JOISIE Journal Of Information System And Informatics Engineering
Vol. 7, No.2, Desember 2023, Hlm 358-363
p- ISSN: 2503-5304

PENERAPAN ALGORITMA *K-NEAREST NEIGHBOUR (K-NN)* SEBAGAI KLASIFIKASI *TWEETS* PADA *TWITTER*

Cian Ramadhona Hassolthine[1*], Riad Sahara[2] , Fesa Asy Syifa Nurul Haq[3], Syahid Abdullah[4]
[1,2,4] PJJ Informatika, Universitas Siber Asia
[3] PJJ Sistem Informasi Universitas Siber Asia
Jl. RM Harsono No.1, Pasar Minggu, Jakarta Selatan, (021) 2780-6189
e-mail : [1]cianhassolthine@lecturer.unsia.ac.id, [2]riadsahara@lecturer.unsia.ac.id,
[3]fesasyifa@lecturer.unsia.ac.id, syahidabdullah@lecturer.unsia.ac.id

Abstract

As technology develops so rapidly in collecting data, it results in a huge pile of data. Due to the large amount of data, it becomes necessary to utilize this data. The aim of using data is of course to receive crucial news from the data patterns that are formed. Data that can be used can be obtained from social media, one of which is Twitter. Twitter is a social media with approximately 50 million users in Indonesia. With so many users in Indonesia, it can be used to use a lot of data. To get this data, use one of the K-Nearest Neighbor algorithms. The KNN algorithm is a classification of a set of data based on learning data that has been previously classified. The classification result of the KK Algorithm is that the data that has been processed falls into class B because of the three closest neighbors, two are in class B, while only one is in class A. The accuracy produced by the KNN Algorithm is also quite good, namely above 80% . This model provides better sensitivity in the data classification process.

Keywords : Algorrithm, Classification, K-Nearest Neighbor (K-NN), Text Mining, Preprocessing

Abstrak

Seiring berkembangnya teknologi yang begitu pesat dalam melakukan pengumpulan data mengakibatkan sebuah tumpukan data yg sangat banyak. Melalui banyaknya data tersebut,sehingga menjadi suatu kebutuhan untuk memanfaatkan data tersebut. Pemanfaatan data tentunya bertujuan agar menerima berita yg krusial dari pola-pola data yang terbentuk. Data yang bisa digunakan dapat diperoleh dari media sosial, salah satunya twitter. Twitter merupakan media sosial yang tercatat kurang lebih 50 juta orang pengguna di Indonesia. Dengan banyaknya pengguna di Indonesia, maka dapat dimanfaatkan dalam penggunaan data yang banyak. Untuk mendapatkan data tersebut yaitu dengan salah satu algoritma K-Nearest Neighbor. Algoritma KNN merupakan klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasikan sebelumya. Hasil klasifikasi dari Algoritma KK yaitu data yang sudah diolah masuk ke dalam kelas B karena dari tiga tetangga terdekat, ada dua yang masuk kelas B, sementara hanya satu yang masuk kelas A. Akurasi yang dihasilkan oleh Algoritma KNN juga cukup baik yaitu di atas 80%. Model ini memberikan sensitivity yang lebih baik dalam proses klasifikasi data.

Keywords : *Algoritma, K-Nearest Neighbor (K-NN), Klasifikasi, Text Mining, Preprocessing*

