INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v14i2.1216

VOL. 14. NO. 2 FEBRUARI 2020
P-ISSN: 0216-6933 | E-ISSN: 2685-807X

KOMPARASI ALGORITMA KLASIFIKASI TEXT MINING PADA REVIEW RESTORAN

Lila Dini Utami

Sistem Informasi Akuntansi Kampus Kota Bogor
Universitas Bina Sarana Informatika
www.bsi.ac.id
lila.ldu@bsi.ac.id

Abstract—At this time, where the development of technology is developing very rapidly, and everyone has the right to express his opinion on a matter. One of them is conducting a review of a restaurant. The review, can be created from food, decoration, or service. This, is used by business people to find out consumer ratings about the restaurants they manage. However, the review data must be processed using the right algorithm. Then this research is conducted to find out which algorithm is more feasible to use to get the highest accuracy. The method used is Naïve Bayes (NB), and k-Nearest Neighbor (k-NN). From the process that has been done, it is obtained that the accuracy of Naïve Bayes is 75.50% with a Kappa value of 0.510, and the accuracy results when using the k-Nearest Neighbor algorithm is 89.50% with the AUC value of 0.790. The use of the k-Nearest Neighbor algorithm can help in making more appropriate decisions for hotel reviews at this time, because the resulting accuracy is greater than the Naïve Bayes Algorithm.

Keywords*: restaurant review, naïve bayes, k-NN*

Abstrak—Pada saat ini, dimana perkembangan teknologi berkembang sangat pesat, dan setiap orang memiliki hak untuk menyampaikan pendapatnya mengenai suatu hal. Salah satunya adalah melakukan review terhadap sebuah restoran. Review tersebut, tercipta bisa dari makanan, dekorasi, ataupun pelayanannya. Hal ini, dimanfaatkan oleh para pelaku bisnis untuk mengetahui penilaian konsumen tentang restoran yang mereka kelola. Namun data review tersebut harus diolah menggunakan algoritma yang tepat. Maka penelitian ini dilakukan untuk mengetahui algoritma yang lebih layak digunakan untuk mendapatkan akurasi yang paling tinggi. Adapun metode yang digunakan adalah Naïve Bayes (NB), dan k-Nearest Neighbor (k-NN). Dari proses yang telah dilakukan didapatkan hasil akurasi Naïve Bayes adalah 75,50% dengan nilai Kappa adalah 0,510, dan hasil akurasi jika menggunakan algoritma k-Nearest Neighbor adalah 89,50% dengan nilai AUC adalah 0,790. Penggunaan algoritma k-Nearest Neighbor dapat membantu dalam pengambilan keputusan yang lebih tepat untuk review hotel pada saat ini, karena akurasi yang dihasilkan lebih besar dibandingkan dengan Algoritma Naïve Bayes.

Kata kunci: *review restoran, naïve bayes, k-NN*

PENDAHULUAN

Dalam usaha menciptakan sebuah reputasi restoran yang baik, ada beberapa hal yang diperhatikan oleh pemilik restoran yaitu pendapat konsumen mengenai makanan, pelayanan yang diberikan selama berada di restoran, dan juga desain interior yang menjadi bagian dari ciri khas restoran tersebut (Pradini & Wempi, 2019).

Di era maraknya penggunaan internet saat ini, jumlah konsumen yang menulis opini dan pengalaman secara online terus meningkat. Membaca review tersebut secara keseluruhan bisa memakan waktu, namun, jika hanya sedikit review yang dibaca, maka evaluasi akan bias. Klasifikasi sentimen bertujuan untuk mengatasi masalah ini dengan secara otomatis mengelompokkan review pengguna menjadi opini positif atau negatif (Muthia, 2016).

