Jurnal Teknik Komputer AMIK BSI
Volume 8, No.2, Juli 2022
P-ISSN 2442-2436, E-ISSN: 2550-0120
Akreditasi Ristekdikti, No: 36/E/KPT/2019 (Sinta 4)
DOI: 10.31294/jtk.v4i2

Pengujian Algoritma Teks Mining Untuk Klasifikasi Analisis Review Aplikasi Halodoc

Eka Rini Yulia[1], Kusmayanti Solecha[2]

[1] Universitas Nusa Mandiri
eka.erl@nusamandiri.ac.id

[2]Universitas Bina Sarana Informatika
kusmayanti.ksc@bsi.ac.id

Diterima	Direvisi	Disetujui
23-05-2022	08-06-2022	14-06-2022

Abstrak - Memasuki era revolusi industry 4.0 teknologi berkembang sangat pesat, teknologi sistem informasi di bidang kesehatan adalah *e-health* sebagai teknologi informasi dan komunikasi yang efektif dan aman dalam mendukung hal-hal yang berkaitan dengan bidang kesehatan seperti pelayanan kesehatan, pengawasan kesehatan, referensi tentang hal-hal kesehatan. Konsumen yang menulis review, opini dan pengalaman dalam telekonsultasi medis terus meningkat. Dataset halodoc tersebut harus diolah menggunakan algoritma yang tepat. Maka hasil dari penelitian ini untuk mengetahui algoritma yang lebih baik digunakan untuk mendapatkan model algoritma terbaik. Peneliti membandingkan beberapa metode klasifikasi Teks Mining diantaranya yaitu Algoritma C4.5, *K-Nearest Neighbor* (K-NN) dan *Support Vector Machine* (SVM). Tahapan penelitian yang dilakukan dimulai dari pengumpulan data, pengolahan data awal, metode yang diusuilkan yaitu C4.5, K-NN dan SVM, menggunakan pengujian metode 10-*folds cross Validation*, hasil evaluasi dan pengujian menggunakan t-test, metode yang diusulkan. Dari proses yang telah dilakukan didapatkan hasil akurasi terbaik hasil akurasi menggunakan algoritma *K-Nearest Neighbor* (K-NN) adalah **88,50%** dengan nilai **AUC: 0,960**. Sedangkan, hasil model terbaik menggunakan uji t-test yaitu algoritma : *Support Vektor Machine* dan K-NN dalam pengujian dataset halodoc.

Kata Kunci: uji t-tes, algoritma, teks mining

Abstract *- Entering the era of the industrial revolution 4.0 technology is developing very rapidly, information system technology in the health sector is e-health as information and communication technology that is effective and safe in support matters related to the health sector such as health services, health supervision, references on matters relating to health. health matters. Consumers who write reviews, opinions, and experiences in medical teleconsultation continue to increase. Halodoc dataset must be processed using the right algorithm. So the results of this study are to find out which algorithm is better used to get the best algorithm model. Researchers compared several classifications of Mining Text, including the C4.5 Algorithm, K-Nearest Neighbor (K-NN), and Support Vector Machine (SVM). The research stages start with data collection, and data processing, the proposed methods are C4.5, K-NN and SVM, using the 10-fold cross-validation method, evaluation results, and testing using t-test, the proposed method. From the process that has been carried out, the accuracy results obtained using the K-Nearest Neighbor (K-NN) algorithm are 88.50% with an AUC value: of 0.960. Meanwhile, the results of the best model using the t-test test, namely the algorithm: Support Vector Machine and K-NN in testing the Halodoc dataset.*

Keywords: t-test, algorithm, text mining

PENDAHULUAN

Memasuki era revolusi industry 4.0 teknologi informasi memasuki berbagai bidang seperti: pendidikan, ekonomi, sosial budaya dan kesehatan. Perkembangan teknologi menyebabkan peningkatan pesat pada bidang telekomunikasi dengan hadirnya internet. Kemudahan dalam penggunaan teknologi informasi khusunya internet membuat para pengguna internet dengan mudah mendapatkan informasi yang diinginkan.

Dari tahun ke tahun peningkatan dibidang teknologi menyebabkan perubahan dalam gaya hidup masyarakat. Semakin sadarnya masyarakat

dengan banyaknya kemudahan dan manfaat yang disajikan akan adanya internet, sehingga munculnya peluang bisnis *e-commerce* dan perusahana startup berbasis telekonsultasi medis di Indonesia.

