Jurnal Inovasi Pembelajaran Progresif
https://ejurnals.com/ojs/index.php/jipp Vol. 7, No. 3, September 2026

ANALISIS KINERJA K-NEAREST NEIGHBORS (KNN) DALAM KLASIFIKASI DESKRIPSI CUACA BERBASIS METODE N-GRAM

Nitamasi Finowa'a[1], Zakarias Situmorang[2]
[1,2]Universitas Katolik Santo Thomas
Email: nitamasi057@gmail.com[1], zakarias65@ust.ac.id[2]

Abstrak: Penelitian ini berfokus pada analisis kinerja algoritma K-Nearest Neighbors (KNN) dalam mengklasifikasikan deskripsi cuaca dari Badan Meteorologi, Klimatologi dan Geofisika (BMKG) dengan menggunakan metode ekstraksi fitur N-Gram. Berbeda dengan metode konvensional di mana data anomali biasanya dibuang, penelitian ini mengoptimalkan proses pra-pemrosesan sehingga tidak ada data observasi historis yang terbuang, dengan 100% dari 731 data curah hujan berhasil dipertahankan dengan menormalisasi data yang tidak terukur ke dalam kelas komputasi yang valid. Data tekstual kemudian diproses melalui langkah - langkah pra-pemrosesan teks, yang meliputi pembersihan, pengubahan huruf besar/kecil, tokenisasi, normalisasi, dan penghapusan kata - kata penghenti (stopword). Penghapusan dan stemming. Transformasi fitur teks dilakukan dengan variasi N-Gram (Unigram, Bigram, Trigram, dan Kombinasi) dan direpresentasikan dalam bentuk numerik menggunakan metode Binary Bag-of -Words. Proses klasifikasi dengan jarak Euclidean untuk variasi parameter K =1, 3, 5, dan 7 dengan pembobotan jarak. Evaluasi dengan matriks kebingungan pada 20% data pengujian menunjukkan bahwa konfigurasi KNN dengan fitur Unigram pada nilai K=3 adalah model yang paling optimal dan efisien. Model ini memberikan Akurasi 99,19%, Presisi Makro 0,9948, Recall Makro 0,9688, dan Skor F1 Makro 0,9807. Hasil menunjukkan bahwa metode Unigram jauh lebih efisien untuk mengurangi dimensi ruang fitur komputasi daripada metode Kombinasi N -Gram tanpa mengurangi tingkat akurasi klasifikasi cuaca.
Kata Kunci: K-Nearest Neighbors, N-Gram, Klasifikasi Teks, Deskripsi Cuaca, Text Mining.

Abstract: *This study focuses on analyzing the performance of the K-Nearest Neighbors (KNN) algorithm in classifying weather descriptions from the Meteorology, Climatology, and Geophysics Agency (BMKG) using the N-Gram feature extraction method. Unlike conventional methods where anomalous data is typically discarded, this study optimizes the preprocessing stage to ensure no historical observation data is lost; 100% of the 731 rainfall data entries were retained by normalizing unmeasured data into valid computational classes. The textual data underwent preprocessing steps, including cleaning, case conversion, tokenization, normalization, and stopword removal. Text feature transformation was performed using various N-Gram configurations (Unigram, Bigram, Trigram, and Combinations) and represented numerically via the Binary Bag-of-Words method. Classification was conducted using Euclidean distance with K-parameter variations of 1, 3, 5, and 7, incorporating distance weighting. Evaluation using a confusion matrix on 20% of the test data revealed that the KNN*

configuration using Unigram features with K=3 was the most optimal and efficient model. This model achieved an accuracy of 99.19%, a Macro Precision of 0.9948, a Macro Recall of 0.9688, and a Macro F1-Score of 0.9807. The results demonstrate that the Unigram method is significantly more efficient at reducing the computational feature space dimensionality compared to N-Gram Combination methods, without compromising weather classification accuracy.
Keywords: *K-Nearest Neighbors, N-Gram, Text Classification, Weather Description, Text Mining.*

