ANALISIS SENTIMEN REVIEW RESTORAN DI BANDUNG DENGAN TEKS BAHASA INDONESIA MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR (KNN)

ANALYSIS OF RESTAURANT REVIEW SENTIMENT IN BANDUNG WITH INDONESIAN TEXT USING K-NEAREST NEIGHBOR (KNN) CLASSIFICATION

Ryan Ramadhani1, Budhi Irawan2, Casi Setianingsih3
1Prodi S1 Teknik Komputer, Fakultas Teknik, Universitas Telkom
2Prodi S1 Teknik Komputer, Fakultas Teknik, Universitas Telkom
3Prodi S1 Teknik Komputer, Fakultas Teknik, Universitas Telkom
1**ryanramadhanii@student.telkomuniversity.ac.id**, 2**budhiirawan@telkomuniversity.co.id**, 3**setiacasie@gmail.com**

Abstrak

Restoran adalah salah satu tempat yang sering dikunjungi untuk berkumpul bersama kerabat. Pertimbangan seseorang untuk mengunjungi sebuah restoran adalah dari makanan, tempat dan harga. Saat ini mencari hal tersebut tidak susah karena sudah ada beberapa website yang berguna untuk menjadikan wadah untuk menampung ulasan dari pengguna untuk sebuah restoran. Ulasan sendiri memiliki sentimen positif, negatif, dan netral didalamnya, dalam menentukan sentiment tersebut dibutuhkan ketelitian dalam mengartikan setiap kata yang ada pada kalimat. Permasalahan tersebut menciptakan ide untuk membuat sistem untuk menganalisa sentiment dari ulasan agar lebih akurat.

Kata Kunci : analisis sentimen, K-NN, ulasan,akurasi.

Abstract

The restaurant is one of the places frequented to visit with relatives. One's consideration for reporting a restaurant is where to eat, where and price. Currently searching for this is not difficult because there are already several websites that are useful for creating a place to get user reviews for a restaurant. The commentary itself has positive, negative, and neutral sentiments in it, in determining the sentiment it requires carefulness in interpreting every word in the sentence. These problems created the idea of creating a system to analyze sentiment from the review to be more accurate.

Keywords: ***sentiment analysis, K-NN, reviews, accuracy.***

1. Pendahuluan

Saat ini teknologi semakin cepat dalam perkembangannya sehingga dapat merubah gaya hidup manusia dalam menjalani aktivitasnya sehari-hari. Salah satu contohnya adalah dengan munculnya sosial media, saat ini pengguna internet atau disebut netizen menjadikan sosial media sebagai wadah mereka untuk mengutarakan pendapatnya terhadap suatu hal, salah satunya restoran. Dikutip dari berita Kumparan 'Beda Pandangan Food Blogger dan jurnalis Kuliner' dahulu mencari referensi terhadap makanan atau restoran yang enak disebarkan dari beberapa orang yang sudah berkunjung dan mereka menyebarkan ke beberapa orang lainnya. Saat ini, Google, perusahaan asal amerika dapat menghimpun banyak ulasan yang ditulis oleh pengguna internet dan menjadikannya dalam satu wadah yaitu Google Review

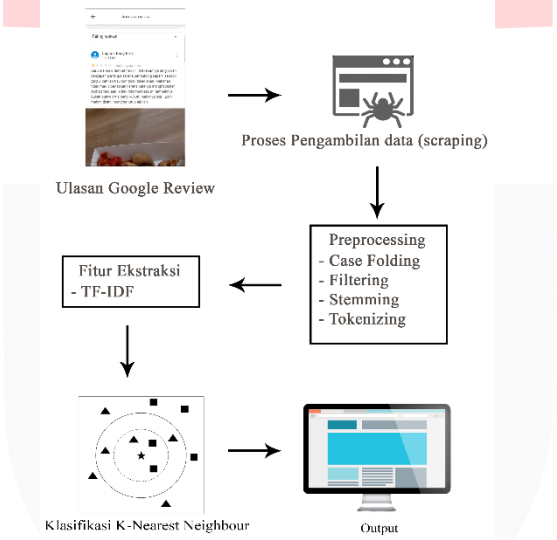
Pengunjung memiliki pertimbangan sendiri sebelum menentukan restoran mana yang akan dikunjungi, contohnya adalah harga dan kualitas produk [1]. Dengan adanya sosial media seperti sekarang, persebaran mengenai informasi detil restoran seperti alamat, waktu operasional maupun ulasan pengunjung dapat diakses melalui internet kapanpun dan dimanapun melalui internet. Salah satunya dengan Google Review. Google Review merupakan sebuah wadah yang didalamnya berisi kumpulan ulasan yang sudah dihimpun dari berbagai situs. Ulasan-ulasan tersebut bisa di

manfaatkan untuk mencari sebuah informasi, dan dalam pemanfaatannya membutuhkan analisis yang tepat sehingga informasi yang dihasilkan dapat membantu banyak pihak untuk mendukung suatu keputusan atau pilihan [2]. Analisa sentimen sering juga di kenal sebagai opinion mining adalah studi komputasi dari pendapat, sentimen, sikap serta emosi yang disajikan dalam sebuah teks [3].

