Klasifikasi Berita Berdasarkan Kategori Menggunakan Convolutional Neural Network dengan IndoBERT

Jonathan Federico Tantoro[a1], I Dewa Made Bayu Atmaja Darmawan[a2]

Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Jalan Raya Kampus Udayana, Bukit Jimbaran, Kuta Selatan, Badung, Bali, Indonesia
[1]jonathanft2108@gmail.com
[2]dewabayu@unud.ac.id

Abstract

The advancement of technology information has led to a significant increased the volume of digital news, that makes needs for automatic news classification. This study aims to design a model capable of caterogizing Indonesian language news articles into six predefined categories, such as News, Money, Bola, *Health,* Tekno, *and Tren. To achieve this goal, the method used combines IndoBERT as the embedding technique with Convolutional Neural Network (CNN) as the classification algorithm. The dataset consists of 3.000 news articles collected from Kompas.com and is divided into training data and testing data using four different data split ratios: 60:40, 70:30 80:20, and 90:10 . The evaluation results show that the best performance was achieved using the 80:20 ratio, where the model reached an accuracy of 91%, along with high precision, recall, and F-1 Score These result prove that the combination of IndoBERT and CNN is effective for the automatic classification of Indonesian new texts.*

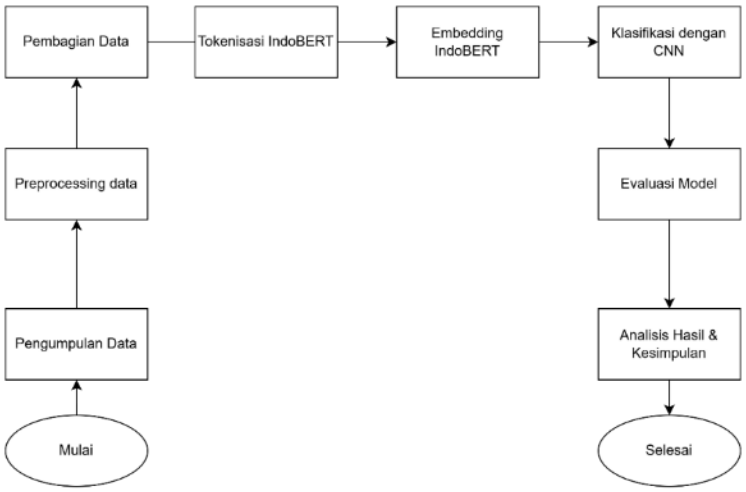
Keywords: *IndoBert, Convolutional Neural Network, Text Classification, Indonesian News, Contextual embedding*

1. Pendahuluan

Perkembangan teknologi informasi telah membawa perubahan signifikan dalam kecepatan akses informasi [1]. Berita sebagai salah satu sumber informasi utama kini tidak hanya tesedia dalam bentuk cetak, tetapi juga tersebar melalui portal berita *online* dan media sosial [2]. Kemudahan akses berita inilah yang menyebabkan peningkatan volume dan variasi berita yang dipublikasikan setiap harinya. Peningkatan volume berita tesebut menghadirkan tantangan, khususnya dalam pengelompokan dan klasifikasi berita. Proses klasifikasi manual menjadi tidak efisien karena membutuhkan waktu yang lama sehingga tidak lagi mampu mengimbangi laju pertumbuhan berita [3]. Hal ini mengakibatkan pembaca kesulitan dalam menemukan berita yang relevan sesuai dengan kategori yang diinginkan, sehingga menggangu pengalaman pembaca. Untuk menjawab permasalahan tersebut, berbagai pendekatan klasifikasi teks telah digunakan. Salah satu pendekatan yang banyak digunakan adalah CNN (*Convolutional Neural Network),* yang telah terbukti efektif dalam mengekstraksi fitur spasial dari representasi teks. Penelitian yang dilakukan oleh Ramdhani menunjukkan bahwa CNN mampu mencapai akurasi 90,74% dalam klasifikasi berita berbahasa Indonesia [4]. Selain CNN, pengunaan model *transformer* seperti IndoBERT juga mulai digunakan sebagai teknik *embedding* yang mampu menangkap konteks dalam kalimat dengan akurat. Penelitian yang dilakukan oleh Prisscilya menggunakan IndoBERT sebagai *embedder* dan berhasil memperoleh hasil yang baik dengan nilai presisi 87% dan F1-*score* 88% [5]. Berdasarkan penemuan tersebut, penelitian ini mengusulkan kombinasi antara CNN sebagai algoritma klasifikasi dan IndoBERT sebagai *embedder.* Tujuannya untuk membangun sistem klasifikasi berita yang mampu mengelompokkan berita ke dalam kategori seperti “News”, “Health”, “Bola”, “Money”,”Tekno”, dan “Tren” secara otomatis.