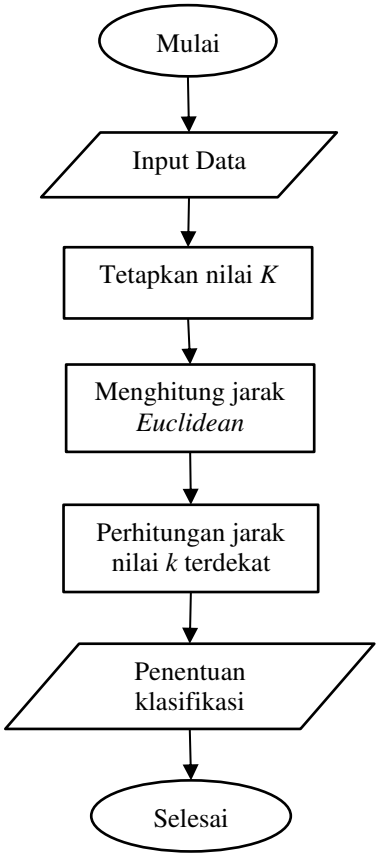
1. PENDAHULUAN

Twitter merupakan layanan *microblogging* yang memungkinkan pengguna untuk berkomunikasi melalui pesan singkat yang dikenal sebagai "*tweet*," adalah salah satu jejaring sosial paling populer saat ini (Muhammad Naja Maskuri, Harliana, Kadek Sukerti, 2022) *Tweet* dapat dilihat oleh siapa saja yang mengunjungi profil pengguna, tetapi mereka juga dapat disembunyikan dari semua orang kecuali

pengikut pengguna (Yusuf et al., 2021). Dengan basis pengguna yang terus berkembang, Twitter telah menjadi alat yang ampuh untuk menyebarkan berita dan konten lainnya dengan cepat. Informasi, termasuk berita, pemikiran, pertanyaan, pernyataan, dan kritik (baik yang menguntungkan maupun yang buruk), dikumpulkan dan disebarluaskan secara bebas melalui media (Wie & Siddik, 2022). Analisis sentimen memungkinkan kita untuk melihat bagaimana perasaan orang lain tentang suatu topik, yang dapat mempercepat dan meningkatkan akurasi pengambilan keputusan kita sendiri (Amrizal, 2018). Kurangnya klasifikasi *tweet* adalah penghalang utama bagi pengguna yang ingin membaca buletin atau artikel di halaman utama Twitter.

Dalam penambangan teks, klasifikasi adalah aturan yang membantu mengumpulkan data berdasarkan seberapa mirip skrip tersebut (Hardiyanto & Rozi, n.d.). Dengan menggunakan kriteria ini, Twitter dapat mengelompokkan *tweet* serupa bersama-sama, seperti tentang olahraga yang berbeda (seperti sepak bola, bola voli, dan tenis) ke dalam satu kategori. Karena buletin dan liputan Twitter diatur menurut kategori, kebiasaan mengkategorikan *tweet* dapat membuat mendengarkan buletin atau laporan jauh lebih sederhana (Suharno et al., 2017).

K-Nearest Neighbor digunakan sebagai kriteria kategorisasi (KNN) (Ghani Muttaqin et al., 2020). Salah satu prinsip pembelajaran mesin, KNN menggunakan data pelatihan yang paling mirip dengan item yang diklasifikasikan untuk menentukan kategorisasinya (Hadi & Sukamto, 2020). Ada dua elemen untuk alasan di balik pengamatan ini: tujuan yang luas dan yang lebih sempit. Tujuan utamanya adalah untuk mengoptimalkan kriteria *K-Nearest Neighbor* untuk mengkategorikan pesan di Twitter. Tujuan penelitian ini adalah teknik KNN untuk mengkategorikan *tweet* di Twitter. Penulis tertarik untuk mempelajari kategorisasi *tweet* di Twitter menggunakan pendekatan K-Nearest Neighbor karena studi yang diselesaikannya berdasarkan melihat tantangan yang dilakukan oleh sarjana lain dalam penelitian klasifikasi twitter. Bahasa pemrograman R akan digunakan untuk membangun sistem ini.



Gambar 1. Tahapan Penelitian

2. METODE PENELITIAN

Pada penelitian ini melalui beberapa tahapan yaitu Input data, tetapkan nilai K, Menghitung jarak *Euclidean*, perhitungan jarak nilai k terdekat dan penentuan klasifikasi. Tahapan penelitian dapat dilihat pada Gambar berikut :

2.1 KLASIFIKASI

Klasifikasi teks adalah proses penambangan teks dengan tujuan menetapkan teks ke kategori yang paling sesuai dengan fiturnya sesuai dengan kriteria yang telah ditentukan (Triyanto et al., 2021). Dengan klasifikasi teks di tangan, itu melukiskan gambaran mental dari proses klasifikasi arsip, yang merupakan langkah penting dalam membuat kontribusi dunia nyata (Hadi & Sukamto, 2020). *Naive Bayes, Artificial Neural Networks, Support Vector Machines, Genetic Techniques, K-Nearest*

Neighbors, dan *Fuzzy C-means* adalah contoh algoritma klasifikasi populer.

