Jurnal Elektronik Ilmu Komputer Udayana
Volume 11, No 3. February 2023

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Article Classification Using Convolutional Neural Network (CNN) And Chi-Square Feature Selection

I Gede Laksmana Yudha[a1], Ngurah Agus Sanjaya ER[a2], Anak Agung Istri Ngurah Eka Karyawati[a3], Ida Bagus Gede Dwidasmara[a4], I Gusti Ngurah Anom Cahyadi Putra[a5], Ida Bagus Made Mahendra[a6]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Udayana
Bali, Indonesia
[1]laksmanayudha22@gmail.com
[2]agus_sanjaya@unud.ac.id
[3]eka.karyawati@unud.ac.id
[4]dwidasmara@unud.ac.id
[5]anom.cp@unud.ac.id
[6]ibm.mahendra@unud.ac.id

Abstract

News articles are reports or information about events that are actual, reliable, and based on facts or reality. The increase in internet users has resulted in the growth of the amount of available information increasing rapidly. Easy internet access makes many types of Indonesian news articles published digitally. With a very large number of news articles, it will be easier to find a news article if the news has been organized and has been grouped according to its respective categories. Text classification is a problem that aims to determine the topic or theme of a document. In achieving this goal, the classification process forms a model that can distinguish data into different classes based on certain rules. The method used to build the model is Convolutional Neural Network (CNN) with Chi-Square feature selection. News articles are divided into six classes, namely news, technology, football, health, lifestyle, and automotive. In this study, the best CNN model was obtained with the number of filters used was 200 and the feature selection being 40%. The test results on the test data provide an accuracy value of 96,074%, precision of 96,079%, recall of 96,074%, and an f-1 score of 96,070%.

Keywords: *Convolutional Neural Network, Text Classification, Chi-Square, K-Fold Cross Validation, Article*

1. Pendahuluan

Artikel berita adalah laporan atau informasi mengenai kejadian/peristiwa yang bersifat aktual, dapat dipercaya, dan berdasarkan fakta atau realita. Karena kemajuan teknologi dan internet, artikel berita semakin sedikit ditemui pada media cetak seperti koran dan banyak beralih pada media digital misalnya website. Peningkatan jumlah informasi yang pesat diakibatkan oleh meningkatnya pula penggunaan internet oleh masyarakat. Mudahnya akses internet menyebabkan beragam jenis artikel berita Bahasa Indonesia secara digital banyak dipublikasikan, hal ini dapat memberi manfaat dalam akses informasi berbahasa Indonesia secara mudah dan cepat. Namun, dengan meningkatnya jumlah artikel berita Bahasa Indonesia menimbulkan masalah lain yaitu pada pengkategorian topik pada setiap artikel berita yang dapat memakan waktu.

Terdapat beberapa permasalahan dalam pengelompokan dokumen, salah satunya yaitu klasifikasi. Selain klasifikasi, klasterisasi juga merupakan permasalahan dalam pengelompokan dokumen [1]. Namun pada kasus kali ini, klasifikasi menjadi permasalahan yang tepat, karena dokumen ditentukan berdasarkan kategori yang telah ditentukan. Klasifikasi teks merupakan permasalahan yang memiliki tujuan untuk menentukan topik atau tema suatu dokumen [2]. Untuk menentukan topik ke dalam kelasnya masing - masing, maka pada proses klasifikasi

akan dibentuk suatu model berdasarkan aturan tertentu sehingga dapat membedakan topik dari suatu dokumen.

Berbagai macam metode telah digunakan dalam penelitian - penelitian sebelumnya terkait dengan klasifikasi teks. Salah satunya penelitian yang dilakukan oleh Ramdhani yaitu klasifikasi berita indonesia menggunakan CNN. Hasil yang diperoleh dari pengujian adalah akurasi terbaik sebesar 90,74% dan nilai loss sebesar 29,05% [3].

Terdapat beberapa cara yang dilakukan untuk dapat meningkatkan performa algoritma klasifikasi. Salah satunya dengan menggunakan seleksi fitur. Pada jurnal Suharno melakukan klasifikasi menggunakan *K-Nearest Neighbors* dan *Chi-Square* sebagai seleksi fitur pada dokumen pengaduan Sambat. Dari hasil pengujian dengan seleksi fitur sebesar 25%, didapatkan nilai *precision*, *recall*, dan *f-measure* secara berturut - turut sebesar 90%, 78%, dan 78%. Dari hasil penelitian tersebut, nilai *f-measure* dapat ditingkatkan melalui penggunaan seleksi fitur dan Metode KNN pada klasifikasi dokumen pengaduan SAMBAT[4].