2. Metode Penelitian

Bagian ini menggambarkan secara umum tahapan yang dilakukan oleh peneliti selama proses penelitian. Alur ini mencakup seluruh proses mulai dari tahap awal hingga diperolehnya hasil dan kesimpulan. Visualisasi dari keserluruhan tahapan dapat dilihat pada Gambar 1. Visualisasi bertujuan untuk memudahkan pemahaman terhadap urutan dan keterkaitan antar proses dalam penelitian.



Gambar 1. Diagram Alur Penelitian

Langkah-langkah yang dilakukan pada penelitian ini dimulai dari pengumpulan data yang akan digunakan sebagai objek penelitian. Selanjutnya dilakukan tahap *preprocessing* untuk membersihkan dan menyiapkan data agar dapat diolah secara optimal oleh model. Proses ini umumnya mencakup penghapusan karakter-karakter khusus, penghilangkan kata umum, dan konversi semua karakter huruf menjadi huruf kecil. Setelah melalui tahap *preprocessing* data kemudian dibagi menjadi dua yaitu data *training* dan data *testing.* Selanjutnya data akan dilakukan tokenisasi oleh IndoBERT yang kemudian akan diubah menjadi representasi vektor. Representasi vektor tadi akan digunakan sebagai input arsitektur CNN yang bertugas melakukan klasifikasi.

2.1. Pengumpulan Data

Penelitian ini menggunakan data berupa artikel berita berbahasa Indonesia yang diperoleh secara otomatis menggunakan metode *web scrapping* dari situs berita daring Kompas.com. Pemilihan Kompas.com didasarkan pada reputasinya sebagai salah satu situ portal berita terpercaya dan terbesar di Indonesia. Dalam penelitian ini, data dikumpulkan dari enam kategori utama yang tersedia di Kompas.com yaitu: *News, Money,* Bola, *Health,* Tekno, dan Tren. Jumlah data yang dikumpulkan berjumlah 3.000 data teks, dengan masing-masing kategori terdiri dari 500 data . Data kemudian dibagi menjadi dua bagian, yaitu data latih dan data uji. Untuk mengevaluasi performa model secara menyeluruh pembagian data latih: data uji dilakukan dalam beberapa variasi rasio, yaitu 60:40, 70:30, 80:20, dan 90:10.

2.2. Preprocessing

Pada penelitian ini, tahap *preprocessing* dilakukan dengan beberapa tahapan yang bertujuan untuk membersihkan data agar lebih mudah diproses oleh model klasifikasi. Adapun tahapan *preprocessing* yang dilakukan adalah sebagai berikut:

a. *Case-Folding*
   Seluruh karakter huruf dalam artikel berita diubah menjadi huruf kecil untuk menyamakan