PENDAHULUAN

Cuaca merupakan salah satu faktor lingkungan yang mempengaruhi beberapa aspek kehidupan manusia, seperti perencanaan kegiatan sehari-hari, efisiensi transportasi, produktivitas pertanian, hingga persiapan menghadapi risiko bencana [1]. Oleh karena itu, ketersediaan informasi cuaca yang tepat dan mudah dipahami merupakan kebutuhan dasar. Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) Indonesia merupakan otoritas resmi yang menyediakan data meteorologi, termasuk pengamatan curah hujan harian [2]. Data curah hujan seringkali didistribusikan secara numerik, dalam satuan milimeter (mm). Selain nilai numerik, kondisi cuaca juga direpresentasikan dalam bentuk deskriptif tekstual untuk memberikan penjelasan yang lebih komprehensif bagi masyarakat [3].

Namun, bidang ilmu komputasi menghadapi tantangan analitis dalam klasifikasi otomatis laporan tekstual ini ke dalam kategori intensitas curah hujan. Teks laporan cuaca tidak terstruktur dan tidak dapat dioperasikan secara langsung oleh algoritma klasifikasi matematika. Teks tersebut membutuhkan serangkaian Metode penambangan teks, khususnya pra-pemrosesan teks dan ekstraksi fitur, untuk menerjemahkan frasa linguistik ke dalam representasi numerik yang dapat dievaluasi[4].

Algoritma K - Nearest Neighbors (KNN) adalah salah satu metode pembelajaran terawasi yang dikenal dapat diandalkan dalam menyelesaikan masalah klasifikasi berdasarkan tingkat kesamaan atau kedekatan antara vektor fitur. Untuk tujuan klasifikasi teks, kombinasi metode representasi N - Gram dengan algoritma KNN dimungkinkan. Tinjauan literatur sebelumnya telah mengkonfirmasi bahwa penambangan teks efektif untuk ekstraksi informasi [5], Tahap pra-pemrosesan yang tepat dapat secara substansial mengurangi noise tekstual dan dengan demikian meningkatkan akurasi algoritma pembelajaran mesin [6]. Metode N-Gram dapat menangkap pola urutan kata dari satu kata (unigram), hingga urutan kata berurutan (bigram

dan trigram) sehingga konteks semantik suatu dokumen dapat ditangkap dengan lebih baik daripada penghitungan frekuensi kata tunggal [7]. Kombinasi KNN dan N-Gram juga telah menunjukkan kekokohan dalam klasifikasi teks multikelas lainnya, seperti deteksi ujaran kebencian[8] hingga klasifikasi dokumen hukum [9].

Studi ini bertujuan untuk menganalisis kinerja algoritma KNN dalam mengklasifikasikan teks deskripsi cuaca ke dalam label intensitas hujan dengan variasi metode ekstraksi N-Gram. Secara khusus, studi ini menyelidiki dampak variasi jumlah tetangga terdekat (K) pada tingkat akurasi dan efisiensi ruang fitur yang dihitung, sambil menjaga seluruh data riwayat cuaca tetap utuh secara ideal.

METODE PENELITIAN

Penelitian ini mengadopsi pendekatan eksperimental dengan pendekatan kuantitatif komputasional. Pendekatan ini didasarkan pada transformasi data tekstual menjadi vektor matematika untuk mengevaluasi kelayakannya melalui penggunaan metrik evaluasi klasifikasi [10]. Tahapan penelitian meliputi akuisisi dataset, pembersihan data, pemrosesan teks awal, pembangkitan fitur N -Gram, vektorisasi representasi, serta tahap klasifikasi dan evaluasi.

A. Pengumpulan dan Pembersihan Data

Objek penelitian adalah data arsip observasi curah hujan BMKG Wilayah I Medan. Dataset terdiri dari 731 catatan data harian. Sistem dirancang untuk menjaga integritas informasi pada fase pembersihan data. Data yang mengandung kode anomali observasi 8888 (curah hujan tidak terukur) dan curah hujan jejak (rentang 0,1 mm sampai 0,4 mm) tidak dibuang. Data disimpan dan dipetakan ke dalam satu kelas intensitas yang divalidasi sistem, yaitu kelas “hujan sangat ringan”. Sehingga, 100 % dari 731 sampel data awal tersebutdinyatakan valid untuk diproses lebih lanjut ke tahap klasifikasi.