Kemudian, menurut Farid pada penelitiannya tentang deteksi hoax pada twitter dengan metode CNN dan seleksi fitur information gain. Dalam penelitian ini Farid et al (2020) menyimpulkan bahwa metode *Convolutional Neural Network* (CNN) dapat dengan baik mengklasifikasikan berita hoax. Penggunaan TF-IDF dan *Information Gain* untuk seleksi fitur juga sangat mempengaruhi hasil klasifikasi karena pada pengujian diperoleh nilai akurasi tertinggi sebesar 95,56% [5].

Sehingga dalam melakukan penelitian ini, penulis berlandaskan pada permasalahan dan penelitian terkait yang telah disebutkan. Sehingga, penulis bermaksud untuk melakukan implementasi algoritma *Convolutional Neural Network* (CNN) pada klasifikasi artikel berita bahasa Indonesia dengan seleksi fitur *Chi-Square*.

2. Metodologi Penelitian

Pada metode penelitian, secara umum terdapat alur penelitian mulai dari input data yang berupa teks berita, yang setelah itu akan dilakukan tahap prepocessing dan seleksi fitur *Chi-Square*. Kemudian, dilakuakn TF-IDF untuk mengubah bentuk data dokumen menjadi vektor sebelum dimasukan ke algoritma CNN. Selanjutnya akan masuk ke tahap klasifikasi *Convolutional Neural Network* (CNN) yang menghasilkan output berupa nama kelas dari probabilitas tertinggi hasil klasifikasi. Kemudian untuk pengukuran performa model digunakan *k-fold cross validation* pada tahap evaluasi.

2.1. Pengumpulan Data

Data yang digunakan adalah artikel berita berbahasa Indonesia yang bersumber dari situs Kompas.com, hal ini dikarenakan situs-situs tersebut merupakan situs portal berita terpercaya dan situs-situs tersebut memiliki beberapa kategori berita yang sama dan kedua website tersebut dirasa cukup untuk memenuhi kebutuhan jumlah data. Data diperoleh secara sekunder menggunakan web-scraping. Kelas yang digunakan adalah kategori yang telah tersedia pada situs berita sebanyak enam kategori yaitu news, teknologi, bola, health, lifestyle, dan otomotif. Kelas – kelas tersebut dipilih berdasarkan kategori yang paling sering muncul pada kebanyakan situs portal berita. Jumlah data yang akan digunakan adalah 13.500 data teks berita dengan 2.250 data pada setiap kelasnya.

2.2. Preprocessing

Untuk menyiapkan data menjadi siap diolah dan bersih maka memerlukan proses preprocessing. Urutan proses *preprocessing* secara berturut turut dimulai dari input data dokumen, *case folding*, *punctuation removal*, stemming, tokenisasi, stopword removal. Pada tabel 1 diperlihatkan contoh proses *preprocessing* pada teks.
Pada *case folding* dokumen diubah ke dalam bentuk huruf kecil, untuk menyeragamkan ukuran teks. Sehingga, kata yang sama dengan salah satu kata menggunakan huruf kapital akan dianggap berbeda. Kata ‘Sukses’ akan sama dengan ‘sukses’ pada *case folding*. Selanjutnya, untuk menghilangkan tanda baca dilakukan punctuation removal, agar mempermudah proses tokenisasi. Kemudian, kata - kata yang memiliki imbuhan akan diubah ke bentuk dasarnya melalui proses stemming. Selanjutnya, kata - kata dipecah menjadi potongan token pada

p-ISSN: 2301-5373
e-ISSN: 2654-5101

proses tokenisasi. Terakhir, dilakukan penghapusan kata - kata umum yang tidak memberikan perbedaan konten secara signifikan melalui proses stopword removal [6].

