Jurnal Ilmu Komputer dan Informatika (JIKI)
Vol. 3, No. 2, Desember 2023, Hal. 99-106
https://jiki.jurnal-id.com

DOI: https://doi.org/10.54082/jiki.65
P-ISSN: 2807-6664
E-ISSN: 2807-6591

Analysis Sentiment Terhadap Ginjal Akut pada Twitter Menggunakan Algoritma Random Forest

Hadi Mirojul Falah[*1], Muhammad Ridwan Jamil[2], Ahmad Taufik[3], Marten Botha[4], Nova Agustina[5]

[1,2,3,4,5]Informatika, Fakultas Teknologi Informasi, Universitas Adhirajasa Reswara Sanjaya, Indonesia
Email: [1]hadimirojulf41@gmail.com, [2]zamzamridwan06@gmail.com, [3]ahmdtfk7@gmail.com, [4]marthenbota34@gmail.com, [5]nova@sttbandung.ac.id

Abstrak

Analisis sentimen berkaitan dengan identifikasi dan klasifikasi pendapat atau sentimen yang diungkapkan dalam teks sumber. Sosial media menghasilkan sejumlah besar data yang penuh dengan sentimen berupa tweet, update status, postingan blog maupun yang lainnya. Analisis sentimen dari data yang dihasilkan pengguna ini sangat berguna dalam mengetahui pendapat dari kerumunan. Dalam makalah ini, dilakukan analisis postingan twitter tentang Ginjal Akut menggunakan pendekatan *Machine Learning*, yaitu dengan menggunakan Algoritma Random Forest dan Particle Swarm Optimization. Hasil penelitian menunjukkan akurasi yang di dapat pada saat melakukan klasifikasi menggunakan algoritma *Random Forest* dengan *Cross Validation* sebesar 94.47%.

Kata kunci: Cross Validation, Klasifikasi, Machine Learning, Particle Swarm Optimization, Random Forest, Twitter

Abstract

Sentiment analysis is concerned with the identification and classification of opinions or sentiments expressed in source texts. Social media produces large amounts of data full of sentiment in the form of tweets, status updates, blog posts and others. Sentiment analysis of user-generated data is very useful in knowing the opinions of the crowd. In this paper, an analysis of Twitter posts about Acute Kidney is carried out using a Machine Learning approach, namely using the Random Forest Algorithm and Particle Swarm Optimization. The research results show that the accuracy obtained when classifying using the Random Forest algorithm with Cross Validation was 94.47%.

Keywords: Cross Validation, Klasifikasi, Machine Learning, Particle Swarm Optimization, Random Forest, Twitter

1. PENDAHULUAN

Dengan pesatnya perkembangan media sosial, semakin banyak orang yang menuliskan pendapatnya tentang suatu hal. Karena hal ini mendorong penelitian untuk menggunakan media sosial Twitter sebagai sumber data yang akan dianalisis berupa sentiment Analisis, yaitu proses memahami, mengekstraksi, dan mengolah data tekstual secara otomatis untuk memperoleh makna informasi yang terkandung dalam opini tersebut[1], [2]. Analisis sentimen (disebut juga Opinion mining) menggunakan komputasi linguistik dan pemrosesan bahasa alami model untuk memahami teks[3], [4].

Machine Learning adalah salah satu yang paling cepat berkembang bidang ilmu komputer, dengan jangkauan sangat yang jauh . Hal ini mengacu pada deteksi otomatis pola yang berarti dalam data. Machine Learning merupakan kemampuan untuk belajar dan beradaptasi[5], [6] .

Studi kasus yang diambil oleh peneliti adalah kasus Ginjal Akut. Mengambil studi kasus Ginjal Akut untuk analisis sentimen karena permasalahan ini sudah banyak dibicarakan banyak oleh masyarakat Indonesia di media sosial. Berdasarkan fakta tersebut, dapat disimpulkan bahwa Ginjal Akut menarik minat banyak orang untuk dibahas. Hal ini dibuktikan dengan banyaknya komentar atau tweet di media sosial khususnya Twitter oleh masyarakat Indonesia terhadap Ginjal Akut. Berdasarkan

DOI: https://doi.org/10.54082/jiki.65
P-ISSN: 2807-6664
E-ISSN: 2807-6591

komentar di Twitter, sentiment analisis dapat dilakukan. Salah satu cara untuk melakukan analisis sentimen dapat dilakukan dengan menggunakan data dari media sosial. Ribuan pengiriman terjadi setiap hari di setiap media sosial. Setiap orang dapat mengutarakan pendapatnya melalui media sosial dengan bebas. Pendapat tersebut mengandung sentimen positif, negatif dan netral pada suatu topik. Sentimen positif mengungkapkan pendapat yang baik tentang suatu konteks, negative sentimen mengungkapkan pendapat buruk dalam suatu konteks, sedangkan sentimen netral menyatakan hal-hal yang tidak mendukung baik atau buruk[1].