B. Text Preprocessing

Deskripsi cuaca diproses baris demi baris secara sistematis untuk mencapai struktur linguistik yang seragam.

1. *Cleaning*: Penghapusan angka, simbol, dan tanda baca yang berpotensi noise.
2. *Case Folding*: Konversi seluruh karakter alfabet menjadi huruf kecil (*lowercase*).
3. *Tokenisasi*: Memecah kalimat lengkap menjadi serangkaian token kata penyusun.

4. *Normalisasi*: Keseragaman leksikal istilah cuaca yang non-standar [11], misalnya istilah "gerimis" dinormalisasi menjadi "hujan ringan".
5. *Stopword Removal*: Menghapus konjungsi dan kata-kata umum secara selektif. Kata - kata penentu makna utama seperti " tidak " dan " sangat" dipertahankan agar esensi kondisi cuaca tidak bergeser.
6. *Stemming*: Reduksi kata berimbuhan ke morfem dasarnya.

C. Ekstraksi dan Vektorisasi Fitur N-Gram

Ekstraksi teks yang telah melewati tahap pra-pemrosesan ke dalam pola N-Gram untuk mengenali struktur susunan kata. Rancang eksperimen menggunakan 4 varian ekstraksi. Unigram (n=1), Bigram (n=2), Trigram (n=3), dan Kombinasi (gabungan ruang fitur Unigram menjadi Trigram). Fitur-fitur tekstual fiturini dikodekan ke dalam model matematika menggunakan metode Binary Bag-of-Words. dikodekan ke dalam model matematika menggunakan metode Binary Bag-of-Words. Matriks akan ditugaskan dengan biner biner nilai 1 jika fitur N-Gram tertentu terdeteksi pada deskripsi cuaca dan dengan nilai 0 jika fitur tersebut tidak ada. Atribut dengan nilai 1 jika fitur N-Gram tertentu terdeteksi pada deskripsi cuaca dan dengan nilai 0 jika fitur tersebut tidak ada.

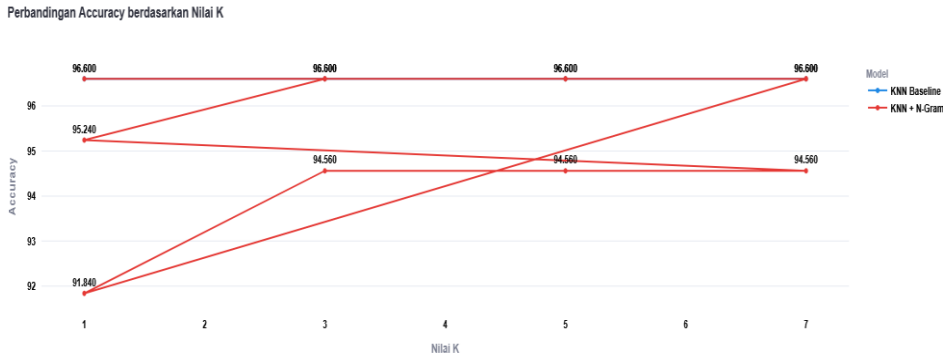
D. Klasifikasi K-Nearest Neighbors (KNN)

Seluruh matriks data biner di-split dengan metode hold-out, sehingga diperoleh perbandingan 80% data pelatihan (585 catatan) dan 20% data pengujian (146 catatan). Algoritma KNN bekerja dengan mengukur seberapa besar kuantitas proksimitas antara vektor data uji dengan data latih dengan menggunakan rumus Euclidean Distance. Setelah diperoleh jarak kalkulasi, jarak tersebut diurutkan secara menaik untuk mencari tetangga terdekat sejumlah nilai K yang diuji (K=1,3,5, dan 7). Teknik jarak pembobotan pembobotan diestimasi untuk klasifikasi akhir. Jarak vektor terdekat dari tetangga klasifikasi akan memiliki penentuan bobot kelas yang lebih unggul daripada tetangga yang lebih jauh.Teknik ini diestimasi untuk klasifikasi akhir. Jarak vektor terdekatKlasifikasi tetangga akan memiliki penentuan bobot kelas yang lebih unggul daripada tetangga yang jauh.

