine* (SVM) dan *Convolutional Neural Network* (CNN), mencapai performa klasifikasi yang tinggi dalam mendeteksi website perjudian. SVM mencatat akurasi sekitar 99%, sedangkan CNN mencapai akurasi sempurna yaitu 100%. Perlu diperhatikan bahwa hasil akurasi sempurna pada model CNN ini dicapai dalam kondisi tertentu, termasuk dataset yang relatif bersih dan proses pelatihan yang optimal pada penelitian. Model SVM menunjukkan metrik evaluasi yang sangat baik, termasuk presisi, *recall*, dan F1-*score* yang hampir sempurna. Namun, CNN unggul dalam akurasi dan kemampuan menangkap pola kompleks berkat arsitektur konvolusinya, serta penggunaan Word2Vec untuk representasi kata yang lebih kaya. Kedua model menunjukkan konsistensi selama pelatihan dan evaluasi. Pada proses evaluasi, model memperlihatkan penurunan *loss* signifikan dan peningkatan akurasi. Metode *cross-validation* diterapkan sebagai metode validasi model. Hasilnya menunjukkan bahwa kedua model berfungsi dengan baik pada data pelatihan dan pengujian, menandakan kemampuan generalisasi yang baik. Akurasi tinggi yang dicapai juga dipengaruhi oleh proses *pre-processing* yang menyeluruh, termasuk pembersihan data dan teknik *undersampling* untuk menyeimbangkan dataset. Langkah-langkah seperti *lowercasing*, penghapusan *stop words,* dan *stemming* juga berkontribusi secara signifikan dalam membuat model lebih efisien dan meningkatkan akurasi. Meskipun demikian, penelitian ini dapat ditingkatkan dengan teknik *grid search* untuk optimasi *hyperparameter* dan eksplorasi variasi arsitektur. Penambahan data latih yang lebih beragam juga dapat membantu generalisasi model di dunia nyata. Penelitian lebih lanjut dapat mempertimbangkan algoritma lain seperti *Recurrent Neural Networks* (RNN) atau *Transformer* untuk perbandingan efektivitas.

REFERENCES

[1] Y. Chen, R. Zheng, A. Zhou, S. Liao, and L. Liu, "Automatic detection of pornographic and gambling websites based on visual and textual content using a decision mechanism," *Sensors (Switzerland)*, vol. 20, no. 14, pp. 1–21, 2020, doi: 10.3390/s20143989.

[2] R. T. Huang, C. H. Shih, T. C. Lin, and W. Y. Lin, "The Study on Illegal Online Gambling Investigation in Taiwan," *Procedia Comput. Sci.*, vol. 207, no. Kes, pp. 2901–2910, 2022, doi: 10.1016/j.procs.2022.09.348.

[3] Kominfo, "Hindari Jerat Judi Online, Menteri Budi Arie Sarankan Empat Langkah," *Komdigi*, 2024. https://www.komdigi.go.id/berita/pengumuman/detail/hindari-jerat-judi-online-menteri-budi-arie-sarankan-empat-langkah (accessed Oct. 04, 2024).

[4] A. Rachman, "PPATK: Perputaran Uang Judi Online Rp 327 Triliun di 2023," *CNBC Indonesia*, 2024. https://www.cnbcindonesia.com/news/20240112090558-4-505064/ppatk-perputaran-uang-judi-online-rp-327-triliun-di-2023 (accessed Jan. 12, 2024).

[5] D. Savitri, "RI Jadi Negara dengan Pemain Judi Online Terbanyak, Ini Penyebabnya Kata Dosen UI," *Detik*, 2024. https://www.detik.com/edu/edutainment/d-7470624/ri-jadi-negara-dengan-pemain-judi-online-terbanyak-ini-penyebabnya-kata-dosen-ui.

[6] B. H. K. Kominfo, "Menkominfo Budi Arie Gas Pol Basmi Judi Online," *Kominfo*, 2024. https://www.komdigi.go.id/berita/pengumuman/detail/menkominfo-budi-arie-gas-pol-basmi-judi-online (accessed Jun. 14, 2024).

[7] E. Benavides-Astudillo, W. Fuertes, S. Sanchez-Gordon, D. Nuñez-Agurto, and G. Rodríguez-Galán, "A Phishing-Attack-Detection Model Using Natural Language Processing and Deep Learning," *Appl. Sci.*, vol. 13, no. 9, 2023, doi: 10.3390/app13095275.

