

Klasifikasi Opini Dengan Menggunakan Algoritma K-Nearest Neighbor Pada Berita Vaksinasi Di Twitter

Dede Sandi¹, Ema Utami², Agus Fatkhurohman³

¹Universitas Amikom Yogyakarta

^{2,3}Pascasarjana Magister Teknik Informatika Universitas Amikom Yogyakarta

E-mail: ¹dedesandi1994@gmail.com, ²ema.u@amikom.ac.id, ³agusfatkhurohman@amikom.ac.id

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 dipergunakan 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 pula. Untuk mendapatkan data tersebut yaitu dengan salah satu algoritma K-Nearest Neighbor Classifier atau KNN. Sistem kerja KNN ialah dengan menghitung jarak terdekat asal record uji ke record testing menggunakan memakai metode skenario pengujian. Hasil dari proses KNN tersebut berupa jarak terdekat dari record uji ke record testing sebesar K sesuai dengan yang diperlukan.

Kata Kunci— twitter, data mining, klasifikasi, k-nearest neighbor classifier, euclidean distance.

Abstract

Rapid technology development leads to heap of massive data. The big amount of the data should be utilized properly. The purpose of data utilization is to help users to get crucial information from the formed data patterns. The amount of crucial data can be obtained from social media i.e., twitter. Twitter is a social media that has approximately 50 million users in Indonesia. By having so many users in Indonesia, the twitter data can be utilized for many purposes. This research is interested to formed data from the twitter by using one of the algorithms which is K-Nearest Neighbor Classifier or KNN. The KNN work system is to calculate the closest distance from the test record to the test record using the test scenario method. The result of the KNN process is the shortest distance from the test record to the test record of K as needed.

Keywords— twitter, data mining, classification, k-nearest neighbor classifier, euclidean distance.

1. PENDAHULUAN

1.1 Latar Belakang

Dengan perkembangan media sosial di Indonesia cukup populer, dan salah satunya adalah *Twitter*. *Twitter* merupakan media sosial yang cukup digemari banyak kalangan, tercatat sekitar 50 juta orang di Indonesia menggunakan *Twitter*. Pada tahun

2016 telah tercatat 4,1 milyar *tweet* dari para pengguna di Indonesia.

Melalui data yang sangat banyak, sehingga dapat digunakan data tersebut untuk informasi yang diperlukan. Salah satunya ialah informasi mengenai opini masyarakat atau pengguna *twitter* tentunya di Indonesia. Opini merupakan ungkapan

pendapat dalam perasaan suka atau tidak suka seseorang terhadap suatu objek. Dalam hal ini saya sebagai penulis akan memfokuskan hanya kepada opini atau pendapat mengenai Vaksinasi di Indonesia.

Berita atau unggahan kabar mengenai Vaksinasi di Indonesia tentunya memuat banyak opini, entah itu dari pujian hingga kritikan dari pengguna *twitter* di Indonesia. Dengan dasar itu, penulis melakukan pengklasifikasian terhadap *tweet* tersebut, untuk klasifikasi pujian dan kritikan. Berita mengenai Vaksinasi begitu sangat krusial sehingga penulis membuat penelitian ini bertujuan untuk membuka suatu

sudut pandang terhadap tanggapan variatif dari masyarakat di Indonesia yang pastinya dapat memengaruhi mengenai tindakan pemerintah dalam menangani vaksinasi untuk selanjutnya.

2. METODE PENELITIAN

2.1. Algoritma *K-Nearest Neighbor*

Algoritma klasifikasi yang digunakan ialah *K-Nearest Neighbor* atau KNN yang melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. KNN melakukan penyimpanan vektor-vektor fitur yang sama dihitung untuk *data testing* (yang kelas tidak diketahui). Dari vektor yang baru ini pembelajaran dihitung, dan sejumlah k yang paling dekat diambil.

Maka itu, dalam penelitian ini digunakan metode KNN untuk mengklasifikasikan atau mengolah *tweet* untuk diketahui klasifikasi dari *tweet* tersebut (positif atau negatif) dalam berita Vaksinasi di Indonesia.