Rating merupakan bagian terpenting yang akan dilihat oleh konsumen yaitu mencerminkan dengan benar baik atau buruknya dalam peringkat *rating* (Paramitha, Cholissodin, & Dewi, 2019). Salah satunya pada penelitian ini menggunakan *review* beserta *rating* restoran. *Rating* dari *review* restoran dapat membantu konsumen dan pemilik restoran dalam mengetahui kualitas restoran, baik itu menu ataupun pelayanannya, layak atau tidak untuk dipilih. *Review* ini tidak hanya dibutuhkan oleh konsumen, tapi juga pemilik restoran. Pemilik

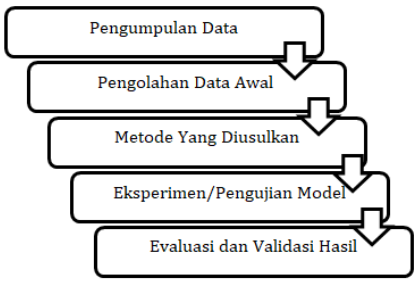
restoran dapat melihat bagaimana respon dari konsumen. *Rating* yang digunakan yaitu *rating* 1 hingga *rating* 5, semakin rendah *rating* yang diberikan semakin kurang bagus begitupun sebaliknya jika *rating* yang diberikan tinggi maka semakin baik kualitasnya (Paramitha et al., 2019).

Beberapa tahun terakhir, telah banyak dilakukan penelitian terkait dengan analisis review, diantaranya:

1. Review restoran yang diambil berasal dari www.zomato.com. Akurasi yang didapat dari review positif dan negatif menggunakan Algoritma Naïve Bayes adalah sebesar 86,50% (Muthia, 2017)
2. Penelitian berikutnya dari (Pratama, Sari, & Adikara, 2018), nilai akurasi dari data uji didapatkan setelah menghitung jumlah dokumen dengan kelas benar dibagi dengan jumlah dokumen keseluruhan. Jumlah data uji yang digunakan dalam pengujian adalah 100 *review* konsumen yang terdiri dari 50 *review* positif dan 50 *review* negatif. Pada pengujian pertama menggunakan persentase 25% dari seleksi fitur didapatkan nilai akurasi sebesar 81%. Pada pengujian kedua menggunakan persentase 50% seleksi fitur didapatkan nilai akurasi sebesar 80%. Pada pengujian ketiga menggunakan persentase 75% dari seleksi fitur didapatkan nilai akurasi sebesar 77%. Pada pengujian terakhir tanpa menggunakan seleksi fitur didapatkan nilai akurasi sebesar 80%.

Berbagai macam situs review restoran, salah satunya adalah situs www.yelp.com. Situs ini menyediakan fitur pencarian restoran yang dilengkapi dengan review serta pemberian tingkatan bintang mulai dari bintang 1 hingga bintang 5 yang diberikan oleh pengguna. Review yang dituliskan punbermacam-macam, mulai dari keadaan makanan yang disajikan, wadah penyajian makanan atau serta kenyamanan dan layanan yang diberikan oleh pihak restoran.

BAHAN DAN METODE



Sumber: (Septiani, 2017)

Gambar 1. Tahapan Penelitian

1. Pengumpulan Data

Pada penelitian ini, penulis melakukan pengumpulan data dengan cara mengambil *sample* secara acak yakni berupa review restoran yang ada di Amerika dari sebuah website, yakni www.yelp.com yang terdiri dari 100 komentar positif dan 100 komentar negatif

2. Pengolahan Data

Processing yang dilakukan, diantaranya adalah:

a. *Tokenization*

Tokenization adalah proses memecah teks menjadi sebuah frasa, kata, simbol atau elemen bermakna, lainnya disebut token (Verma, Renu, & Gaur, 2014). Dalam proses *tokenization* ini, semua kata dikumpulkan kemudian dihilangkan tanda baca, spasi, simbol atau apapun yang bukan huruf.