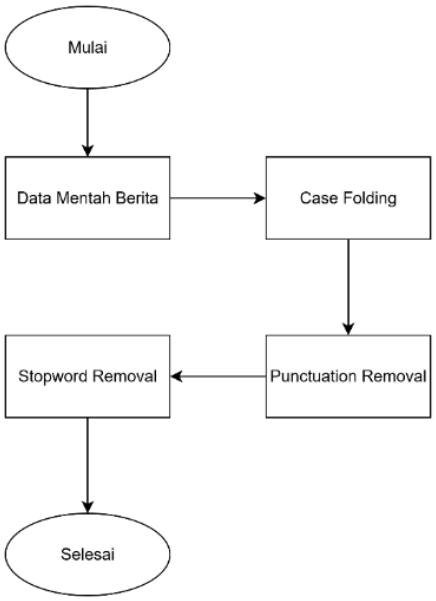
representasi kata [6].

b. Tokenisasi
Proses memecah teks menjadi satuan kata atau token, yang nantinya akan digunakan sebagai input model klasifikasi [7].

c. *Stopword Removal*
Menghapus kata-kata umum (“yang”, “di”, “dan”, “ke”, dan sebagainya) yang tidak memberikan kontribusi pada proses analisis data [8].

d. *Punctuation Removal*
Menghapus simbol atau tanda baca yang ada di dalam artikel berita, dikarenakan tanda baca pada umumnya tidak memberikan kontribusi signifikan dalam klasifikasi teks [9].

e. *Label Encoding*
Mengubah kategori teks yang sebelumnya berbentuk “news”, “money”, “bola”, “health”, “tekno”, dan “tren” menjadi bentuk numerik. Proses ini dilakukan karena algoritma klasifikasi hanya dapat memproses data numerik [10]

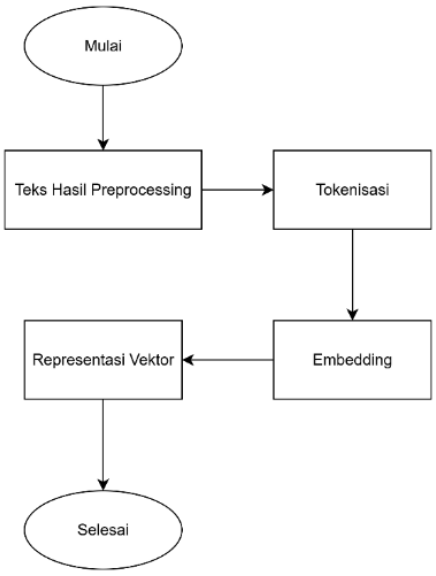


Gambar 2. Diagram Alur Preprocessing

2.3. IndoBERT Embedding

IndoBERT adalah model *pre-trained* dari arsitektur BERT(Bidirectional Encoder Representations from Transformers) yang telah dilatih secara khusus menggunakan korpus berbahasa Indonesia seperti Wikipedia dan beberapa artikel berita dalam skala besar [11]. Dengan kemampuannya untuk memahami konteks kata dari kedua arah (kiri dan kanan) dalam sebuah kalimat memungkinkannya untuk menangkap makna kata secara lebih menyeluruh, berbeda dengan metode seperti Word2vec atau GloVe yang menghasilkan kata satu per satu [12]. Pendekatan ini menjadikan IndoBERT efektif dalam berbagai pengolahan bahasa, khususnya pada klasifikasi teks, sehingga model ini banyak digunakan dalam berbagai penelitian klasifikasi teks berbahasa Indonesia.