3. HASIL DAN PEMBAHASAN

3.1. *Dataset Twitter*

Dataset yang digunakan merupakan *tweet* sebanyak 1000 *tweet* yang didapatkan melalui situs penyedia *dataset* kaggle.com, diperlukan untuk melakukan penelitian klasifikasi menggunakan algoritma KNN. Di bawah ini adalah sample dari dataset yang digunakan:

Tabel 3.1. Sampel *Dataset Twitter*

Pengguna	<i>Tweet</i>
Randy Aldino	Masker bekas rawan tertular virus. Pakai masker baru utk mencegah secara maksimal penularan virus co...
Pikiran Rakyat	Berharap Tak Muncul Lagi Gelombang Pandemi, Tiongkok Gratiskan Vaksin Covid-19 untuk Warganya #Pande...
Geisha Ayu	Ayo Dukung!! Hidup sehat tanpa Covid-19 dengan melakukan vaksinasi covid-19 #vaksinasi https://t.co/...
Marsianus Eka Noviantus	kenapa harus nakes, kan kita ada wakil rakyat, sebagai wakil nya rakyat, seharusnya mereka yang pertama #covid19
yugo notobandino	udahlah, ayok berubah dan bergerak umur kita takkan panjang hidup kita sudah ditentukan mau sampai kapan kayak gini
BaBe - Baca Berita	China Manfaatkan Sertifikat Halal MUI untuk Memasarkan Vaksin Sinovac ke Negara-Negara Islam

3.2. *Preprocessing Dataset*

Preprocessing merupakan proses untuk mengolah data yang belum sesuai dengan bentuk data yang diperlukan untuk proses klasifikasi dengan *case folding* untuk mengubah semua *upper case* menjadi *lower case* dan *stopwords removal* untuk menghapus kata, *space* dan karakter yang tidak dipakai. Berikut ini adalah data yang telah di *preprocessing*. Berikut adalah dataset yang telah melalui proses *preprocessing dataset*:

Tabel 3.2. *Preprocessing* Sampel Dataset
Twitter

<i>Preprocessing</i>
randy aldino masker bekas rawan tertular virus pakai masker baru utk mencegah secara maksimal penularan virus
pikiran rakyat berharap tak muncul lagi gelombang pandemi tiongkok gratisan vaksin untukarganya
geisha ayu ayo dukung hidup sehat tanpa covid-19 dengan melakukan vaksinasi covid-19
narsianus eka noviantus kenapa harus nakes kan kita ada wakil rakyat sebagai wakil nya rakyat seharusnya mereka yang pertama
yugo notobandino udahlah, ayok berubah dan bergerak umur kita takkan panjang hidup kita sudah ditentukan mau sampai kapan kayak gini
babe baca berita china memanfaatkan sertifikat halal mui untuk Memasarkan vaksin sinovac ke negara islam

3.3. *Split Dataset*

Data *split* bertujuan untuk mendapatkan data *training* dan data *testing*. Dalam penelitian ini penulis menggunakan 3 skenario yaitu 80%:20%, 70%:30% dan 60%:40%.

3.4. Pengujian dan Analisis Skenario

Pengujian yang dilakukan penulis ialah terhadap data *split* dan jumlah $k = 3$, $k = 5$, $k = 7$, $k = 9$ dengan tiap pengujian jumlah digunakan skenario data *split*.

Tabel di bawah ini adalah hasil pengujian metode KNN dengan $k = 3$ dan 3 skenario *split* yaitu 80% data *training* dan 20% data *testing*, 70% data *training* dan 30% data *testing*, 60% data *training* dan 40% data *testing*.

Tabel 3.3. Perbandingan *Accuracy Data*

$k = 3$	
Data <i>Split</i>	<i>Accuracy</i>
80-20	90.00%
70-30	73.00%
60-40	62.00%

Tabel di bawah ini adalah hasil pengujian metode KNN dengan $k = 5$ dan 3 skenario *split* yaitu 80% data *training* dan 20% data *testing*, 70% data *training* dan 30% data *testing*, 60% data *training* dan 40% data *testing*.

Tabel 3.4. Perbandingan *Accuracy Data*

$k = 5$	
Data <i>Split</i>	<i>Accuracy</i>
80-20	90.00%
70-30	73.00%
60-40	62.00%

Tabel di bawah ini adalah hasil pengujian metode KNN dengan $k = 7$ dan 3 skenario *split* yaitu 80% data *training* dan 20% data *testing*, 70% data *training* dan 30% data *testing*, 60% data *training* dan 40% data *testing*.

Tabel 3.5. Perbandingan *Accuracy Data*

$k = 7$	
Data <i>Split</i>	<i>Accuracy</i>
80-20	86.00%
70-30	70.67%
60-40	61.50%

Tabel di bawah ini adalah hasil pengujian metode KNN dengan $k = 9$ dan 3 skenario *split* yaitu 80% data *training* dan 20% data *testing*, 70% data *training* dan 30% data *testing*, 60% data *training* dan 40% data *testing*.