Masyarakat kini tengah berfokus tentang kesehatan banyak yang menggunakan internet untuk mencari solusi kesehatan dirinya, Salah satu perkembangan teknologi sistem informasi di bidang kesehatan adalah *e-health* sebagai teknologi informasi dan komunikasi yang berbiaya efektif dan aman dalam mendukung hal-hal yang berkaitan dengan bidang kesehatan seperti pelayanan kesehatan, pengawasan kesehatan, referensi tentang hal-hal kesehatan, pendidikan tentang kesehatan untuk pengetahuan.(Putra & Suryanata, 2021)

Dengan maraknya masyarakat yang menggunaan internet, banyaknya pengguna yang menulis opininya dan berbagi cerita pengalamannya secara online terus meningkat. Memakan banyak waktu dalam membaca review-review tersebut, sementara, jika membaca hanya sebagian review, akan bias hasil evaluasinya. Klasifikasi sentimen bertujuan dalam mengatasi setiap permasalahan review maka dari itu secara otomatis akan membuat kelompok review dari pengguna menjadi opini positif atau negative. (Muthia, 2017)

Klasifikasi merupakan cara pengelompokkan benda berdasarkan ciri-ciri yang dimiliki oleh objek klasifikasi. Dalam prosesnya, klasifikasi dapat dilakukan dengan banyak cara baik secara manual ataupun dengan bantuan teknologi. Klasifikasi yang dilakukan secara manual adalah klasifikasi yang dilakukan oleh manusia tanpa adanya bantuan dari algortima cerdas komputer. Sedangkan klasifikasi yang dilakukan dengan bantuan teknologi. (Wibawa, Guntur, Purnama, Akbar, & Dwiyanto, 2018)

Banyaknya penelitian terkait dengan analisis review, beberapa tahun terakhir, diantaranya,

1. (Taufik, 2018) membandingkan NB, SVM, Decission Tree (C4.5) dan Metode NB dengan Seleksi Fitur *Particle Swarm Optimization*, hasil yang didapat dari perbandingan keempat metode algoritma tersebut, tingkat akurasi yang lebih baik dalam review klasifikasi hotel indonesia menggunakan algoritma Decission Tree (C4.5) dengan tingkat akurasi sebesar 96,94%.
2. (L. D. Utami et al., 2018) menerapkan NB, SVM dan k-Nearest Neighbor (k-NN) dalam melakukan analisis sentimen review hotel, hasil yang didapat dari perbandingan ketiga metode tersebut tingkat akurasi penggunaan algoritma k-Nearest Neighbor mencapai 75,00% dapat membantu dalam pengambilan keputusan yang lebih tepat untuk review hotel.
3. (Ipmawati, Kusrini, & Luthfi, 2017) menerapkan NB, SVM dan k-NN hasil dari komparasi menunjukkan SVM memperoleh hasil yang baik dalam akurasi pada data imdb review film 78,55%.
4. (L. A. Utami, 2017) menerapkan SVM berbasis PSO dan K-NN, Algoritma SVM berbasis PSO terbukti memberikan solusi terhadap dalam permasalahan klasifikasi opini berita kebakaran hutan agar berita lebih akurat dan optimal.

Peneliti membandingkan beberapa metode klasifikasi Teks Mining, diantaranya yaitu Algoritma C4.5, K-NN dan SVM. Hasil klasifikasi dari masing-masing algoritma akan dibandingkan dan dilihat tingkat akurasinya kemudian di lakukan uji t-test dalam menentukan model algoritma terbaik.

METODOLOGI PENELITIAN

Dalam penelitian ini, tiga algoritma akan digunakan oleh peneliti yaitu k-NN, C4.5 dan SVM. Maksud dari penelitian ini adalah untuk mengetahui algoritma mana yang mendapatkan nilai akurasi yang terbaik antara algoritma. Pada tahapan penelitian yang akan dilakukan adalah :

1. Pengumpulan data
   Data yang diambil peneliti berasal dari aplikasi halodoc, yaitu review tentang bagaimana pelayanan aplikasi tersebut. Data yang digunakan sebanyak 200 data dengan rincian 100 data negatif dan 100 data positif.