2.2 TEXT MINING

Text Mining adalah metode berbantuan komputer untuk mengekstraksi informasi dari dokumen; Dokumen-dokumen itu sendiri biasanya terdiri dari data tekstual (Ma'rifah et al., 2020). Data teks membutuhkan tahapan praprocessing untuk mengidentifikasi dan merepresentasikan data tidak terstruktur menjadi data terstruktur sehingga mudah untuk dipahami.

Menurut (Salahudin et al., 2020) *text mining* memiliki beberapa tahapan praproses yaitu sebagai berikut :

1. *Case folding*: mengubah semua karakter huruf menjadi huruf non-kapital *(lower casing).*
2. *Remove punctuation*: menghapus semua karakter tanda baca seperti tanda titik (.), titik dua (:), koma (,), tanda tanya (?), tanda seru (!), tanda petik (“), dan sebagainya.
3. *Remove hashtag*: simbol *hashtag* (#) dalam *tweet* digunakan untuk judul topik pembicaraan dan pengelompokan percakapan.
4. *Clean number*: menghapus angka yang terdapat pada teks serta angka yang terdapat di depan dan di belakang kata. Penyertaan angka di depan atau di belakang kata menunjukkan kata tersebut diulang seperti lari2 maksudnya larilari, namun hal tersebut tidak dibenarkan dalam Kamus Besar Bahasa Indonesia sehingga perlu dihapus.
5. *Remove* URL: URL atau alamat web perlu dihapus karena dijadikan halaman promosi bagi sebagian pengguna. URL yang muncul pada *tweet* tidak memiliki arti.
6. *Remove stopword*: menghapus kata yang tidak mencirikan isi dari suatu dokumen (teks) seperti kata “di”, “pada”, “oleh”, dan sebagainya. Proses *remove stopword* dilakukan setelah dibuat daftar *stopword (stoplist).*

2.3 ALGORITMA K-NEAREST NEIGHBOR (KNN)

Salah satu cara termudah untuk menyelesaikan masalah klasifikasi adalah menggunakan algoritma *K-Nearest Neighbor* (KNN) (Setiyorini et al., n.d.). Klasifikasi teks dan data adalah aplikasi umum dari pendekatan ini (Handayani et al., 2022). Mengklasifikasikan hal-hal menggunakan informasi yang secara geografis dekat digunakan dalam pendekatan ini (Qaiser et al., 2018). Untuk menentukan kesenjangan antara dua posisi x1 dan x2, kami menggunakan rumus:

$$dis\left(x_{1,}x_2\right) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2} \quad (1)$$

Ket :

Jarak dalam ruang 1 dimensi ditentukan oleh variabel independen tunggal, dalam ruang 2 dimensi oleh dua, dan dalam ruang multi-dimensi oleh lebih dari dua (Rizkya Rani et al., 2019).

3. HASIL DAN PEMBAHASAN

Berikut terdapat data pelanggan yang tercantum dalam tabel di bawah ini:

Tabel 1. Data Pelanggan

Age	Income	Class
29	350	A
51	430	B
33	290	A
24	255	A
40	410	B
45	380	B
34	390	?

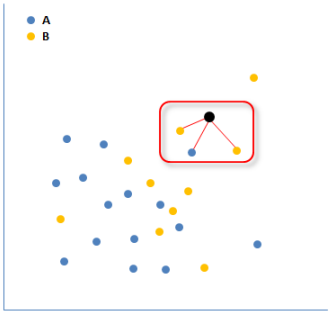
Ada enam data yang telah diberi label dan satu data yang perlu diidentifikasi oleh kelas berdasarkan informasi yang diberikan pada tabel di atas.

Ada 2 kelas yaitu kelas A dan kelas B

1. Nilai usia dan pendapatan, yang keduanya merupakan variabel independen, akan digunakan dalam

perhitungan jarak karena tidak terpengaruh oleh faktor lain.