p-ISSN: 2986-3929
e-ISSN: 3032-1948



Gambar 3. Diagram Alur IndoBERT

2.4. Convolutional Neural Network (CNN)

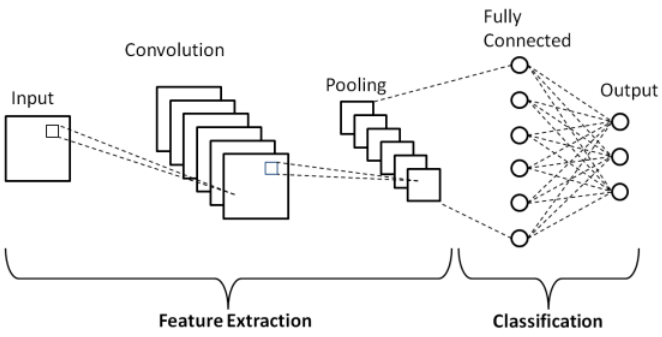
Convolutional Neural Network (CNN) merupakan pengembangan dari *multilayer perceptron* yang termasuk ke dalam *Deep Neural Network.* CNN memiliki kedalaman jaringan yang tinggi dikarenakan memiliki sebuah lapisan bernama *convolution layer[13].* Secara umum, CNN terdiri dari beberapa lapisan utama yakni:

a. *Convolution Layer*
Berfungsi untuk mengekstraksi fitur lokal dari input dengan menggunakan kernel. Proses ini akan menghasilkan *feature map* yang menggambarkan pola-pola dalam data [14].

b. *Pooling Layer*
Bertugas melakukan reduksi dimensi pada *feature map* melalui metode seperti *max pooling* atau *average pooling*. *Pooling layer* bertujuan untuk mengurangi kompleksitas komputasi dan menghindari *overfitting [14] [15]*.

c. *Fully Connected Layer*
Lapisan yang menghubungkan semua neuron dan bertanggung jawab dalam proses klasifikasi akhir berdasarkan fitur-fitur yang telah diekstrak oleh lapisan sebelumnya [14] [15].

Alur algoritma CNN digambarkan pada Gambar 4.



Gambar 4. Arsitektur CNN

Proses dimulai dari input data yang dimasukkan ke dalam *convolution layer*, di mana beberapa filter diterapkan untuk mengekstraksi fitur lokal dari data[14]. Hasil proses ini disebut *feature map,* selanjutnya *feature map* akan diproses oleh *pooling layer* yang berfungsi untuk mereduksi dimensi data dan mengurangi kompleksitas komputasi dengan tetap mempertahankan informasi utama[14][15]. Setelah melalui beberapa kali konvolusi dan proses *pooling*, representasi data yang telah didapatkan akan diteruskan ke *fully connected layer.* Pada tahap inilah semua neuron saling terhubung untuk memproses hasil ekstraksi fitur dan menghasilkan output akhir berupa klasifikasi[14][15].

2.5. Evaluasi

Tahap ini diperlukan untuk mengetahui performa model yang digunakan. Penilaian performa dilakukan dengan membandingkan hasil prediksi model terhadap label pada dara uji. Salah satu metode penilaian performa yang digunakan adalah *confusion matrix* yang digunakan untuk merepresentaskan jumlah prediksi yang benar dan salah dihitung ke dalam *True Positive*(TP), False Positive (FP), True Negative (TN), dan False Negative (FN) [16]. Nilai besar dan salah tersebut akan dihitung secara berturut-turut melalui persamaan *accuracy, precision, recall,* dan f-1 score.

Tabel 1. Confusion Matrix

Kelas Sebenarnya	Prediksi Positif	Prediksi Negatif
Positif	TP	FN
Negatif	FP	TN

Keterangan:

TP = Jumlah data yang diprediksi benar untuk kelas positif
TN = Jumlah data yang diprediksi benar sebagai bukan kelas positif
FP = Jumlah data yang diprediksi salah sebagai kelas positif
FN = Jumlah data yang diprediksi salah sebagai bukan kelas positif

Rumus yang dapat digunakan untuk menghitung metrik evaluasi adalah sebagai berikut:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F-1\ Score = \frac{2 \times precision \times recall}{precision+recall} \quad (3)$$

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

3. Hasil dan Diskusi

Evaluasi performa model dilakukan untuk mengetahui seberapa baik model dalam melakukan klasifikasi artikel berita, berdasarkan rasio pembagian data latih dan data uji yang berbeda. Dalam penelitian terda empat rasio yang diuji yakni: 60:40, 70:30, 80:20, dan 90:10, pemisahan rasio ini bertujuan untuk mengevaluasi pengaruh variasi proporsi data terhadap hasil klasifikasi. Setiap rasio akan diuji menggunakan metrik evaluasi yang mencakup akurasi, *precision, recall,* dan F-1 *Score* Hasil evaluasi terhadap keempat skenario ditunjukkan pada Tabel 2.