Tabel 1. Hasil Proses Tokenization

Review	***Tokenization***
This place is good but not worth the money they charge. The lobster roll I got was the size of a hot-dog. It isn't worth the $15 per roll to not be full afterwards. There are plenty of great places in the area that are way cheaper, just as good, and more filling. No one wants to go back to work hungry.	*This place is good but not worth the money they charge The lobster roll I got was the size of a hot dog It isn t worth the per roll to not be full afterwards There are plenty of great places in the area that are way cheaper just as good and more filling No one wants to go back to work hungry*

Sumber: (Utami, 2019)

b. *Stopwords Removal*

Stopwords Removal didefinisikasn sebagai sekumpulan kata yang tidak berhubungan (irrelevant) dengan subyek utama yang dimaksud, meskipun kata tersebut sering muncul didalam data yang digunakan (Setiawan, Kurniawan, & Handiwidjojo, 2013). Dalam proses *stopwords removal* ini, penulis melakukan penghapusan kata-kata yang tidak relevan seperti *the*, *of*, *for*, *with*, dan sebagainya.

Tabel 2. Hasil Proses *Stopwords Removal*

Review	***Tokenization***
This place is good but not worth the money they charge. The lobster roll I got was the size of a hot-dog. It isn't worth the $15 per roll to not be full afterwards. There are plenty of great places in the area that are way cheaper, just as good, and more filling. No one wants to go back to work hungry.	*This place is good but not worth the money they charge. The lobster roll I got was the size of a hot-dog. It isn't worth the $15 per roll to not be full afterwards. There are plenty of great places in the area that are way cheaper, just as good, and more filling. No one wants to go back to work hungry.*

Sumber: (Utami, 2019)

VOL. 14. NO. 2 FEBRUARI 2020
P-ISSN: 0216-6933 | E-ISSN: 2685-807X

c. *Stemming*

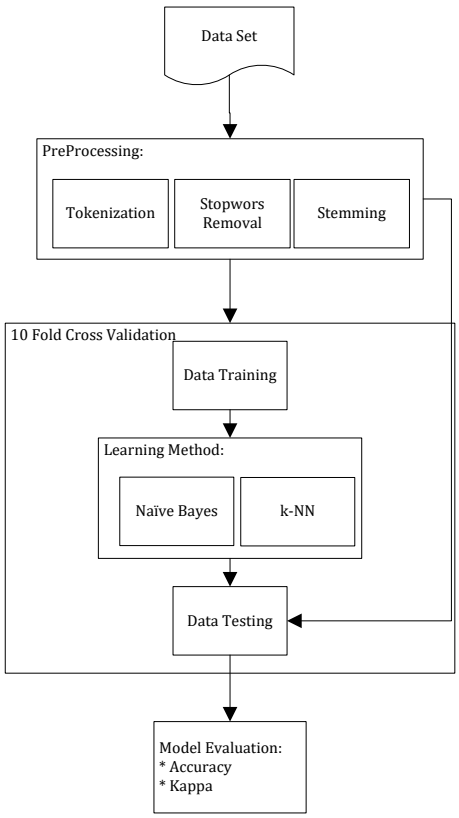
Stemming adalah proses menghilangkan imbuhan dari suatu kata dari suatu dokumen atau query menjadi root word (Novitasari, 2016). Pada proses *stemming* ini, dilakukan pengelompokkan kata kedalam beberapa kelompok yang memiliki kata dasar yang sama, seperti drug, drugged, dan drugs dimana kata dasar dari semuanya adalah kata drug.

Tabel 3. Hasil Proses *Stemming*

Review	***Tokenization***
This place is good but not worth the money they charge. The lobster roll I got was the size of a hot-dog. It isn't worth the $15 per roll to not be full afterwards. There are plenty of great places in the area that are way cheaper, just as good, and more filling. No one wants to go back to work hungry.	*this place is good but not worth the money they charge. the lobster roll i got was the size of a hot-dog. it isn't worth the $15 per roll to not be full afterwards. there are plenty of great places in the area that are way cheaper, just as good, and more filling. no one wants to go back to work hungry.*

Sumber: (Utami, 2019)

3. Metode Yang Diusulkan



Sumber: (Utami, 2019)