Tabel 1. *Text Preprocessing*

Preprocessing	Hasil
Data Awal	Setelah sukses dengan penjualan perdana terbatas pada model sebelumnya, Katalis kembali luncurkan skuter listrik yang unik. Motor ramah lingkungan ini disebut terinspirasi dari robot.
Case Folding	setelah sukses dengan penjualan perdana terbatas pada model sebelumnya, katalis kembali luncurkan skuter listrik yang unik. motor ramah lingkungan ini disebut terinspirasi dari robot.
Punctuation Removal	setelah sukses dengan penjualan perdana terbatas pada model sebelumnya katalis kembali luncurkan skuter listrik yang unik motor ramah lingkungan ini disebut terinspirasi dari robot
Stemming	telah sukses dengan jual perdana batas pada model belum katalis kembali luncur skuter listrik yang unik motor ramah lingkung ini sebut inspirasi dari robot
Tokenisasi	['telah', 'sukses', 'dengan', 'jual', 'perdana', 'batas', 'pada', 'model', 'belum', 'katalis', 'kembali', 'luncur', 'skuter', 'listrik', 'yang', 'unik', 'motor', 'ramah', 'lingkung', 'ini', 'sebut', 'inspirasi', 'dari', 'robot']
Stopword	['sukses', 'jual', 'perdana', 'batas', 'model', 'katalis', 'luncur', 'skuter', 'listrik', 'unik', 'motor', 'ramah', 'lingkung', 'inspirasi', 'robot']

2.3. Seleksi Fitur

Seleksi fitur juga diperlukan dalam klasifikasi, ini sangat penting dalam klasifikasi teks karena tingginya dimensi fitur teks dan keberadaan fitur yang tidak relevan (*noisy*) [7]. Chi-Square adalah salah satu teknik filter dalam seleksi fitur [8]. Hasil klasifikasi dipengaruhi oleh jumlah fitur yang digunakan, seleksi fitur membantu dalam mengurangi jumlah fitur yang dianggap tidak relevan pada kumpulan dokumen. Sehingga, kinerja metode klasifikasi dapat ditingkatkan efektifitas dan efisiensinya. Pengujian derajat kepentingan sebuah term terhadap kategorinya dilakukan menggunakan seleksi fitur *chi-square*. Berikut ini merupakan fungsi persamaan dari Chi-Square :

$$x^2(t,c) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (1)$$

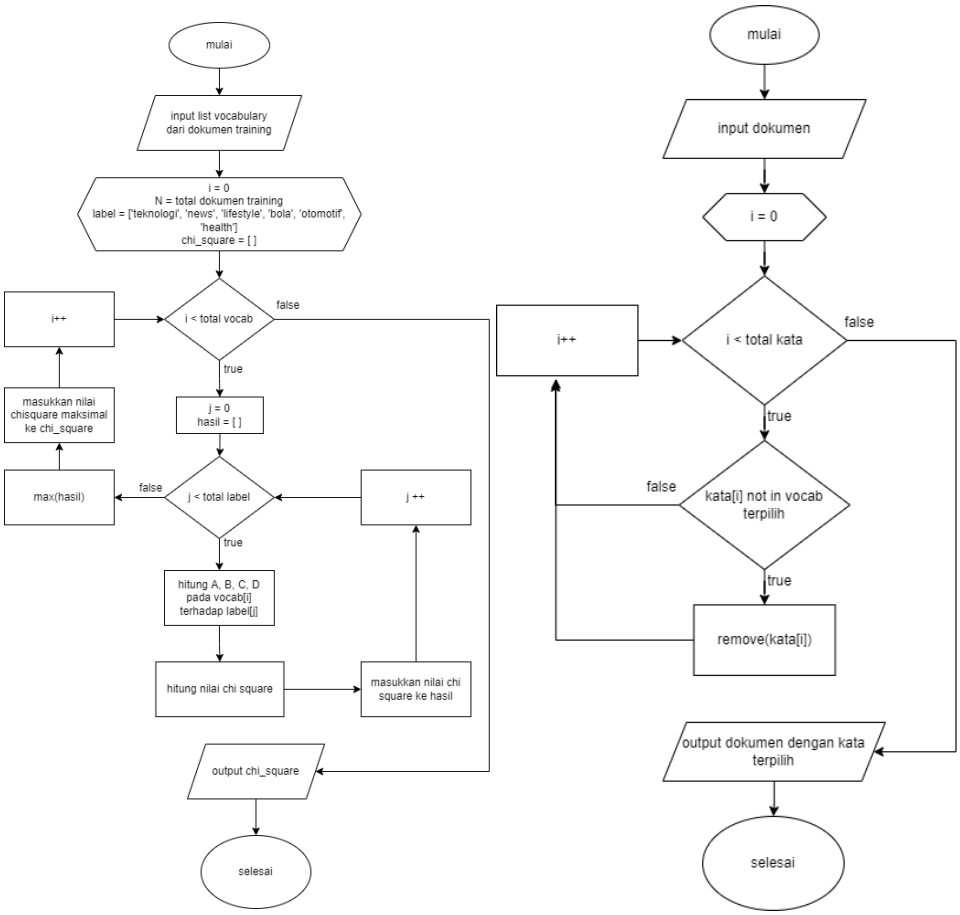
keterangan :
t = Term
c = Kelas/kategori topik dokumen
N = Total dokumen latih/*training*
A = Total dokumen pada kelas/kategori topik c yang mengandung t
B = Total dokumen bukan kelas/kategori topik c yang mengandung t
C = Total dokumen pada kelas/kategori topik c yang tidak mengandung t
D = Total dokumen bukan kelas/kategori topik c yang tidak mengandung t