Tabel 3.6. Perbandingan *Accuracy Data*

$k = 9$	
Data Split	Accuracy
80-20	85.50%
70-30	70.33%
60-40	60.50%

Akurasi terbaik dari tabel di atas adalah skenario dengan data *split* 80% data *training* dan 20% data *testing* dan jumlah $k = 3$, $k = 5$. Dengan demikian dapat dianalisa jika semakin banyak data *training* yang digunakan maka hasil akurasi yang didapatkan akan semakin tinggi, namun untuk penentuan k tidak menentukan akurasi karena belum mencoba untuk k lainnya.

4. KESIMPULAN

Pada pengujian skenario dalam proses klasifikasi semakin banyak data *training* yang digunakan maka akurasi yang didapat akan semakin bagus karena banyak proses data *training*. Nilai k yang besar akan memperbesar jumlah kebenaran pada proses klasifikasi, dan setiap skenario pengguna *twitter* lebih banyak berpendapat bahwa berita vaksinasi di Indonesia untuk saat ini masih tergolong bagus dengan menggunakan algoritma *K-Nearest Neighbor* atau KNN.

5. SARAN

Pada penelitian selanjutnya diharapkan peneliti lebih mengacu dan memperoleh dari beberapa bagian besar data *testing* yang lebih banyak dengan studi kasus yang berbeda.

DAFTAR PUSTAKA

- [1] Sesilia Novita R, Prihastuti Harsani, and Arie Qur'ania, "Penerapan *K-Nearest Neighbor (KNN)* Untuk Klasifikasi Anggrek Berdasarkan Karakter Morfologi Daun Dan Bunga", *KOMPUTASI*, vol. 15, no. 1, Januari 2018, pp. 118-125.
<https://journal.unpak.ac.id/index.php/komputasi/article/view/1267/1074>
- [2] Chandra Arief Rahardja, Try Juardi, and Halim Agung, "Implementasi Algoritma *K-Nearest Neighbor* Pada Website Rekomendasi Laptop", *JURNALBUANA INFORMATIKA*, vol. 10, no. 1 (2019)
<https://ojs.uaaj.ac.id/index.php/jbi/article/view/1847/1352>
- [3] Ni Luh Gede Pivin Suwirmayanti, "Penerapan Metode *K-Nearest Neighbor* Untuk Sistem Rekomendasi Pemilihan Mobil", *Techno.COM*, vol. 16, no. 2, Mei 2017 : 120-131.
<https://core.ac.uk/download/pdf/289792396.pdf>
- [4] Agus Panoto, Yustina Retno Wahyu Utami, and Wawan Laksito YS, "Penerapan Metode *K-Nearest Neighbor* Untuk Prediksi Kelulusan Mahasiswa Pada STMIK Sinar Nusantara Surakarta",
<https://p3m.sinus.ac.id/jurnal/index.php/TIKomSiN/article/viewFile/284/256>
- [5] Mus Mulyadi Baharuddin, Huzain Azis, and Tasrif Hasanuddin, "Analisis Performa Metode *K-Nearest Neighbor* Untuk Identifikasi Jenis Kaca", *ILKOM Jurnal Ilmiah*, vol. 11, no. 3 (2019).
<https://doi.org/10.33096/ilkom.v11i3.489.269-274>

- [6] Yahya and Winda Puspita Hidayanti, “Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Efektivitas Penjualan Vape (Rokok Elektrik) pada ‘Lombok Vape On’”, *INFOTEK*, vol. 3, no. 2, Juli 2020, hal. 104-114.
<https://dx.doi.org/10.29408/jit.v3i2.2279>
- [7] Rama Aji Pangestu, Sabar Rudiarto, Devi Fitriana, “Aplikasi Web Berbasis Algoritma K-Nearest Neighbor Untuk Menentukan Klasifikasi Barang. Studi Kasus: Perum Peruri”, *JITKOM*, vol. 2, no. 1, Januari 2018
<https://publikasi.mercubuana.ac.id/index.php/jitkom/article/view/5184/2357>
- [8] Muhammad Rivki and Adam Mukharil Bachtiar, “Implementasi Algoritma K-Nearest Neighbor Dalam Pengklasifikasian Follower Twitter Yang Menggunakan Bahasa Indonesia”, *JURNAL SISTEM INFORMASI* vol. 13, no. 1 (2017) 31-37.
<http://dx.doi.org/10.21609/jsi.v13i1.500>
- [9] Ni Luh Ratniasih, “Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Penentuan Mahasiswa Berpotensi Drop Out”, *JuTIK*, vol. 5, no. 3 (2019).
<https://jurnal.undhirabali.ac.id/index.php/jutik/article/view/804/pdf>
- [10] Slamet Wiyono and Taufiq Abidin, “Implementation Of K-Nearest Neighbor (KNN) Algorithm To Predict Student’s Performance”, *SIMETRIS*, vol. 9, no. 2 (2018).
<https://doi.org/10.24176/simet.v9i2.2424>