Penelitian ini bermanfaat untuk menciptakan sistem untuk menganalisa sentiment yang terdapat pada Ulasan restoran di Google Review. Proses analisa akan mendeteksi teks menggunakan metode klasifikasi K-Nearest Neighbor (KNN) yaitu sebuah algoritma yang biasa digunakan untuk 2 klasifikasi teks dan data. Pada metode ini dilakukan klasifikasi terhadap objek berdasarkan data yang jaraknya paling dekat dengan objek tersebut [4]. Berdasarkan latar belakang diatas, maka fokus tugas akhir ini mendeteksi sentiment yang dimiliki restoran di daerah bandung berdasarkan ulasan yang ada pada Google Review menggunakan metode klasifikasi K-Nearest Neighbor.

2. Perancangan Sistem

Sistem yang akan dibuat pada Tugas Akhir ini adalah sistem analisa sentiment dari kolom ulasan Google Review. Sistem ini akan menggunakan algoritma klasifikasi K-Nearest Neighbor.



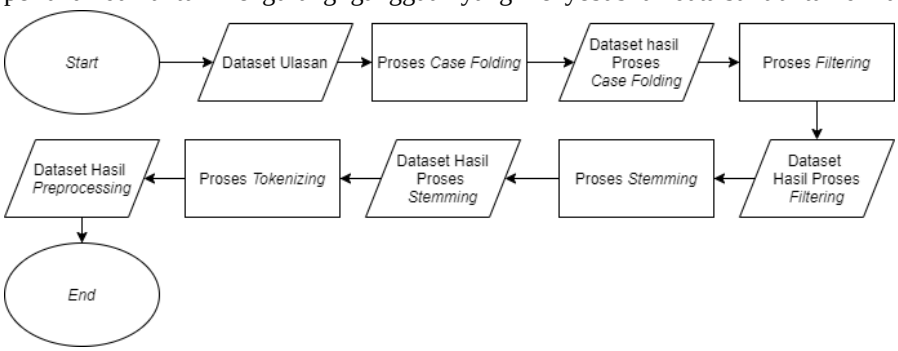
Gambar 1. Gambaran Umum Sistem

Pada gambar 1, dijelaskan alur sistem sebagai berikut :

1. *User* melakukan scraping data ulasan pada Google Review
2. Data tersebut dilakukan proses *pre-processing* dan ekstraksi menggunakan TF-IDF.
3. *User* melakukan klasifikasi data ulasan menggunakan algoritma KNN.
4. Hasil klasifikasi berupa label sentiment dan di tampilkan pada *website*.

2.1 Teks *Pre-Processing*

Teks *Pre-processing* adalah tahap untuk membersihkan data dari kata atau symbol yang tidak diperlukan dan untuk mengurangi gangguan yang menyebabkan data sulit untuk diklasifikasi



Gambar 2. Flowchart Pre-Processing

Pada gambar 3 dapat diketahui bahwa proses dari preprocessing untuk dataset latih dan dataset uji dimulai dengan proses Case Folding, proses Filtering, proses Stemming dan proses Tokenizing. Data yang sudah melewati semua prosesnya akan di pakai untuk tahap selanjutnya.

Sebelum Pre-Processing	Setelah Pre-Processing
Makanannya semua enak, pemandangan bagus, suasana nyaman, udara segar	“makan”, “semua”, “enak”, “pemandangan”, “bagus” , “suasana”, “nyaman” ,”udara” ,“segar”

2.2 TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) adalah suatu metode untuk memberikan bobot kepada suatu istilah. Metode ini banyak digunakan pada klasifikasi teks. Dalam metode merupakan gabuungan dari 2 konsep untuk menghitung bobot istilah. Term Frequency adalah jumlah suatu istilah muncul dalam sebuah dokumen. Inverse Document Frequence adalah pengukur seberapa penting istilah tersebut. Term Frequency dan Inverse Document Frequence digabung untuk menghasilkan bobot dari setiap istilah dalam setiap dokumen [5]. 6 Dalam menghitung Term Frequency pada setiap istilah yang dicari (t) di sebuah dokumen (d) menggunakan persamaan dibawah ini:

$$Tf(t,d) = c(t,d) \quad (2.1)$$

Keterangan:
Tf = Term Frequency
t = term (istilah yang dicari)
d = dokumen (data keseluruhan)
c = frekuensi t dalam d
c(t,d) merupakan jumlah frekuensi term(t) dalam sebuah dokumen(d), untuk menormalkan persamaan 2.1 terhadap Panjang dokumen, maka:

$$Tf(t,d)=(c(t,d))/l(d) \quad (2.2)$$

Keterangan:
Tf = Term Frequency
t = term (istilah yang dicari)
d = dokumen (data keseluruhan)
c(t,d) = jumlah frekuensi t dalam dokumen d
l(d) = jumlah data dala dokumen

dari persamaan (2.2) bisa disimpulkan semakin istilah yang dicari (t) muncul dalam dokumen (d) maka nilai TF yang didapatkan akan semakin besar. Untuk menghitung nilai IDF sebagai berikut:

$$Idf(t) = 1 + \log N/k \quad (2.3)$$

Keterangan:
Idf = Inverse Document Frequency
N = Jumlah total dokumen yang dianalisa
K = Jumlah dokumen yang terdapat istilah (t)
Dapat dilihat dari persamaan diatas bawha IDF mengukur seberapa sering istilah (t) muncul [], Jika dihubungkan antara TF dan IDF maka persamaannya:

$$weight(t,d) = tf(t,d) * idf(t) \quad (2.4)$$

$$\begin{cases} \frac{c(d,t)}{l(d)} * (1 + log\ \frac{N}{k}) & Jika\ c(t,d) \geq 1 \\ 0 & selain\ itu \end{cases} \quad (2.5)$$

Keterangan:
Weight(t,d) = bobot keseluruhan t terhadap d

2.3 K-Nearest Neighbor

K-Nearest Neighbour adalah metode klasifikasi data yang bekerja berdasarkan pembelajaran (*training datasets*), yang diambil dari k tetangga terdekatnya. Data baru yang diklasifikasi diproyeksikan pada ruang dimensi dengan mencari c terdekat dengan c-baru (tetangga) dengan Teknik pencarian yang menggunakan rumus Euclidean.

Tahapan umum dalam *K-Nearrest Neighbor* (KNN) adalah:

1. Menentukan kemiripan antara dokumen uji setiap tetangganya.
2. Berdasarkan hasil pengurutan, dilakukan proses perhitungan menggunakan persamaan *Euclidean Distance* untuk menentukan jarak antara dua titik pada data latih dan data uji.

$$d = \sqrt{\sum_{i=0}^{x}(Xi - Yi)^2} \quad (2.9)$$

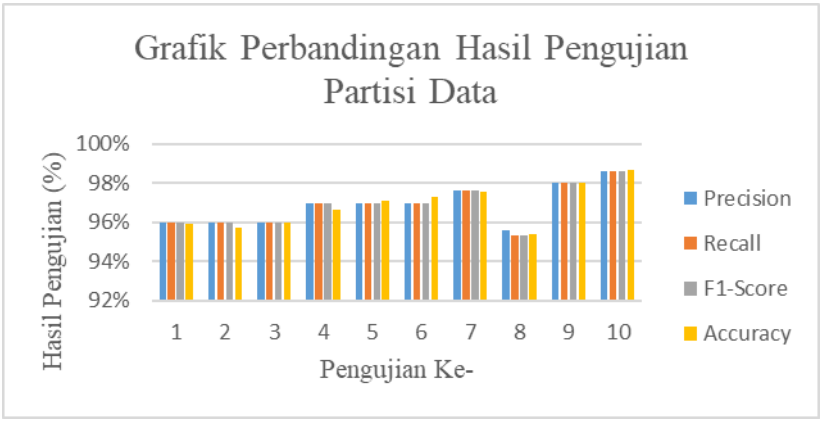
Keterangan:
d = jarak *Euclidean*
X = data 1
Y = data 2
i = data ke-i
n = jumlah data

3. Menentukan parameter K yang akan digunakan untuk menentukan jumlah tetangga acuan (jumlah tetangga paling dekat yang akan dijadikan acuan)
4. Mengurutkan data pada tahap no 2 dengan cara *ascending* kemudian tentukan tetangga yang terdekat berdasarkan jarak minimum ke-K
5. Berdasarkan hasil pengurutan diambil sejumlah nilai K data yang memiliki jarak terkecil. Maka, didapatkan hasil berupak *output* data yang merupakan sentiment positif, negatif, atau netral.

3. Pengujian sistem

Data yang dibutuhkan yaitu berupa teks ulasan. Semua data ulasan tersebut berasal dari Google Review. Terdapat 1619 data yang digunakan dalam pengujian ini. Data tersebut sebelumnya sudah divalidasi kepada pihak Balai Bahasa Jawa Barat dan telah diberi label.