Gambar 2. Metode Yang Diusulkan

a. Algortimat Naïve Bayes

Menurut Markov dan Daniel dalam, tahapan dalam algoritma Naïve Bayes, adalah:

a) Hitung probabilitas bersyarat/likelihood:

$$P(x|C) = P(x1, x2, \dots, xn\,|C) \quad \dots\dots (1)$$

C : Class
x : Vektor dari nilai atribut n
P(xi|C) : Proporsi dokumen dari class C yang mengandung nilai atribut xi

b) Hitung probabilitas prior untuk tiap class:

$$P(C) = \frac{Nj}{N} \quad \dots\dots (2)$$

Nj : Jumlah dokumen pada suatu class
N : Jumlah total dokumen

c) Hitung probabilitas posterior dengan rumus:

$$P(C|x) = \frac{P(x|C)P(C)}{P(x)} \quad \dots\dots (3)$$

Menurut Santoso dalam (Muthia, 2017), dengan kata-kata yang lebih umum, rumus Bayes bisa diberikan sebagai berikut:

$$\text{Posterior} = \frac{likelihood\ x\ prior}{evidence} \quad \dots\dots (4)$$

b. Algoritma K-Nearest Neighbor (K-NN)

Algoritma K-Nearest Neighbor (K-NN) adalah suatu metode yang menggunakan algoritma supervised. K-NN termasuk kelompok instance-based learning.Algoritma ini juga merupakan salah satu teknik lazy learning.K-NN dilakukan dengan mencari kelompok k objek dalam data trainingyang paling dekat (mirip) dengan objek pada data baru atau data testing (Mustakim & Oktaviani, 2016)

Persamaan berikut ini adalah rumus untuk menghitung jarak pada algoritma KNN dengan metode cosine similiarity (Rivki & Bachtiar, 2017):

$$\cos(\theta_{QD}) = \frac{\sum_{i=1}^{n} D_i Q_i}{\sqrt{\sum_{i=1}^{n}(Q_i)^2}.\sqrt{\sum_{i=1}^{n}(D_i)^2}} \quad \dots\dots (3)$$

Dimana,
$\cos(\theta_{QD})$: Kemiripan Q terhadap dokumen D
Q : Data Uji
D : Data Latih
n : Banyaknya Data

HASIL DAN PEMBAHASAN

1. Hasil Menggunakan Algoritma Naïve Bayes

Dari 100 review positif dan 100 review negatif, sebanyak 93 review yang sesuai yaitu negatif, dan sebanyak 7 review diprediksi negatif tetapi ternyata positif. Sebaliknya, sebanyak 58 review yang sesuai yaitu positif, dan sebanyak 42 review diprediksi positif tetapi ternyata negatif. Hasil akurasi dari penerapan Rapid Miner 9.5 dengan menggunakan Algoritma Naïve Bayes adalah 75,50% dan kappa sebesar 0,510.

Tabel 4. *Accuracy* menggunakan Algoritma Naïve Bayes

***Accuracy* : 75,50% +/- 7,98% (*micro average*: 75,50%)**			
	True Negative	***True Positive***	***Class Precision***
Pred. Negative	93	42	68,89%
Pred. Posotive	7	58	89.23%
Class Recall	93,00%	58,00%	

Sumber: (Utami, 2019)

Tabel 5. *Kappa* menggunakan Algoritma Naïve Bayes

***Kappa* : 0,510 +/- 0,160 (*micro average*: 0,510)**			
	True Negative	***True Positive***	***Class Precision***
Pred. Negative	93	42	68,89%
Pred. Posotive	7	58	89.23%
Class Recall	93,00%	58,00%	

Sumber: (Utami, 2020)

$$Accuracy = \frac{(TN + TP)}{(TN + FN + TP + FP)}$$
$$Accuracy = \frac{(93 + 58)}{(93 + 7 + 58 + 42)}$$
$$Accuracy = \frac{151}{200} = 0{,}755 = 75{,}50\%$$

2. Hasil Menggunakan Algoritma K-Nearest Neighbor (K-NN)

Dari 100 review positif dan 100 review negatif, sebanyak 87 review yang sesuai yaitu negatif, dan sebanyak 13 review diprediksi negatif tetapi ternyata positif. Sebaliknya, sebanyak 92 review yang sesuai yaitu positif, dan sebanyak 8 review diprediksi positif tetapi ternyata negatif.