2. Pengelolahan data awal
   Dari data yang terlah tersedia sebanyak 200 data, 100 komentar positif dan 100 komentar negatif. Dataset tersebut akan melalui tiga tahapan proses yaitu:
   a. Tokenization
      Tokenize digunakan oleh peneliti untuk menghapus, memisahkan dan menyempurnakannya kata, simbol dari tanda baca (JianQiang & Xiaolin, 2017).

Tabel 1. Hasil Proses Tokenization

Review	Tokenization
Keperluan untuk kesehatan sekarang lebih mudah, transaksi mudah & pengiriman cepat (gak nyampe 1 jam) udah langsung dapet barangnya. Bisa konsultasi seputar kesehatan pula dengan dokter dokter ahli. Kemudahan dalam 1 genggaman saja. Mantapp	Keperluan untuk kesehatan sekarang lebih mudah transaksi mudah pengiriman cepat gak nyampe jam udah langsung dapet barangnya Bisa konsultasi seputar kesehatan pula dengan dokter dokter ahli Kemudahan dalam genggaman saja Mantapp

   b. Stopword Removal
      Kata-kata berhenti dianggap peneliti sangat berperan negative dalam klasifikasi, kata-kata umum seperti is, the, at, if, or, etc

(JianQiang & Xiaolin, 2017).

Tabel 2. Hasil Proses Stopword Removal

Review	Stopword Removal
Keperluan untuk kesehatan sekarang lebih mudah, transaksi mudah & pengiriman cepat (gak nyampe 1 jam) udah langsung dapet barangnya. Bisa konsultasi seputar kesehatan pula dengan dokter dokter ahli. Kemudahan dalam 1 genggaman saja. Mantapp	keperluan untuk kesehatan sekarang lebih mudah transaksi mudah pengiriman cepat gak nyamp jam udah langsung dapet barangnya bisa konsultasi seputar kesehatan pula dengan dokter dokter ah kemudahan dalam genggaman saja mantapp

c. Stemming
Menghilangkan kata akhir dalam mendeteksi kata dasar, banyak kata yang berkurang, dan memberikan hasil yang baik (Symeonidis, Effrosynidis, & Arampatzis, 2018)

Tabel 3. Hasil Proses Stemming

Review	Stemming
Sangat Membantu. Terutama di kondisi Pandemi. Tapi tolong diEvaluasi Ketersediaan Obat di Apotik rekanan HALODOC di Kota MANADO. Untuk resep obat batuk biasa saja Apotik suka habis. Jadi sia2 konsultasi dgn dokter ujung2nya obat yg diresepkan ga ada di Apotik. Jadi rugi dong saya harus chat kembali	sangat membantu terutama di kondisi pandemi tapi tolong dievaluasi ketersediaan obat di apotik rekanan halodoc di kota manado untuk resep obat batuk biasa saja apotik suka habi jadi sia konsultasi dgn dokter ujung nya obat yg diresepkan ga ada di apotik jadi rugi dong saya harus chat kembali

3. Metode yang diusulkan
Untuk pengujian algoritma kalsifikasi review halodoc, peneliti mengusulkan metode yang dipakai adalah k-NN, C4.5 dan SVM dalam mendapatkan hasil akurasi yang terbaik, kemudian akan dilakukan uji pemodelan algoritma yang terbaik dengan menggunakan uji t-test.
4. Pengujian dan Eksperimen Metode
Proses pengujian yg dilakukan peneliti menggunakan RapidMiner Studio Versi 9.10.1. Dataset yang digunakan diambil adalah review dari situs https://play.google.com/store/apps/details?id=com.linkdokter.halodoc.androi d&hl=in dan dataset tersebut dikelompokan menjadi dua yang terdiri dari 100 dataset review negatif dan 100 dataset review positif.
5. Validasi dan Evaluasi Hasil Penelitian
Dalam *Preprocessing* data akan dilakukan *tokenize, stopwords removal* dan *stemming*. Kemudian masuk ke tahap *cross-validation*. *Cross-validation* digunakan untuk menghindari *overlapping* pada data testing (Ipmawati et al., 2017). Peneliti menggunakan validasi standart, 10 *folds cross-validation* karena proses tersebut membagi data secara acak ke 10 bagian. Sehingga penelitian ini akan menghasilkan nilai AUC dan akurasi serta mendapatkan model algoritma terbaik.
6. Uji T-Test
Metode pengujian t-test ini dalam pengujiannya menentukan dua sample yang tidak memiliki hubungan dengan nilai rata-rata yang berbeda. (Rahayuningsih, 2019).