Nilai Chi-Square ini dapat diglobalisasikan pada semua kategori dengan memilih nilai maksimal atau dengan menggunakan nilai rata – rata di antara semua kategori [4].

$$x^2{}_{avg}(t) = \sum_{i=1}^{m} P(c_i) \,.\, x^2(t,c_i) \quad (2)$$

$$x^2{}_{max}(t) = max_{i=1}^{m}\{x^2(t, c_i)\} \quad (3)$$

Gambar 1 adalah flowchart untuk menghitung nilai chi-square tunggal untuk setiap kata pada vocabulary. Suatu kata i pada satu kategori akan dihitung nilai A, B, C, D sesuai dengan persamaan (1). Kemudian, nilai – nilai tersebut digunakan untuk menghitung nilai *chi-square* menggunakan persamaan (1). Kemudian, perhitungan nilai *chi-square* dilanjutkan untuk semua kategori. Setelah mendapatkan nilai *chi-square* kata i dengan semua kategori, selanjutnya dipilih nilai chi-square maksimal yang akan digunakan sebagai nilai *chi-square* tunggal untuk kata i. Proses ini dilakukan terhadap semua kata yang ada pada *vocabulary*. Pada proses akhir, kata – kata pada *vocabulary* akan diurutkan berdasarkan nilai *chi-square* tertinggi ke terendah. Kemudian pada *flowchart*, kata – kata dalam dokumen yang tidak termasuk dalam kata – kata rasio seleksi fitur akan dihilangkan.



Gambar 1. Seleksi Fitur *Chi-Square*

2.4. Pembobotan TF-IDF

Setelah melalui tahap *preprocessing*, agar dapat diproses data perlu diubah ke dalam bentuk numerik. TF-IDF mentransformasi data hasil preprocessing ke dalam bentuk numerik yaitu vektor. Pembobotan ini menentukan tingkat hubungan kata terhadap dokumen dengan cara memberikan nilai bobot pada setiap kata. TF-IDF mengkombinasikan konsep tingkat kemunculan term atau kata pada sebuah dokumen (*term frequency*) dan tingkat kepentingan dari term atau kata pada dokumen (*inverse document frequency*) [9]. Persamaan dari TF-IDF dapat dilihat pada rumus (4), (5), dan (6).

p-ISSN: 2301-5373
e-ISSN: 2654-5101

a. Term Frequency

$$tf_{ik} = \frac{f_{ik}}{\sum_{j=1}^{t} f_{ij}} \quad (4)$$

yang mana, f_{ik} merupakan kemunculan kata atau term *k* pada dokumen ke-i.

b. Inverse Document Frequency

$$idf_k = log\frac{N}{n_k} \quad (5)$$

dimana *N* adalah jumlah total dokumen, n_k merupakan jumlah total dokumen dengan kemunculan kata atau term *k*.