Hasil akurasi dari penerapan Rapid Miner 9.5 dengan menggunakan Algoritma K-Nearest Neighbor (K-NN) adalah 89,50% dan kappa sebesar 0,790.

Tabel 6. *Accuracy* menggunakan Algoritma K-Nearest Neighbor (K-NN)

***Accuracy* : 89,50% +/- 5,99% (*micro average*: 89,50%)**			
	True Negative	***True Positive***	***Class Precision***
Pred. Negative	87	8	91,58%
Pred. Posotive	13	92	87,62%
Class Recall	87,00%	92,00%	

Sumber: (Utami, 2019)

Tabel 7. *Kappa* menggunakan Algoritma K-Nearest Neighbor (K-NN)

***Kappa* : 0,790 +/- 0,120 (*micro average*: 0,790)**			
	True Negative	***True Positive***	***Class Precision***
Pred. Negative	87	8	91,58%
Pred. Posotive	13	92	87,62%
Class Recall	87,00%	92,00%	

Sumber: (Utami, 2019)

$$Accuracy = \frac{(TN + TP)}{(TN + FN + TP + FP)}$$
$$Accuracy = \frac{(87 + 92)}{(87 + 13 + 92 + 8)}$$
$$Accuracy = \frac{179}{200} = 0{,}895 = 89{,}50\%$$

3. Hasil Penelitian

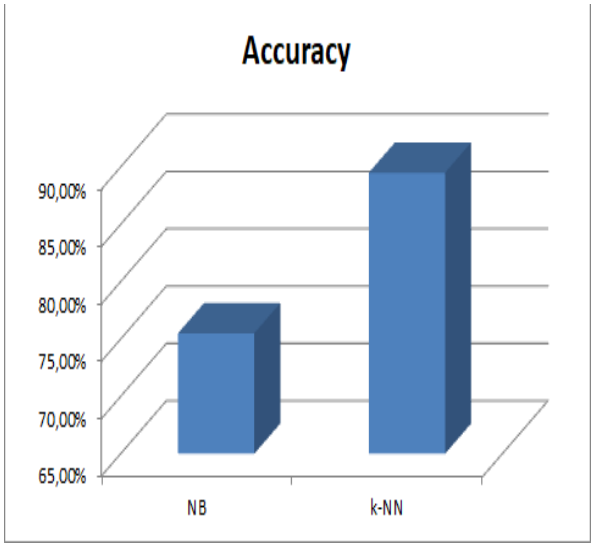
Dari hasil kedua algoritma Naïve Bayes dan K-Nearest Neighbor (K-NN)

Tabel 8. Komparasi *Accuracy* dan *Kappa*

	Accuracy	**Kappa**
NB	75,50%	0,510
k-NN	89,50%	0,790

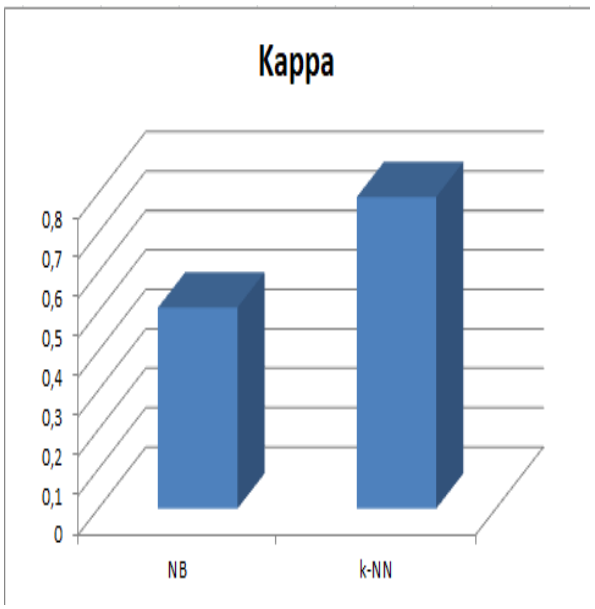
Sumber: (Utami, 2019)

VOL. 14. NO. 2 FEBRUARI 2020
P-ISSN: 0216-6933 | E-ISSN: 2685-807X



Sumber: (Utami, 2019)