HASIL DAN PEMBAHASAN

Penelitian ini dimaksud untuk menguji 3 algoritma klasifikasi yaitu k-NN, C4.5 dan SVM pada klasifikasi sentimen analisis serta mencari model algoritma yang terbaik dengan melakukan uji t-test.

Penelitian menggunakan 1 dataset yang berjumlah 200 terdiri dari 100 review negatif dan 100 review positif. Berikut hasil akurasi dari 3 algoritma yang digunakan dalam pengujian dataset halodoc.

1. Hasil dari metode Algoritma K-Nearest Neighbor (k-NN).

Tabel 4. Hasil Akurasi Algoritma KNN

Accuracy : 88.50%, AUC: 0.960			
	true Data_Positif	true Data_Negatif	class precision
pred. Data_Positif	94	17	84.68%
pred. Data_Negatif	6	83	93.26%
class recall	94.00%	83.00%	

Nilai Accuracy Algoritma K-NN ialah:

$$\text{Accuracy} = \frac{(TN+TP)}{(TN+FN+TP+FP)}$$

$$\text{Accuracy} = \frac{83+94}{83+17+94+6}$$

$$\text{Accuracy} = \frac{177}{200}$$

Accuracy = 0.885 = 88.50%

2. Hasil dari metode Algoritma C4.5

Tabel 5. Hasil Akurasi Algoritma C4.5

Accuracy : 74.00%, AUC: 0.723

	true Data_Positif	true Data_Negatif	class precision
pred. Data_Positif	65	17	79.27%
pred. Data_Negatif	35	83	70.34%
class recall	65.00%	83.00%	

Nilai Accuracy Algoritma C4.5 ialah:

$$\text{Accuracy} = \frac{(TN+TP)}{(TN+FN+TP+FP)}$$

$$Accuracy = \frac{83 + 65}{83 + 17 + 65 + 35}$$

$$Accuracy = \frac{148}{200} = 0.74 = 74.00\%$$

3. Hasil dari metode Algoritma Support Vektor Machine

Tabel 6. Hasil Akurasi Algoritma Support Vektor Machine

Accuracy : 87.00%, AUC: 0.972			
	true Data_Positif	true Data_Negatif	class precision
pred. Data_Positif	97	23	80.83%
pred. Data_Negatif	3	77	96.25%
class recall	97.00%	77.00%	

Nilai Accuracy Algoritma Support Vektor Machine ialah:

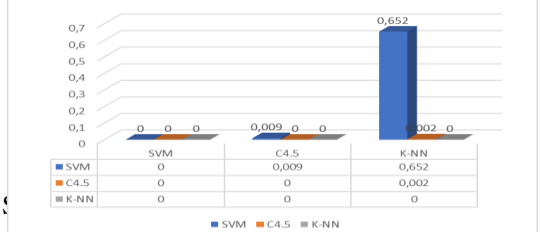
$$\text{Accuracy} = \frac{(TN+TP)}{(TN+FN+TP+FP)}$$

$$Accuracy = \frac{77 + 97}{77 + 23 + 97 + 3}$$

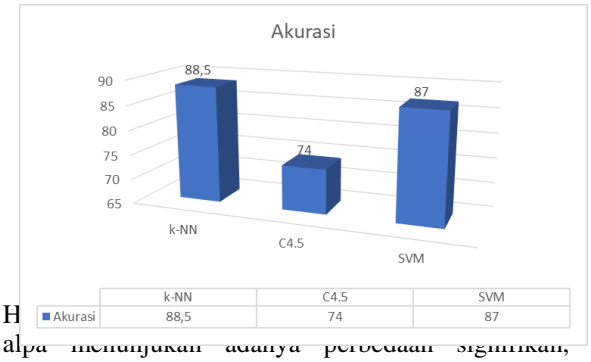
$$Accuracy = \frac{174}{200}$$

Accuracy = 0.87 = 87.00%

Dari hasil penelitian ini, dengan menggunakan aplikasi rapid miner, menunjukan algoritma K-NN dengan akurasi terbaik yang mendapatkan nilai akurasi sebesar 88.50%. Algoritma Support Vektor Machine dengan nilai akurasi 87.00% sedangkan algoritma C.45 mendapatkan nilai akurasi 74.00%.