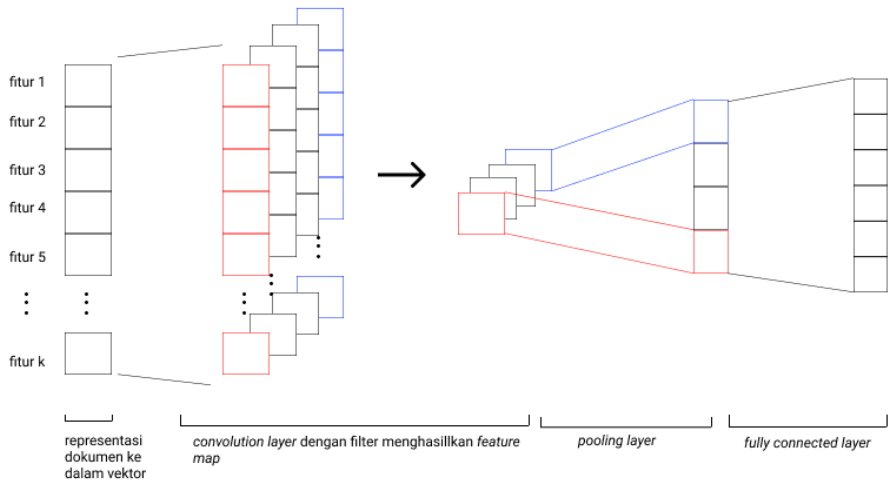
c. Term Frequency - Inverse Document Frequency

Kombinasi dari nilai TF dan IDF akan menghasilkan nilai TF-IDF.

$$W_{ik} = tf_{ik} \times idf_k \quad (6)$$

2.5. Convolutional Neural Network

Algoritma CNN merupakan metode pengembangan *multilayer perceptron* yang memiliki tingkat kedalaman jaringan yang tinggi sehingga tergolong ke salah satu jenis *Deep Neural Network* [10]. CNN umumnya terdiri dari beberapa lapisan, yaitu lapisan konvolusi, lapisan *pooling*, dan lapisan *fully connected*. Data yang menjadi masukkan pada CNN akan dipelajari fiturnya pada lapisan konvolusi dan *pooling*, yang selanjutnya akan dilakukan klasifikasi pada lapisan *fully connected*. Menurut Goldberg, ide utama di balik konvolusi dan *pooling* adalah untuk menerapkan filter pada setiap instansiasi dari h-word sliding window pada kalimat [11].



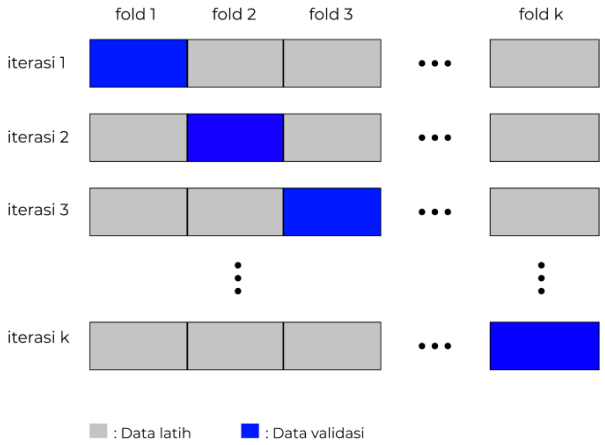
Gambar 2. Arsitektur CNN

Salah satu ciri utama dari CNN dengan algoritma lainnya adalah lapisan konvolusi. Lapisan ini menghasilkan *feature map* dengan menggeser sebuah filter yang melakukan proses konvolusi terhadap data input pada lapisan sebelumnya.

Setelah *feature map* dihasilkan, maka selanjutnya feature map tersebut menjadi masukan pada lapisan pooling. Lapisan pooling bertugas untuk meringkas/mereduksi dimensi dari feature map, sehingga ukuran data menjadi lebih kecil yang berpengaruh pada kecepatan komputasi. Max Pooling adalah salah satu contoh dari lapisan pooling yang memperoleh informasi penting dari feature map dengan mengambil nilai maksimal pada elemen-elemen yang berada pada lingkup window [12].

Kemudian, lapisan fully connected akan melakukan proses klasifikasi terhadap fitur - fitur yang dihasilkan dari proses sebelumnya. Lapisan ini terdiri atas lapisan input, lapisan tersembunya atau hidden layer, dan lapisan output yang masing - masing lapisan memiliki neuron – neuron yang saling terhubung. Proses pemetaan fitur ke dalam kategori akan dibantu dengan nilai bobot yang dimiliki oleh setiap neuron - neuron tersebut.

2.6. K-Fold Cross Validation



Gambar 3. *K-Fold Cross Validation*

K-Fold Cross Validation membagi keseluruhan dataset dalam k bagian, sehingga terdapat sebanyak k iterasi yang setiap iterasinya secara bergantian bagian – bagian data akan menjadi data latih dan data validasi [13]. Gambar 3 merupakan Ilustrasi dari *k-fold cross validation*.