Gambar 3. Grafik *Accuracy* Algoritma Klasifikasi



Sumber: (Utami, 2019)

Gambar 3. Grafik *Kappa* Algoritma Klasifikasi

KESIMPULAN

Data review restoran dapat diklasifikasi dengan baik kedalam bentuk positif dan negatif. Akurasi algoritma Naïve Bayes (NB) mencapai 75,50% dengan Kappa 0,510. Jika menggunakan algoritma k-Nearest Neighbor (k-NN) mencapai 89,50% dengan AUC 0,790, yang artinya meningkat 14% dari akurasi Naïve Bayes. Hasil dari komparasi algoritma klasifikasi ini adalah, antara algoritma Naïve Bayes (NB), dan k-Nearest Neighbor (k-NN) didapatkan k-NN dengan hasil terbaik dengan akurasi 89,50% dengan Kappa adalah 0,790.

REFERENSI

Mustakim, & Oktaviani, G. (2016). Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa. *Jurnal Sains, Teknologi Dan Industrii*, *13*(2), 195–202.

Muthia, D. A. (2016). Opinion Mining Pada Review Buku Menggunakan Algoritma Naive Bayes. *Jurnal Teknik Komputer AMIK BSI*, *2*(1), 1–8. Retrieved from http://ejournal.bsi.ac.id/ejurnal/index.php/jtk/article/viewFile/357/266

Muthia, D. A. (2017). Analisis Sentimen Pada Review Restoran Dengan Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes. *Jurnal Ilmu Pengetahuan Dan Teknnologi Komputer*, *2*(2), 39–45.

Novitasari, D. (2016). Perbandingan Algoritma Stemming Porter Dengan Arifin Setiono Untuk Menentukan Tingkat Ketepatan Kata Dasar. *Jurnal String*, *1*(2), 120–129. Retrieved from http://journal.lppmunindra.ac.id/index.php/STRING/article/view/1031

Paramitha, D. T. A., Cholissodin, I., & Dewi, C. (2019). Prediksi Rating Otomatis Berdasarkan Review Restoran pada Aplikasi Zomato dengan menggunakan Extreme Learning Machine (ELM). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(5), 4687–4693. Retrieved from http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/5314

Pradini, R. P., & Wempi, J. A. (2019). Desain Interior Sebagai Medium Komunikasi Nonverbal Restoran Eat Happens dalam Membentuk Reputasi. *Profesi Humas*, *3*(2), 177–201. Retrieved from http://journal.unpad.ac.id/profesi-humas/article/view/18734

Pratama, N. D., Sari, Y. A., & Adikara, P. P. (2018). Analisis Sentimen Pada Review Konsumen Menggunakan Metode Naive Bayes Dengan Seleksi Fitur Chi Square Untuk Rekomendasi Lokasi Makanan Tradisional. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(9), 2982–2988.

Rivki, M., & Bachtiar, M. (2017). Implementasi Algoritma K-Nearest Neighbor Dalam Pengklasifikasian Follower Twitter Yang

Menggunakan Bahasa Indonesia. *Jurnal Sistem Informasi*, *13*(1), 31–37.

Septiani, W. D. (2017). Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 dan Naive Bayes Untuk Predikasi Penyakit Hepatitis. *Jurnal Pilar Nusa Mandiri*, *13*(1), 76–84.

Setiawan, A., Kurniawan, E., & Handiwidjojo, W. (2013). Implementasi Stop Word Removal Untuk Pembangunan Aplikasi Alkitab Berbasis Windows 8. *Jurnal EKSIS*, *6*(2), 1–11.

Utami, L. D. (2019). *Laporan Akhir Penelitian: Komparasi Algoritma Klasifikasi Text Mining Pada Review Restoran*. Jakarta.

Verma, T., Renu, & Gaur, D. (2014). Tokenization and Filtering Process in RapidMiner. *International Journal of Applied Information System (IJAIS)*, *7*(2), 16–18. Retrieved from https://pdfs.semanticscholar.org/d024/33ad6b77f740fb4f43673eed9b80b0ccb199.pdf