HASIL DAN PEMBAHASAN

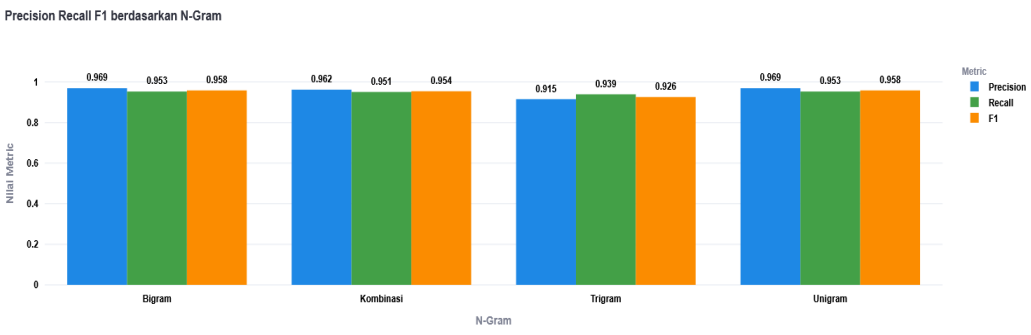
Pengujian komputasi dilakukan pada 146 data uji terhadap 585 data pelatihan. Eksperimen ini membandingkan dampak parameter kedekatan K dan dimensi ekstraksi N-

Gram pada metrik kualitas klasifikasi: Akurasi, Presisi, Recall, dan F1-Score. Berdasarkan pemantauan hasil pergerakan akurasi, kinerja algoritma pada K=1 cenderung menyebar, berfluktuasi, dan kurang stabil pada semua jenis N-Gram. Stabilitas yang rendah pada K=1 disebabkan karena sistem hanya bergantung pada kemiripan data sampel tunggal, yang menyebabkan probabilitas kesalahan klasifikasi yang sangat tinggi karena data yang bising. Dengan meningkatkan jumlah tetangga menjadi K=3, K=5, dan K=7, kinerja keseluruhan varian N-Gram yang dikoreksi meningkat dan membentuk garis konvergensi pada titik akurasi puncak.



Gambar 1. Grafik Perbandingan Akurasi Berdasarkan Nilai K

Gambar 1 menunjukkan tren pergerakan model akurasi terhadap variasi nilai K. Terlihat jelas bahwa seluruh lini performa dari model KNN baik yang menggunakan fitur N-Gram tunggal maupun Kombinasi secara bertahap menjadi stabil dan konvergen menyentuh titik akurasi tertinggi saat nilai K terbentuk dari angka 3. Kemudian, perbandingan rata-rata kinerja dari setiap variasi N-Gram disajikan secara lebih rinci pada Gambar 2.



Gambar 2. Grafik Perbandingan Kinerja Berdasarkan Jenis N-Gram

Gambar 2 menunjukkan trade - off antara kinerja akurasi dan ukuran overhead komputasi. Dari segi performa, metode Kombinasi N-Gram (kombinasi 1-3) mencatat persentase akurasi kumulatif tertinggi, yaitu 99,19%. Namun, untuk mencapai akurasi ini, metode Kombinasi membangun 777 ruang fitur komputasi. Jumlah fitur ini sangat besar sehingga tidak proporsional dengan teks observasi yang pendek dan akan menyebabkan penurunan efisiensi waktu pengiriman (curse of Dimensionality).