2.7. Evaluasi

Pengukuran performa didasarkan pada jumlah pengujian yang diprediksi dengan jumlah benar dan salah oleh model yang ditabulasikan dalam tabel yang disebut sebagai confusion matrix [14]. Kemudian nilai benar dan salah tersebut dihitung kedalam TP, FN, FP, dan TN yang selanjutnya secara berturut – turut diukur melalui persamaan akurasi, *precision, recall,* dan *f-1 score* melalui rumus (7), (8), (9), dan (10). Tabel 2 adalah contoh dari *confusion matrix*.

Tabel 2. *Confusion Matrix*

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

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$F - 1\ Score = \frac{2 \times precision \times recall}{precision + recall} \quad (10)$$

p-ISSN: 2301-5373
e-ISSN: 2654-5101

True Positive (TP) terjadi ketika data positif berhasil dengan benar diklasifikasikan oleh model. *False Negative* (FN) terjadi ketika data positif diklasifikasikan salah sebagai data negatif oleh model. *False Positive* (FP) terjadi ketika data negatif diklasifikasikan salah sebagai data positif oleh model. *True Negative* (TN) terjadi ketika data negatif berhasil diklasifikasikan dengan benar oleh model.

3. Hasil dan Diskusi

Tabel 3. Hasil Uji Coba Data Validasi

Rasio	Filter	Akurasi	*Precision*	*Recall*	*F-1*
10 %	50	94,657%	94,664%	94,665%	94,642%
	100	94,926%	94,945%	94,937%	94,913%
	200	94,787%	94,805%	94,793%	94,774%
	400	94,889%	94,879%	94,902%	94,879%
20%	50	95,065%	95,058%	95,084%	95,061%
	100	94,870%	94,881%	94,876%	94,853%
	200	94,889%	94,885%	94,904%	94,876%
	400	95,065%	95,062%	95,080%	95,056%
30%	50	95,213%	95,213%	95,233%	95,204%
	100	95,315%	95,322%	95,327%	95,308%
	200	95,176%	95,188%	95,195%	95,164%
	400	95,232%	95,242%	95,242%	95,220%
40%	50	95,185%	95,181%	95,195%	95,177%
	100	95,278%	95,271%	95,297%	95,273%
	200	**95,315%**	**95,314%**	**95,336%**	**95,312%**
	400	95,157%	95,159%	95,178%	95,153%
50%	50	95,102%	95,101%	95,115%	95,100%
	100	95,120%	95,113%	95,134%	95,115%
	200	95,074%	95,080%	95,084%	95,067%
	400	95,065%	95,080%	95,088%	95,062%
60%	50	95,093%	95,093%	95,110%	95,095%
	100	95,037%	95,035%	95,051%	95,023%
	200	95,102%	95,114%	95,121%	95,098%
	400	95,204%	95,208%	95,222%	95,194%
70%	50	95,129%	95,122%	95,147%	95,125%
	100	95,009%	95,023%	95,013%	94,995%
	200	95,278%	95,283%	95,293%	95,269%
	400	95,000%	95,003%	95,019%	94,996%
80%	50	95,083%	95,079%	95,099%	95,076%
	100	95,102%	95,110%	95,117%	95,093%
	200	95,065%	95,057%	95,089%	95,062%
	400	95,074%	95,083%	95,090%	95,068%
90%	50	95,102%	95,100%	95,123%	95,096%
	100	95,111%	95,107%	95,124%	95,106%
	200	95,120%	95,121%	95,137%	95,115%
	400	95,167%	95,169%	95,192%	95,162%
100%	50	94,796%	94,801%	94,812%	94,789%
	100	94,870%	94,865%	94,885%	94,862%
	200	94,972%	94,975%	94,999%	94,972%
	400	94,778%	94,785%	94,786%	94,764%

Pada pengujian, kombinasi parameter yang digunakan adalah rasio seleksi fitur dan jumlah filter pada CNN. Rasio seleksi fitur adalah 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, dan 100% dari total fitur yang diperoleh. Kemudian Jumlah filter pada CNN adalah 50, 100, 200, dan 400. Pengujian menggunakan *k-fold cross validation* dengan nilai *k* yang digunakan adalah 5. Hasil *k-fold cross validation* selanjutnya akan dihitung rata – rata nilai evaluasi yaitu akurasi, *precision, recall*, dan *f-1 score* yang kemudian dipilih kombinasi parameter dengan nilai evaluasi terbaik.