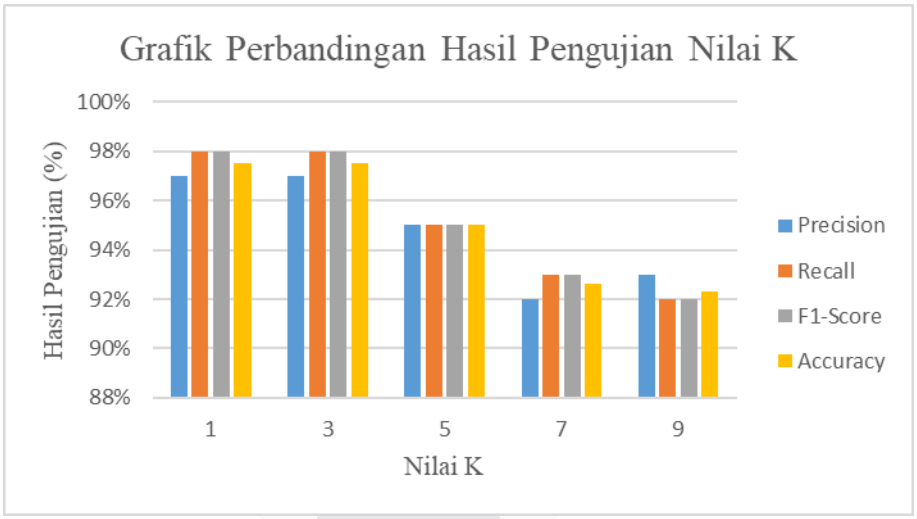
a. Partisi data



Gambar 3. Hasil Pengujian Partisi Data

Pada gambar 4.3 dapat dilihat pengujan kesepuluh dengan partisi data latih sebesar 95 % dan data uji sebesar 5% nilai accuracy yang didapatkan sebesar 98,7%.

b. Nilai K



Gambar 4. Hasil Pengujian Nilai K

Pada gambar 4.4 menunjukan grafik dari hasil pengujian nilai K 1, 3, 5, 7, 9 terhadap nilai precision, recall, F1-score, accuracy. Dari grafik tersebut bisa dilihat nilai K terbaik adalah K=1 dan K=3 dengan precision sebesar 97%, recall sebesar 98%, F1-score sebesar 98% dan accuracy sebesar 97,53%.

Pengujian yang telah dilakukan membuktikan bahwa dalam KNN nilai K sangat berpengaruh terhadap hasil klasifikasinya. Nilai k terbaik tidak selalu sama setiap kondisinya karena tergantung dari data latih yang dimiliki, menggunakan K ganjil disarankan agar menghindari hasil klasifikasi yang sama. Semakin besar nilai K yang dipilih, maka semakin banyak data yang diambil untuk menentukan mayoritas ketetanggaan. Jadi pemilihan nilai K berpengaruh dalam proses klasifikasi menggunakan algoritma KNN.

4. Kesimpulan

Berdasarkan hasil pengujian yang telah dilakukan pada tugas akhir ini, dapat disimpulkan sebagai berikut:

1. Sistem analisa sentimen ulasan restoran dalam Bahasa Indonesia dengan menggunakan metode klasifikasi K-Nearest Neighbor berbasis web berhasil mengklasifikasikan ulasan pada Google Review berupa sentimen positif, sentimen negatif, atau sentimen netral.
2. Berdasarkan pengujian partisi data, Partisi data latih 95% dan data uji 5% merupakan yang terbaik diantara pengujian partisi lainnya dengan nilai akurasi sebesar 98,7% dan nilai precision, recall, F1-score sebesar 98,6% .
3. Berdasarkan pengujian nilai K, nilai K terbaik adalah 1 & 3 dengan mendapatkan tingkat akurasi sebesar 97,53%.

Daftar Pustaka:

[1] I. Nadiati, "PENGARUH HARGA DAN KUALITAS PRODUK TERHADAP KEPUASAN KONSUMEN PADA KAMBINC EATABLES RESTO & KAFE," 2017.

[2] F. Afshoh, "ANALISA SENTIMEN MENGGUNAKAN NAÏVE BAYES UNTUK MELIHAT PERSEPSI MASYARAKAT TERHADAP KENAIKAN HARGA JUAL ROKOK PADA MEDIA SOSIAL TWITTER," 2017.

[3] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012.

[4] R. S. P. M. A. F. Winda Estu Nurjanah, "Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 1, no. 12, pp. 1750-1757 , 2017.

[5] Y. A. S. A. A. Syafitri Hidayatul Annur Aini, "Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* , vol. 2, no. 9, pp. 2546-2554 , 2018.