Random Forest terdiri dari beberapa pohon keputusan, Setiap pohon keputusan tumbuh sepenuhnya, tidak diperlukan pemangkasan semakin banyak pohon, semakin akurat hasilnya menjadi, dan itu tidak cocok dengan baik. Algoritme hutan acak melakukan ini membuat penilaian umum dan memiliki keuntungan otomatis pemilihan fungsi dll. Jadi kita memiliki masalah utama untuk di pecahkan[7]

Algoritma yang digunakan dalam melakukan analisis sentimen pada penelitian ini adalah Random Forest.Tujuan dari penelitian ini hanya untuk mengetahui akurasi pada sebuah algoritma Klasifikasi. Hal ini berpengaruh atau tidak berpengaruh pada penelitian terdahulu.

2. METODE PENELITIAN

2.1. Identifikasi Masalah

Masyarakat menggunakan media sosial sebagai sarana untuk mengekspresikan pikiran, minat, dan pendapat mereka tentang berbagai hal. Ribuan kiriman terjadi setiap hari di setiap media sosial. Setiap orang dapat mengutarakan pendapatnya melalui media sosial dengan bebas. Studi kasus yang diambil peneliti adalah Gagal Ginjal di Indonesia. Kasus tersebut diambil karena Gagal Ginjal ramai diperbincangkan oleh masyarakat Indonesia di media sosial Twitter. Jika ingin mengetahui kecenderungan komentar masyarakat terhadap Gagal Ginjal di Indonesia apakah banyak berita fakta atau berita hoax kemudian dilakukan analisis sentimen. Sentiment Analysis (SA) atau Opinion Mining (OM) adalah kom-studi dugaan tentang pendapat, sikap, dan emosi orang menuju entitas. Entitas dapat mewakili individu, peristiwa atau topik. Topik-topik ini kemungkinan besar akan dibahas oleh ulasan. Dua ekspresi SA atau OM dapat dipertukarkan. Mereka mengekspresikan makna bersama. Namun, beberapa peneliti menyatakan bahwa OM dan SA memiliki pengertian yang sedikit berbeda[8], [9].
Opinion Mining mengekstraksi dan menganalisis pendapat orang tentang entitas sementara Analisis Sentimen mengidentifikasi sentimen diungkapkan dalam sebuah teks kemudian menganalisisnya. Oleh karena itu, sasaran dari SA adalah menemukan pendapat, mengidentifikasi sentimen yang mereka ungkapkan, dan kemudian mengklasifikasikan polaritas[8]

2.2. Dataset

Kumpulan data tidak tersedia untuk umum, format dari data ini yaitu CSV (Comma Separated Values). Seperti yang telah disebutkan sebelumnya, record tersebut memiliki 800 baris beberapa atribut seperti id, status fakta, judul berita, nama dan nim. Kumpulan data mentah berisi metadata dan entri yang tidak lengkap atau hilang disaring preprocessing. Penelitian ini sering membahas berbagai sumber daya digunakan dalam Sentiment Twiter Analysis. Kami meringkas berbagai pendekatan yang diusulkan untuk konstruksi leksikon sentimen untuk Twitter, sajikan dan jelaskan secara singkat evaluasi yang tersedia kumpulan data, jelaskan proses pembuatan kumpulan data Twitter, dan rangkum perbedaan proses diterapkan untuk membubuhi keterangan dataset yang ada[10]. Preprocessing meliputi penentuan median nilai yang tersedia menjadi nilai yang hilang dan mengonversi nilai string menjadi numerik. Misalnya, mengubah jenis kelamin seseorang numerik; Tetapkan 0 untuk pria dan 1 untuk wanita. Selanjutnya, dataset dibagi menjadi set pengujian dan pelatihan untuk memprediksi seberapa baik kinerja algoritme[11].