Tabel 3 merupakan hasil dari pengujian dengan menggunakan *k-fold cross validation* dengan nilai akurasi, *precision, recall,* dan *f-1* score terhadap rasio seleksi fitur dan jumlah filter CNN yang digunakan. Berdasarkan hasil pengujian pada Tabel 3, didapat kombinasi parameter terbaik yaitu rasio seleksi fitur sebesar 40% dan jumlah filter sebanyak 200 dengan nilai rata – rata dari akurasi sebesar 95,315%, *precision* sebesar 95,314%, *recall* sebesar 95,336%, dan f-1 *score sebesar* 95,312%. Berdasarkan Tabel 3 didapat bahwa perubahan filter CNN dan rasio seleksi fitur cenderung tidak memberikan pengaruh yang signifikan terhadap akurasi model.

Tabel 4. Hasil Uji Coba Data Testing

Akurasi	***Precision***	***Recall***	***F-1***
96,074%	96,079%	96,074%	96,070%

Tabel 5. Pengaruh Filter CNN Terhadap Akurasi Model

Rasio Seleksi Fitur	**Filter CNN**	**Akurasi**
100%	50	94,796%
	100	94,870%
	200	94,972%
	400	94,778%

Tabel 6. Pengaruh Rasio Seleksi Fitur Terhadap Akurasi Model

Filter CNN	**Rasio Seleksi Fitur**	**Akurasi**
200	10%	94,787%
	20%	94,889%
	30%	95,176%
	40%	95,315%
	50%	95,074%
	60%	95,102%
	70%	95,278%
	80%	95,065%
	90%	95,120%
	100%	94,972%

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Setelah mendapatkan model dengan kombinasi parameter terbaik, selanjutnya model tersebut akan diujikan pada data testing. Tabel 4 adalah hasil uji coba model terhadap data *testing*. Berdasarkan hasil uji coba model terhadap data *testing*, nilai evaluasi rata – rata yang diperoleh adalah akurasi sebesar 96,074%, precision sebesar 96,079%, recall sebesar 96,074%, dan f-1 score sebesar 96,070%. Kemudian, mengenai pengaruh yang diberikan melalui perubahan hyperparameter filter CNN dan rasio seleksi fitur terhadap akurasi model dapat dilihat pada Tabel 5 dan 6.

Berdasarkan Tabel 5 dan Tabel 6 didapat bahwa perubahan rasio seleksi fitur dan filter CNN cenderung tidak memberikan pengaruh yang signifikan terhadap akurasi model.

Tabel 7. *Running Time Training* Terhadap Filter CNN

Rasio Seleksi Fitur	Filter CNN	*Running Time* (menit)
100%	50	12,47
	100	30,01
	200	124,05
	400	224,44

Tabel 8. *Running Time Training* Terhadap Rasio Seleksi Fitur

Filter CNN	Rasio Seleksi Fitur	*Running Time* (menit)
200	10%	2,70
	20%	6,78
	30%	12,14
	40%	21,08
	50%	38,58
	60%	44,91
	70%	86,39
	80%	95,87
	90%	121,55
	100%	124,05

Berdasarkan Tabel 7 dan Tabel 8 didapat bahwa semakin tinggi rasio seleksi fitur dan semakin banyak filter pada CNN yang digunakan maka semakin lama waktu yang diperlukan oleh model untuk proses training.

4. Kesimpulan

Berdasarkan hasil pengujian yang dilakukan, maka dapat ditarik kesimpulan bahwa:

a. Dari hasil validasi model dengan *k-fold cross validation* diperoleh bahwa model klasifikasi artikel berita Bahasa Indonesia terbaik adalah model *Convolutional Neural Network* (CNN) dengan jumlah filter 200 dan seleksi fitur sebesar 40%, dimana akurasi, *precision*, *recall*, dan *f-1 score* dari validasi adalah secara berturut – turut sebesar 95,315%, 95,314%, 95,336%, dan 95,312%.