Selanjutnya 3 algoritma tersebut, peneliti dengan langkah selanjutnya melakukan pengujian dengan menggunakan uji t-tes. Pada pengujian ini, nilai dari uji t-test ini memiliki nilai yang sigifikan berbeda pada tiap algoritma. Nilai alpa 0.05, jika probabilitasnya >0.005 maka H_0 diterima (tidak ada perbedaan signifikan), kemudian jika probabilitasnya <0.05 makan H_0 ditolak (adanya perbedaan signifikan). Berikut grafik dari hasil uji t-tes dari 3 algoritma yang digunakan dalam klafisikasi.



Hasil alpa menunjukan adanya perbedaan signifikan, algoritma tersebut adalah algoritma Support Vektor Machine dengan algoritma C4.5, algoritma C4.5 dengan algoritma K-NN. Sedangkan untuk algoritma Support Vektor Machine dengan algoritma K-NN menunjukan hasil lebih besar dari nilai alpa sehingga tidak ada perbedaan signifikan. Sehingga, model terbaik dari daset halodoc ini ialah algoritma Support Vektor Machine dan algoritma K-NN

KESIMPULAN

Dari hasil penelitian yang sudah dilakukan dengan menggunakan dataset halodoc dapat disimpulkan, Peneliti menggunakan tiga Algoritma C4.5, K-NN dan SVM. Hasil penelitian dari berbagai pengujian menghasilkan nilai akurasi terbaik sebesar 88.50% adalah algoritma K-NN. Kemudian pengujian menggunakan uji t-tes berdasarkan urutan algoritma menghasilkan model algoritma yang terbaik ialah SVM dan K-NN sedangkan C4.5 algoritma yang kurang baik digunakan dalam pengujian dataset halodoc.

REFERENSI

Ipmawati, J., Kusrini, K., & Luthfi, E. T. (2017). Komparasi Teknik Klasifikasi Teks Mining Pada Analisis Sentimen. *Indonesian Journal on Networking and Security*, *6*(1), 28–36.

JianQiang, Z., & Xiaolin, G. (2017). *Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis* (Vol. 5). IEEE Access,. doi:10.1109/ACCESS.2017.2672677

Muthia, D. A. (2017). ANALISIS SENTIMEN

PADA REVIEW RESTORAN DENGAN TEKS BAHASA INDONESIA MENGUNAKAN. *JURNALILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER*, *2*(2), 39–45.

Putra, P. A., & Suryanata, I. G. N. P. (2021). SINERGI HALODOC DALAM MUTU PELAYANAN RUMAH SAKIT DI MASA PANDEMI COVID 19. *E-JURNAL EKONOMI DAN BISNIS*, *10*(04), 211–222.

Taufik, A. (2018). Komparasi Algoritma Text Mining Untuk Klasifikasi Review Hotel. *Jurnal Teknik Komputer*, *IV*(2). https://doi.org/10.31294/jtk.v4i2.3461

Rahayuningsih, P. A. (2019, Mei). Komparasi Algoritma Klasifikasi Data Mining untuk. *Journal of Information Technology and Computer Science (JOINTECS), 4*(2), 63-68.

Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). *A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis* (Vol. 110). Expert Syst. Appl. doi:10.1016/j.eswa.2018.06.022

Utami, L. A. (2017). ANALISIS SENTIMEN OPINI PUBLIK BERITA KEBAKARAN HUTAN MELALUI KOMPARASI ALGORITMA SUPPORT VECTOR MACHINE DAN K-NEAREST NEIGHBOR BERBASIS PARTICLE SWARM OPTIMIZATION. *Jurnal PILAR Nusa Mandiri*, *13*(1), 103–112.

Utami, L. D., Rachmi, H., Nurlaela, D., Akuntansi, S. I., Bina, U., Informatika, S., … Komputer, I. (2018). KOMPARASI ALGORITMA KLASIFIKASI PADA ANALISIS REVIEW HOTEL. *Jurnal PILAR Nusa Mandiri*, *14*(2), 261–266.

Wibawa, A. P., Guntur, M., Purnama, A., Akbar, M. F., & Dwiyanto, F. A. (2018). Metode-metode Klasifikasi. *Prosiding Seminar Ilmu Komputer Dan Teknologi Informasi*, *3*(1), 134–138.