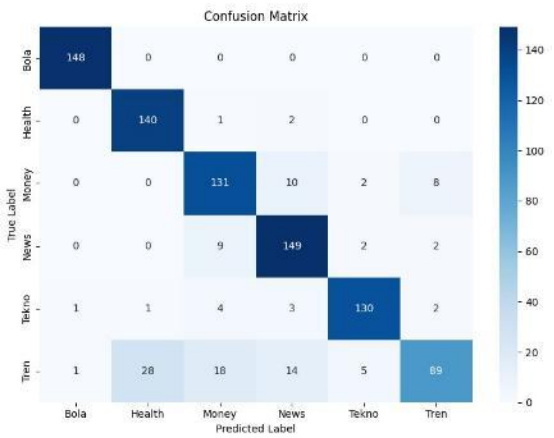
p-ISSN: 2986-3929
e-ISSN: 3032-1948

Tabel 2. Hasil Evaluasi Model CNN

Rasio Data	Akurasi	Precision	Recall	F-1 Score
60 : 40	0,84	0,85	0,83	0,83
70 : 30	0,87	0,88	0.88	0,87
80 : 20	0,91	0,92	0,91	0,91
90 : 10	0,90	0.90	0,90	0,90

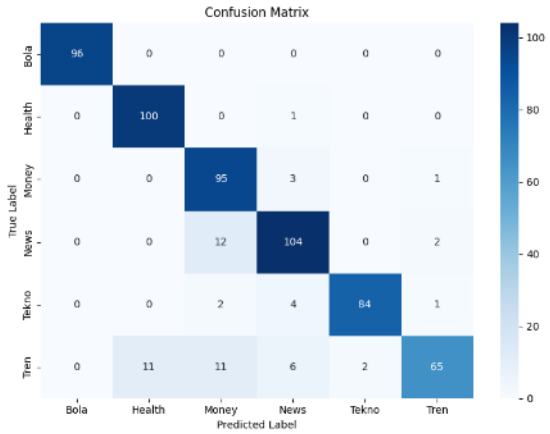
Secara umum, terlihat bahwa performa model cendrerung meingkat seiring dengan bertambahnya proporsi data latih. Rasio 80:20 menghasilkan performa terbaik dengan nilai akurasi sebesar 91%, sementara rasio 60:40 menunjukkan performa terendah, dengan akurasi 84%. Performa yang diperoleh pada rasio 80:20 menunjukkan bahwa model membutuhkan proporsi data latih dan data uji yang besar agar model dapat mempelajari pola-pola dalam data secara efektif. Hal ini dapat dibuktikan dengan melihat rasio 90:10 yang memiliki jumlah data latih yang lebih besar, namun nilai akurasi dan metrik lainnya justru lebih rendah dibandingkan performa rasio 80:20. Dapat dilihat bahwa performa model menurun seiring dengan penurunan performa dara latih. Akurasi tertinggi diperoleh pada rasio 80:20 yang mencapai akurasi sebesar 91%, sedangkan akurasi terendah diperoleh rasio 60:40 dengan akurasi 87%. Hal ini menunjukkan bahwa model memerlukan proporsi data latih yang besar untuk dapat mempelajari pola dengan baik. Untuk memahami performa model secara mendalam berdasarkan masing-masing kategori berita, dilakukan analisis menggunakan *confusion matrix* yang ditampilkan pada Gambar 5(a) hingga Gambar 5(d). Setiap *confusion matrix* merepresentasikan hasil prediksi model terhadap data uji untuk masing-masing rasio data latih dan uji yang digunakan. Representasi ini dilakukan untuk mempermudah proses identifikasi distribusi kesalahan klasifikasi yang terjadi di tiap kategori.