2. Sebagai variabel dependen, nilai kelas tergantung pada ada atau tidaknya faktor-faktor lain (Usia dan Pendapatan).



Gambar 2. Grafik *Data Point*

Dari gambar di atas dapat dianalisa, ada sejumlah data point yang terbagi menjadi dua kelas yaitu A (biru) dan B (kuning). Misalnya ada data baru (hitam) yang akan kita prediksi kelasnya menggunakan algoritma KNN. Dari contoh di atas, nilai K yang digunakan adalah 3. Setelah dihitung jarak antara titik hitam ke masing-masing data point lainnya, didapatkan 3 titik terdekat yang terdiri dari 2 titik kuning dan satu titik biru seperti yang diilustrasikan di dalam kotak merah, maka kelas untuk data baru (titik hitam) adalah B (kuning).

3.1 ANALISIS ALGORITMA KLASIFIKASI TENTUKAN NILAI K

Jika kelasnya adalah bilangan genap, nilai K harus berupa bilangan ganjil, dan sebaliknya. Karena ada tepat dua kelompok, A dan B, setiap nilai K sama-sama mungkin menghasilkan hasil yang sama (karena pasangan tetangga terdekat selalu memiliki tepat dua anggota dari kelompok yang sama). Namun, jika Anda memilih nilai ganjil K, seperti 3 atau 5, hasilnya akan selalu menjadi total populasi yang lebih besar.

Dalam pemrograman R, misalnya, nilai K optimal dapat ditentukan dengan menjalankan serangkaian perhitungan menggunakan rentang nilai K yang mungkin (katakanlah, dari K = 2 hingga K = 10).

3.2 Hitung jarak antara data baru dan masing-masing data lainnya

Gunakan rumus yang diberikan untuk mendapatkan jarak menggunakan pendekatan *Euclidean*. Enam informasi perlu dihitung:

Data 1

$$dis = \sqrt{(34-29)^2 + (390-350)^2} = 40.31$$

Data 2

$$dis = \sqrt{(34-51)^2 + (390-430)^2} = 43.46$$

Data 3

$$dis = \sqrt{(34-33)^2 + (390-290)^2} = 100.01$$

Data 4

$$dis = \sqrt{(34-29)^2 + (390-255)^2} = 135.37$$

Data 5

$$dis = \sqrt{(34-40)^2 + (390-410)^2} = 20.88$$

Data 6

$$dis = \sqrt{(34-45)^2 + (390-380)^2} = 14.87$$

3.2 AMBIL TIGA DATA DENGAN JARAK TERDEKAT

Jika kita memilih jarak terdekat dari perhitungan jarak Euclidean sebelumnya, kita mendapatkan yang berikut.

Tabel 2. Perhitungan *Euclidean distance*

Data	Age	Income	Jarak dengan data baru
6	45	380	14.87
5	40	410	20.88
1	29	350	40.31
2	51	430	43.46
3	33	290	100.01
4	24	255	135.37

Analisa dari Tabel 2 di atas yaitu data baru tersebut masuk ke dalam kelas B karena dari tiga tetangga terdekat, ada dua yang

masuk kelas B, sementara hanya satu yang masuk kelas A.

3.3 AKURASI

Setelah dilakukan tahap klasifikasi menggunakan Algoritma KNN, selanjutnya dilakukan pengujian dengan teknik *Confision Matrix.* Hasil akurasi dapat dilihat pada Gambar 3 di bawah ini.

```
          truth
response    1    0
       1  370  390
       0 1544 7696
```

```
accknn <- train_test_telp_knn$aggregate(list(msr("classif.acc"),
                                msr("classif.specificity"),
                                msr("classif.sensitivity")))
accknn
```

```
      classif.acc classif.specificity classif.sensitivity
        0.8066000          0.9517685          0.1933124
```

Gambar 3. Model Akurasi Confision Matrix

Akurasi yang dihasilkan oleh Algoritma KNN juga cukup baik yaitu di atas 80%. Model ini memberikan *sensitivity* yang lebih baik dalam proses klasifikasi data.

4. SIMPULAN

Berdasarkan data sebelumnya, kita dapat menyimpulkan bahwa dari tiga tetangga terdekat, dua jatuh ke kelas B sementara yang tersisa tetap di kelas A, menempatkan data baru di kelas B. Algoritma KNN menghitung KNN dengan menemukan sampel pelatihan yang paling dekat dengan sampel uji. Setelah KNN dikumpulkan, sebagian besar informasi digunakan untuk membuat prediksi berdasarkan sampel uji.