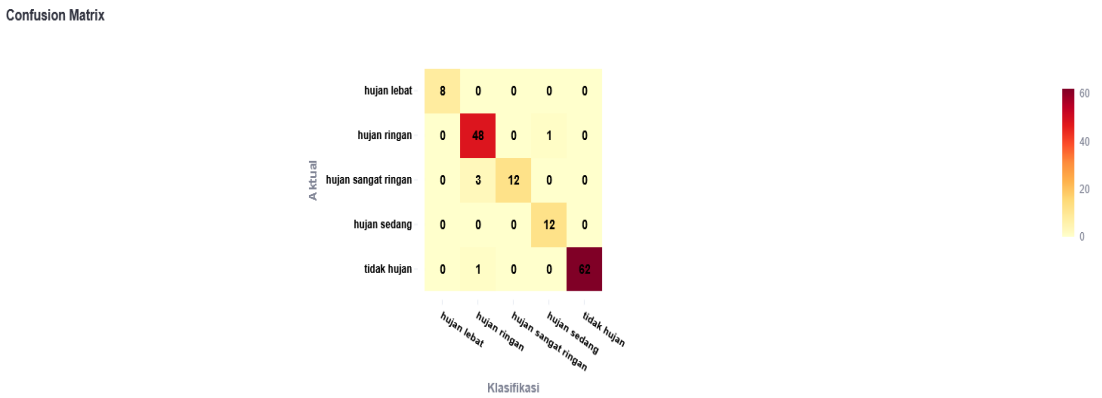
Model terbaik secara keseluruhan berhasil ditetapkan sebagai konfigurasi KNN dengan pendekatan Unigram pada nilai parameter K 3 sebagai titik sistem ideal keseluruhanModel tersebut berhasil diatur sebagai konfigurasi KNN dengan pendekatan Unigram pada nilai parameter K 3 sebagai titik sistem ideal.

Tabel 1. Performa Model KNN (Unigram, K=3)

Metrik Evaluasi	**Nilai Kinerja**
Accuracy	99,19%
Precision Macro	99,48%
Recall Macro	96,88%
F1-Score Macro	98,07%
Jumlah Fitur	111

Seperti yang ditunjukkan pada Tabel 1 menunjukkan, model KNN berdasarkanberbasis Unigram dengan 3 tetangga terdekat sangat memuaskan. Sistem ini hanya menjalankan 111 ruang komputasi fitur tunggal, tetapi dapat mencapai persentase akurasi tertinggi sebesar 99,19%, yang sama dengan kinerja maksimum metode Kombinasi. Model yang dihasilkan seimbang dalam kapabilitas yang tervalidasi dengan perolehan nilai F1-Score Macro sebesar 0,9807. Artinya sistem sudah dapat mengklasifikasikan kelas dengan tepat dan memiliki sensitivitas yang baik dalam mendeteksi persebaran data pada setiap kategori curah hujan.

Analisis lebih rinci dari matriks kebingungan (confusion matrix), yang digunakan sebagai dasar untuk memantau kinerja kelas aktual dan prediksi [12], menunjukkan dominasi yang sangat padat pada area diagonal utama klasifikasi.



Gambar 3. Confusion Matrix Hasil Klasifikasi Model

Gambar 3 menunjukkan bahwa sistem KNN berhasilmengelompokkan sebagian besar data pelatihan secara gugus ke dalam 7 kelas cuaca yang sebenarnya. Sebagian besar data pelatihan dimasukkan secara tepat ke dalam 7 kelas cuaca yang sebenarnya. Hasil normalisasi anomali kelas "hujan sangat ringan" terbukti dapat dipisahkan secara matematis dengan algoritma jarak. Sedikit sekali penyimpangan klasifikasi yang muncul pada frekuensi yang sangat minim pada kelas-kelas dengan spektrum yang berdekatan seperti kelas hujan sedang dan lebat, karena sering digunakan frase beririsan oleh BMKG untuk mendeskripsikan transisi intensitas hujan di lapangan

KESIMPULAN DAN SARAN

Eksperimen penggunaan algoritma K-Nearest Neighbors berdasarkan representasi N-Gram terbukti andal dalam klasifikasi otomatis dokumen pelaporan cuaca ke dalam kategori intensitas curah hujan. Optimalisasi terhadap perlakuan prapemrosesan berhasil meniadakan pembuangan data anomali historis observasi lapangan. Tercatat sebanyak 731 rekaman data secara utuh berhasil divalidasi ke dalam rentang kelas pengukuranonal. Simulasi ruang fitur yang berbeda dilakukan dan konfigurasi Ekstraksi Unigram dengan parameter K = 3 ditemukan sebagai pemodelan terbaik dan paling efisien. Model ini dapat mencapai akurasi maksimum 99,19% dan F1-Score 0,9807, dan dapat menghemat kapasitas ruang fitur secara drastis dibandingkan dengan metode Kombinasi N-Gram. Untuk meningkatkan generalisasi pemodelan di masa mendatang, disarankan untuk menggabungkan volume dataset iklim dari stasiun wilayah observasi yang berbeda untuk memperkuat distribusi sampel untuk kelas kategori curah hujan ekstrem.