Gambar 5(a). Confusion Matrix Rasio 60:40

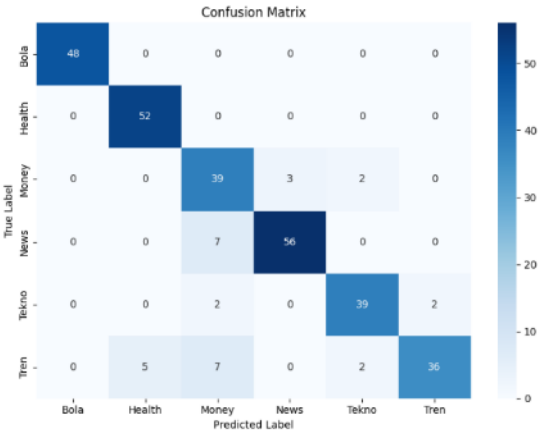


Gambar 5(b). Confusion Matrix Rasio 70:30

p-ISSN: 2986-3929
e-ISSN: 3032-1948



Gambar 5(c). Confusion Matrix Rasio 80:20



Gambar 5(d). Confusion Matrix Rasio 90:10

Dari visualisasi tersebut, dapat dilihat bahwa kategori "bola" dan "health" secara konsisten menunjukkan performa klasifikasi yang baik di semua rasio data. Sebaliknya kategori "tren" menunjukkan tingkat kesalahan klasifikasi yang relatif tinggi pada semua rasio, terutama pada rasio 60:40 dan 70:30. Pada rasio 80:20 distribusi klasifikasi terlihat lebih stabil, menunjukkan bahwa model berhasil mengenali pola fitur dengan baik. Meskipun rasio 90:10 memiliki proporsi data latih lebih besar, jumlah data uji yang terlalu sedikit membuat evaluasi kurang representatif. Oleh karena itu, *confusion matrix* mendukung hasil kinerja CNN yang menunjukkan bahwa rasio 80:20 memberikan hasil klasifikasi paling seimbang dan akurat pada semua kategori.

4. Kesimpulan

Berdasarkan hasil penelitian dapat disimpulkan bahwa kombinasi algoritma klasifikasi *Convolutional Neural Network* (CNN) dan IndoBERT sebagai metode embedding mampu memberikan performa yang baik dalam mengklasifikasikan artikel berita berbahasa Indonesia. Model menunjukkan performa terbaik pada skenario pembagian data 80:20 dengan akurasi mencapai 91%. Hasil ini menunjukkan bahwa representasi vektor dari IndoBERT mampu menangkap konteks Secara keseluruhan penelitian ini berhasil membuktikan bahwa CNN dan IndoBERT dapat menjadi pendekatan yang efektif dalam pengembangan sistem klasifikasi teks berbahasa Indonesia secara otomatis.

Daftar Pustaka

[1] M. Kamayani and D. Mugisidi, "Information Technology Uses In Research: Best Practices and Recommendations," 2016. [Online]. Available: https://www.researchgate.net/publication/316714784

[2] A. Zainudin, P. S.K., N. A. Zulkefli, and S. Raghavan, "From Print to Screen: Motivational Factors Influencing Online Newspaper Consumption among Open and Distance Learners," *International Journal of Academic Research in Business and Social Sciences*, vol. 15, no. 5, May 2025, doi: 10.6007/IJARBSS/v15-i5/25454.

[3] A. Irma Purnamasari and A. Rinaldi Dikananda, "Klasifikasi Kualitas Berita Pada Majalah Menggunakan Metode Decision Tree," *Jurnal Teknologi Ilmu Komputer*, vol. 1, no. 2, pp. 48–54, 2023, doi: 10.56854/jtik.v1i2.52.