DOI: https://doi.org/10.54082/jiki.65
P-ISSN: 2807-6664
E-ISSN: 2807-6591

2.3. Preprocessing

Setiap data set akan selalu memiliki noise, redundansi, dan data yang tidak relevan yang harus dihapus. Preprocessing data bisa memiliki dampak signifikan pada kinerja generalisasi algoritma Machine Learning. Kami menyajikan algoritme yang paling tinggi akurasinya untuk setiap langkah preprocessing data sehingga mencapai yang terbaik kinerja untuk sebuah Data set[12]. Itu tidak berhasil membuat fase pelatihan sangat sulit dan mengarah pada hasil yang buruk kinerja pengklasifikasi. Ini adalah langkah penting mencapai akurasi tinggi dari pengklasifikasi apa pun karena membantu training dengan data sebaik mungkin. Ada beberapa teknik pengolahan dan pembersihan data. Kami mengambil beberapa langkah untuk menyelesaikan proses pendahuluan standar dengan menghapus karakter khusus yang biasanya muncul di tweet, seperti '@, #, \" dll., hapus spasi tambahan dan tambahan kata-kata dan banyak langkah seperti itu. Apabila menggunakan sebuah algoritma, dapat dijelaskan di bagian ini, beserta dengan *state of the art*[5].

2.4. Validasi Data

Teknik klasifikasi seperti Cross-validation [13], [14], split-validasi yaitu bentuk lain dari validasi silang dan validasi silang validasi dengan pohon keputusan telah diterapkan pada sekumpulan data yang berisi catatan individu dari kelompok umur yang berbeda dengan nilai semua faktor yang mungkin menyebabkan sebuah data set dan atas dasar algoritma ini prediksi dilakukan. Cross-validation adalah teknik untuk mengevaluasi bagaimana hasil perhitungan statistik menjadi data umum mengatur. Jenis teknik ini diterapkan ketika tujuan utama model tertentu adalah prediksi dan keakuratannya perlu diperhatikan dihitung[14].Dalam metode validasi data peneliti menggunakan beberapa fitur validasi yaitu Split validation 0,5 – 0,9 dan Cross validation X fold. Untuk mencari nilai akurasi yang paling tinggi di dapat. Split validation dan Cross validation di uji menggunakan tools Rapid Miner.

2.5. Random Forest

Hutan Acak terdiri dari beberapa pohon keputusan,setiap pohon keputusan akan tumbuh penuh, tidak perlu dipotong pemrosesan, semakin banyak pohon yang dimilikinya, semakin akurat hasilnya menjadi, dan itu tidak akan terlalu pas. Algoritme hutan acak akan lakukan perkiraan keseluruhan, dan ini memiliki keunggulan otomatis pemilihan fitur dll. Jadi kami memiliki masalah utama berikut untuk dipecahkan.[15] Definisi ini menunjukkan Random Forest adalah kombinasi dari banyak pengklasifikasi struktur pohon. di dalam Model RF Breiman, setiap pohon ditanam berdasarkan kumpulan sampel pelatihan dan variabel acak[16]

2.6. Optimize Feature Selection (PSO)

PSO adalah teknik EC yang diusulkan oleh Kennedy dan Eberhart pada tahun 1995. PSO dilatarbelakangi oleh perilaku sosial seperti kawanan burung dan kawanan ikan. Fenomena yang mendasari dari PSO adalah bahwa pengetahuan dioptimalkan oleh interaksi social dalam populasi di mana pemikiran tidak hanya bersifat pribadi tetapi juga sosial.[17]–[19] Berdasarkan pbest dan gbest, PSO mencari solusi optimal dengan memperbarui kecepatan dan posisi masing-masing partikel menurut persamaan berikut:[19]

$$x_{id}^{t+1} = x_{id}^{t} + x_{id}^{t+1} \quad (1)$$

$$x_{id}^{t+1} = w * v_{id}^{t} + c_1 * r_1 * \left(p_{id} - x_{id}^{t}\right) + c_2 * r_2 * \left(p_{gd} - x_{id}^{t}\right) \quad (2)$$

2.7. Evaluasi

Pada tahap evaluasi dilakukan setelah data diolah dengan Algoritma Random Forest menghasilkan negative banyak berita hoax dan fakta. Untuk menentukan akurasi algoritma, evaluasi akurasi, presisi, dan recall dilakukan untuk menghasilkan kesimpulan kinerja dari algoritma Random Forest untuk

DOI: https://doi.org/10.54082/jiki.65
P-ISSN: 2807-6664
E-ISSN: 2807-6591

dianalisis negative pada pengguna di Twitter. Selanjutnya melakukan analisis dan terakhir menyimpulkan hasil analisis sentiment yang telah dilakukan dengan algoritma Random Forest.[1]