[8] Y. Cheng, H. Jiang, Z. Zhang, Y. Du, and T. Chai, "Birds of a Feather Flock Together: Generating Pornographic and Gambling Domain Names Based on Character Composition Similarity," *Wirel. Commun. Mob. Comput.*, vol. 2022, pp. 1–17, 2022, doi: 10.1155/2022/4408987.

[9] L. Jiayin, Y. Jie, N. Bowei, Z. Chengchen, and H. Haichen, "Capture Methods of Gambling Related Illegal Websites in Massive Websites," *J. Data Acquis. Process.*, vol. 36, no. 5, pp. 1050–1061, 2021.

[10] C. Wang, M. Zhang, F. Shi, P. Xue, and Y. Li, "A Hybrid Multimodal Data Fusion-Based Method for Identifying Gambling Websites," *Electron.*, vol. 11, no. 16, 2022, doi: 10.3390/electronics11162489.

[11] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," *Telecommun. Syst.*, vol. 76, no. 1, pp. 139–154, 2021, doi: 10.1007/s11235-020-00733-2.

[12] M. Min, J. J. Lee, and K. Lee, "Detecting Illegal Online Gambling (IOG) Services in the Mobile Environment," *Secur. Commun. Networks*, vol. 2022, pp. 1–12, 2022, doi: 10.1155/2022/3286623.

[13] M. O. Raza *et al.*, "Reading Between the Lines: Machine Learning Ensemble and Deep Learning for Implied Threat Detection in Textual Data," *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, pp. 1–17, 2024, doi: 10.1007/s44196-024-00580-y.

[14] R. Nisa, S. Amriza, and D. Supriyadi, "Komparasi Metode Machine Learning dan Deep Learning untuk Deteksi Emosi pada Text di Sosial Media," *JUPITER*, vol. 13, no. 2, pp. 130–139, 2021.

[15] R. A. Y and K. M. L, "Perbandingan Algoritma Cnn Dan Svm Untuk Analisis Sentimen Mengenai Kenaikan Harga Bahan Bakar Minyak," *eProceedings Eng.*, vol. 10, no. 6, pp. 5418–5422, 2023.

[16] P. M. Aryanto and R. Mardhiyyah, "Analisis Sentimen Terhadap Review Google Maps Jogja City Mall Menggunakan Algoritma Support Vector Machine," *J. Comput. Syst. Informatics*, vol. 6, no. 1, pp. 25–35, 2024, doi: 10.47065/josyc.v6i1.6049.

[17] D. Adelia, W. Astuti, and K. M. Lhaksmana, "Election Hoax Detection on X using CNN with TF-RF and TF-IDF Weighting Features," *J. Comput. Syst. Informatics*, vol. 5, no. 4, pp. 912–920, 2024, doi: 10.47065/josyc.v5i4.5778.

[18] A. D. Cahyani, A. K. Ramdani, and Y. Sibaroni, "Hoax Detection of Covid-19 News on Social Media using Convolutional Neural Network (CNN) and Support Vector Machine (SVM)," *Int. J. Inf. Commun. Technol.*, vol. 9, no. 2, pp. 177–185, 2023, doi: 10.21108/ijoict.v9i2.872.

[19] E. Prasetyaningrum, S. Rijal, A. Purwanto, and A. Aziz, "Sentiment Analysis of Rice Price Increase on Facebook Using Naïve Bayes Algorithm," *J. Comput. Sci. Inf. Technol. (CoSciTech)*, vol. 5, no. 2, pp. 381–390, 2024, [Online]. Available: https://ejurnal.umri.ac.id/index.php/coscitech/article/view/5161/2439.

[20] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Inf.*, vol. 14, no. 54, 2023, doi: 10.3390/info14010054.

[21] R. Nurlaely, S. D. Sartika, Kamdan, and I. L. Kharisma, "Analisis Sentimen Twitter Terhadap Cyberbullying Menggunakan Metode Support Vector Machine (SVM)," *J. Comput. Sci. Inf. Technol.*, vol. 4, no. 2, pp. 376–384, 2023, [Online]. Available: http://ejurnal.umri.ac.id/index.php/coscitech/indexhttps://doi.org/10.37859/coscitech.v4i2.5161.

[22] R. Arifiandy and H. Fahmi, "Perbandingan Vektorisasi Deteksi Spam Email Menggunakan Bag of Word, TF IDF, dan Word2Vec pada Multinomial Naïve Bayes," *Syntax J. Inform.*, vol. 13, no. 01, pp. 1–14, 2024, [Online]. Available: https://upgradeojs3.vpsrifqi.com/index.php/syntax/article/download/11254/4585.