[4] M. A. Ramdhani, M. A. Ramdhani, D. S. adillah Maylawati, and T. Mantoro, "Indonesian news classification using convolutional neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 2, pp. 1000–1009, Aug. 2020, doi: 10.11591/ijeecs.v19.i2.pp1000-1009.

[5] V. Prisscilya and A. S. Girsang, "Classification of Indonesia False News Detection Using Bertopic and Indobert," *Jurnal Indonesia Sosial Teknologi*, vol. 5, no. 8, 2024, [Online]. Available: http://jist.publikasiindonesia.id/

[6] D. F. AL-Hafiidh, I. F. Rozi, and I. K. Putri, “Peringkasan Teks Otomatis pada Portal Berita Olahraga menggunakan metode Maximum Marginal Relevance.,” *Jurnal Informatika Polinema*, vol. 8, no. 3, pp. 21–30, Jun. 2022, doi: 10.33795/jip.v8i3.519.
[7] A. T. B. Panjaitan and I. Santoso, “Deteksi Hoaks Pada Berita Berbahasa Indonesia Seputar COVID-19,” *Format : Jurnal Ilmiah Teknik Informatika*, vol. 10, no. 1, p. 76, Feb. 2021, doi: 10.22441/format.2021.v10.i1.007.
[8] A. A. Kurniawan and M. Mustikasari, “Implementasi Deep Learning Menggunakan Metode CNN dan LSTM untuk Menentukan Berita Palsu dalam Bahasa Indonesia,” vol. 5, no. 4, pp. 2622–4615, 2020, doi: 10.32493/informatika.v5i4.7760.
[9] I. Pakpahan and Jasman Pardede, “Analisis Sentimen Penanganan Covid-19 Menggunakan Metode Long Short-Term Memory Pada Media Sosial Twitter,” *Jurnal Publikasi Teknik Informatika*, vol. 2, no. 1, pp. 12–25, Jan. 2023, doi: 10.55606/jupti.v1i1.767.
[10] C. Herdian, A. Kamila, and I. G. Agung Musa Budidarma, “Studi Kasus Feature Engineering Untuk Data Teks: Perbandingan Label Encoding dan One-Hot Encoding Pada Metode Linear Regresi,” *Technologia : Jurnal Ilmiah*, vol. 15, no. 1, p. 93, Jan. 2024, doi: 10.31602/tji.v15i1.13457.
[11] F. Indriani, R. A. Nugroho, M. R. Faisal, and D. Kartini, “Comparative Evaluation of IndoBERT, IndoBERTweet, and mBERT for Multilabel Student Feedback Classification,” *Jurnal RESTI*, vol. 8, no. 6, pp. 748–757, Dec. 2024, doi: 10.29207/resti.v8i6.6100.
[12] J. Hauschild and K. Eskridge, “Word embedding and classification methods and their effects on fake news detection,” *Machine Learning with Applications*, vol. 17, p. 100566, Sep. 2024, doi: 10.1016/j.mlwa.2024.100566.
[13] A. R. Maulana and N. Rochmawati, “Opinion Mining Terhadap Pemberitaan Corona di Instagram menggunakan Convolutional Neural Network,” *Journal of Informatics and Computer Science*, vol. 02, 2020.
[14] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, “A review of convolutional neural networks in computer vision,” *Artif Intell Rev*, vol. 57, no. 4, Apr. 2024, doi: 10.1007/s10462-024-10721-6.
[15] M. M. Taye, “Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions,” Mar. 01, 2023, *MDPI*. doi: 10.3390/computation11030052.
[16] H. Imaduddin, F. Yusfida A’la, and Y. S. Nugroho, “Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach.” [Online]. Available: www.ijacsa.thesai.org