DAFTAR PUSTAKA

H. Sulaiman, M. Ryansyah, K. Widianto, S. Sidik, and A. Nugraha, "Implementasi Machine Learning Dengan Metode Text Mining Pada Twitter," *Infotek: Jurnal Informatika dan Teknologi*, vol. 7, no. 1, pp. 52–62, Jan. 2024, doi: 10.29408/jit.v7i1.23734.

Badan Meteorologi Klimatologi dan Geofisika (BMKG), "Dataset Pengamatan Curah Hujan Harian Wilayah I Medan (Data Online Pusat Database BMKG)." Accessed: Jul. 26, 2026. [Online]. Available: https://dataonline.bmkg.go.id/

D. Alaina Lailatul Maqfiroh and Z. Fatah, "K-Nearest Neighbor untuk Memprediksi Kondisi Cuaca Menggunakan RapidMiner," *Jurnal Sains dan Sistem Teknologi Informasi (SANDI)*, vol. 6, no. 2, pp. 30–41, 2024.

D. Astari Umbu Zaza *et al.*, "Penerapan Text Mining Dalam Klasifikasi Judul Skripsi Mahasiswa Pada Universitas Stella Maris Sumba Menggunakan Metode Naïve Bayes," *Jurnal Informatika dan Bisnis*, vol. 02, no. 03, pp. 327–337, 2024.

A. Hermawan, I. Jowensen, J. Junaedi, and Edy, "Implementasi Text-Mining untuk Analisis Sentimen pada Twitter dengan Algoritma Support Vector Machine," *JST (Jurnal Sains dan Teknologi)*, vol. 12, no. 1, pp. 129–137, Apr. 2023, doi: 10.23887/jstundiksha.v12i1.52358.

C. Suhaeni, S. A. Kamila, F. Fahira, M. Yusran, and G. Alfa Dito, "Exploring a Large Language Model on the ChatGPT Platform for Indonesian Text Preprocessing Tasks," *Indonesian Journal of Statistics and Its Applications*, vol. 9, no. 1, pp. 100–116, Jun. 2025, doi: 10.29244/ijsa.v9i1p100-116.

Y. Setiawan, N. U. Maulidevi, and K. Surendro, "The Optimization of n-Gram Feature Extraction Based on Term Occurrence for Cyberbullying Classification," *Data Sci. J.*, vol. 23, no. 1, 2024, doi: 10.5334/dsj-2024-031.

K. Hadi and E. Utami, "Analysis of K-NN with the Integration of Bag of Words, TF-IDF, and N-Grams for Hate Speech Classification on Twitter," *Jurnal Informatika*, vol. 12, no. 2, pp. 289–298, 2024.

M. Alidin and R. Fadilah, "Optimasi KNN dengan PSO untuk Klasifikasi Kasus Hukum di Australia Menggunakan N-Gram," *Journal of Software Engineering and Information System (SEIS)*, vol. 5, no. 1, 2025, [Online]. Available: https://ejurnal.umri.ac.id/index.php/SEIS/index

A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, " Survey on Text Classification Algorithms: From Text to Predictions," *Information (Switzerland)*, vol. 13, no. 2, Feb. 2022, doi: 10.3390/info13020083.

M. I. Raif, N. N. Hidayati, and T. Matulatan, "Otomatisasi Pendeteksi Kata Baku Dan Tidak Baku Pada Data Twitter Berbasis KBBI," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 2, pp. 337–348, Apr. 2024, doi: 10.25126/jtiik.20241127404.

S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," *African Journal of Biomedical Research*, pp. 4023–4031, Nov. 2024, doi: 10.53555/ajbr.v27i4s.4345.