b. Hasil *k-fold cross validation* menunjukkan kecenderungan bahwa perubahan rasio seleksi fitur tidak memberikan pengaruh yang signifikan terhadap performa model klasifikasi menggunakan *Convolutional Neural Network* (CNN), namun peningkatan rasio seleksi fitur dapat mempengaruhi running time pada sistem. Semakin tinggi rasio seleksi fitur yang digunakan maka semakin lama waktu yang diperlukan oleh model untuk proses training. Sehingga, seleksi fitur dapat mengurangi fitur yang dianggap kurang relevan untuk proses klasifikasi agar proses training menjadi lebih cepat. Hasil terbaik diperoleh pada saat seleksi fitur sebesar 40% dengan nilai akurasi, *precision*, *recall*, dan *f-1 score* model klasifikasi terhadap data *testing* (*unseen* data) adalah berturut-turut sebesar 96,074%, 96,079%, 96,074%, dan 96,070%.

.

Daftar Pustaka

[1] N. A. S. ER, “Implementasi Latent Dirichlet Allocation (Lda) Untuk Implementation of Latent Dirichlet Allocation (Lda) for,” vol. 8, no. 1, pp. 127–134, 2021, doi: 10.25126/jtiik.202183556.

[2] C. Manning and H. Schutze, *Foundations of statistical natural language processing*. MIT press, 1999.

[3] M. A. Ramdhani, M. A. Ramdhani, D. S. adillah Maylawati, and T. Mantoro, “Indonesian news classification using convolutional neural network,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 19, no. 2, pp. 1000–1009, 2020, doi: 10.11591/ijeecs.v19.i2.pp1000-1009.

[4] C. F. Suharno, M. A. Fauzi, and R. S. Perdana, “Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan Sambat Online Menggunakan Metode K-Nearest Neighbors Dan Chi-square,” *Syst. Inf. Syst. Informatics J.*, vol. 3, no. 1, pp. 25–32, 2017, doi: 10.29080/systemic.v3i1.191.

[5] H. K. Farid, E. B. Setiawan, and I. Kurniawan, “Selection for Hoax News Detection on Twitter using Convolutional Neural Network,” *Indones. J. Comput.*, vol. 5, no. December 2020, pp. 23–36, 2020, doi: 10.34818/indojc.2021.5.3.506.

[6] F. Taufiqurrahman, S. Al Faraby, and M. D. Purbolaksono, “Klasifikasi Teks Multi Label pada Hadis Terjemahan Bahasa Indonesia Menggunakan Chi Square dan SVM,” *e-Proceeding Eng.*, vol. 8, no. 5, pp. 10650–10659, 2021.

[7] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[8] I. Listiowarni and N. Puspa Dewi, “Pemanfaatan Klasifikasi Soal Biologi Cognitive Domain Bloom’s Taxonomy Menggunakan KNN Chi-Square Sebagai Penyusunan Naskah Soal,” *Digit. Zo. J. Teknol. Inf. dan Komun.*, vol. 11, no. 2, pp. 186–197, 2020, doi: 10.31849/digitalzone.v11i2.4798.

[9] K. D. Yonatha Wijaya and A. A. I. N. E. Karyawati, “The Effects of Different Kernels in SVM Sentiment Analysis on Mass Social Distancing,” *JELIKU (Jurnal Elektron. Ilmu Komput. Udayana)*, vol. 9, no. 2, p. 161, 2020, doi: 10.24843/jlk.2020.v09.i02.p01.

[10] A. R. Maulana and N. Rochmawati, “Opinion Mining Terhadap Pemberitaan Corona di Instagram menggunakan Convolutional Neural Network,” *JINACS*, vol. 02, pp. 53–59, 2020.

[11] Y. Goldberg, *Neural network methods for natural language processing*, vol. 10, no. 1. Morgan & Claypool Publishers, 2017.

[12] Y. Kim, “Convolutional neural networks for sentence classification,” *EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 1746–1751, 2014, doi: 10.3115/v1/d14-1181.

[13] H. Rhomadhona and J. Permadi, “Klasifikasi Berita Kriminal Menggunakan NaÃ¯ve Bayes Classifier (NBC) dengan Pengujian K-Fold Cross Validation,” *J. Sains dan Inform.*, vol. 5, no. 2, pp. 108–117, 2019, doi: 10.34128/jsi.v5i2.177.

[14] P.-N. Tan, M. Steinbach, and V. Kumar, *Introducing to Data Mining*. Boston: Pearson Education, 2006.