6. DAFTAR PUSTAKA

Amrizal, V. (2018). Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim). *Jurnal Teknik InformatikA*, *11*(2), 149–164. https://doi.org/10.15408/jti.v11i2.8623

Ghani Muttaqin, A., Auliasari, K., & Santi Wahyuni, F. (2020). Penerapan Metode K-Nearest Neighbor Untuk Prediksi Penjualan Berbasis Web Pada Pt.Wika Industry Energy. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 4, Issue 2).

Hadi, H. P., & Sukamto, T. S. (2020). Klasifikasi Jenis Laporan Masyarakat Dengan K-Nearest Neighbor Algorithm. *JOINS (Journal of Information System)*, *5*(1), 77–85. https://doi.org/10.33633/joins.v5i1.3355

Handayani, R. D., Fauzi Anggi, A., Wahyu, Y., & Purtra, S. (2022). Klasifikasi Emosi Pada Sosial Media Menggunakan Support Vector Mahine dan N-Gram. *Jurnal Riset Teknologi Informasi Dan Komputer (JURISTIK)*, *2*(2), 7–10. https://doi.org/10.53863/juristik.v2i2.590

Hardiyanto, B., & Rozi, F. (n.d.). *Prediksi Penjualan Sepatu Menggunakan Metode K-Nearest Neighbor*.

Ma'rifah, H., Prasetya Wibawa, A., & Akbar, M. I. (2020). *Sains, Aplikasi, Komputasi dan Teknologi Informasi Klasifikasi artikel ilmiah dengan berbagai skenario preprocessing*. *2*(2), 70.

Muhammad Naja Maskuri, Harliana, Kadek Sukerti, R. M. H. B. (2022). Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Penyakit Stroke. *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, *4*(1), 130–140.

Qaiser, S., Utara, U., Sintok, M., Kedah, M., Ramsha, A., & Analytics, T. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents Text Mining. In *International Journal of Computer Applications* (Vol. 181, Issue 1).

Rizkya Rani, S., Retno Andani, S., Suhendro, D., Studi Sistem Informasi, P., & Tunas Bangsa Pematangsiantar Jln Jendral Sudirman Blok No, S. A. (2019). *Nearest Neighbor untuk Prediksi Kelulusan Siswa pada SMK (Sekar Rizkya Rani) | 670*.

Salahudin, S., Sulistyaningsih, T., & Lutfi, M. (2020). *Analysis of Government*

Official Twitters during Covid-19 Crisis in Indonesia. https://www.researchgate.net/publication/342145462

Setiyorini, T., Rizky, ;, & Asmono, T. (n.d.). *Penerapan Metode K-Nearest Neighbor Dan Information Gain Pada Klasifikasi Kinerja Siswa*. http://nusamandiri.ac.id

Suharno, C. F., Fauzi, M. A., & Perdana, R. S. (2017). *Klasifikasi Teks Bahasa Indonesia pada Dokumen Pengaduan Sambat Online menggunakan Metode K-Nearest Neighbors (K-NN) dan Chi-Square* (Vol. 1, Issue 10). http://j-ptiik.ub.ac.id

Triyanto, S., Sunyoto, A., & Arief, M. R. (2021). Analisis Klasifikasi Bencana Banjir Berdasarkan Curah Hujan Menggunakan Algoritma Naïve Bayes. *JOISIE Journal Of Information System And Informatics Engineering*, *5*(Desember), 109–117. https://scikit-learn.org/,

Wie, J. V., & Siddik, M. (2022). Penerapan Metode Naïve Bayes Dalam Mengklasifikasi Tingkat Obesitas Pada Pria. *JOISIE Journal Of Information System And Informatics Engineering*, *6*(Desember), 69–77. https://www.kaggle.com/,

Yusuf, M., Rangkuti, R., Alfansyuri, V., Gunawan, W., Informatika, T., Komputer, I., & Mercu Buana, U. (2021). *Penerapan Algoritma K-Nearest Neighbor (Knn) Dalam Memprediksi Dan Menghitung Tingkat Akurasi Data Cuaca Di Indonesia*. *2*(2).