Dalam klasifikasi, selalu ada sesuatu yang disebut False Statement. Misalnya, sebuah kalimat dinyatakan negative genap meskipun kalimat tersebut adalah kalimat positif atau kalimat dinyatakan netral meskipun kalimat tersebut adalah kalimat positif/negative, serta kalimat positif dinyatakan negatif (Lihat Tabel 1). Untuk menghindari ini kita harus melakukan evaluasi kinerja untuk mengetahui Pernyataan Palsu.[1], [20]

Tabel 1. Klasifikasi Sentimen

Classification	***Positive***	***Negative***
Positive	*True Positive (TP)*	*False Negative (FN)*
Negative	*False Positive (FP)*	*True Negative (TN)*

Berapa banyak keberhasilan algoritma yang digunakan, perlu untuk memeriksa akurasi seperti yang terlihat pada Persamaan. 1:

$$Accuracy = \frac{TP+T}{TP+FN+FP+T} \quad (3)$$

Presisi adalah rasio item yang relevan yang dipilih untuk semua item yang dipilih. Presisi dapat diartikan sebagai kecocokan antara permintaan informasi dan jawaban permintaan (Lihat Persamaan, 2).

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Penarikan kembali didefinisikan sebagai rasio item relevan yang dipilih dengan jumlah total item relevan yang tersedia. Penarikan adalah dihitung dengan Persamaan. 3[1], [21] :

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

3. HASIL DAN PEMBAHASAN

Dataset dibuat menggunakan posting twitter Ginjal Akut. Tweet adalah pesan singkat yang penuh dengan kata-kata gaul dan salah eja. Jadi kami melakukan analisis sentimen tingkat kalimat. Ini dilakukan dalam tiga fase. Pada tahap pertama preprocessing, Kemudian Optimize Featur Selection menggunakan algoritma PSO. Akhirnya menggunakan pengklasifikasi yang berbeda, tweet diklasifikasikan menjadi kelas fakta dan hoax. Berdasarkan jumlah tweet di setiap kelas, sentimen terakhir diturunkan. Algoritma yang di gunakan dalam penelitian ini adalah Random Forest. Ketika model dasar diimplementasikan dengan split validation 0,5 - 0,9 dan cross validation x dengan hasil akurasi yang berbeda. Karena perbedaan akurasi pada saat pengujian algoritma Random Forest dengan menggunakan fitur validai antara split validation 0,5 - 0,9 dan cross validation. Akurasi yang di dapat pada saat melakukan penelitian dalam data set Ginjal Akut dengan menggunakan Random Forest di dapatkan nilai akurasi paling tinggi yaitu 94.97% dengan AUC 0.425 memakai fitur split validation 0,8. Nilai akurasi paling rendah terdapat pada pengujian menggunakan split validation 0,5 sebesar 94.34% dengan AUC 0.432. Tujuan melakukan data percobaan ini hanya untuk mengetahui perfoma sebuah algoritma yaitu Random Forest. Setelah melakukan percobaan tersebut peneliti melakukan pengoptimasian pada data set Ginjal Akut menggunakan Random Forest hanya untuk mendapat kenaikan akurasi pada saat di optimasi. Algoritma optimasi yang di pakai dengan tujuan agar menaikan akurasi pada data set Ginjal Akut memakai 2 algoritma optimasi. Algoritma optimasi yang dimaksud yaitu Optimize Selection(Evolutionary) dan Optimize Weighting(PSO). Sedangkan itu pengujian di lakukan menggunakan algoritma Optimize Feature Selection(Evolutionary)+Cross Validation+Random Forest ditemukan hasil akurasi menjadi 94.85% dengan AUC 0.549, dari yang sebelumnya yaitu 94.47% dengan AUC 0.404. Hasil yang mencengangkan pada saat peneliti melakukan pengujian terhadap

DOI: https://doi.org/10.54082/jiki.65
P-ISSN: 2807-6664
E-ISSN: 2807-6591

algoritma Optimize Feature Weighting(PSO)+Cross Validation+Random Forest nilai akurasi yang di dapat 94.85% dengan AUC 0.549.Tidak ada perubahan nilai akurasi pada saat menggunakan algoritma optimasi yang berbeda melainkan sama nilai akurasi meskipun algorita optimasi berbeda. Teknik optimasi ini harus di lakukan pada data set Ginjal Akut dengan tujuan agar meningkatkan nilai akurasi pada saat di optimasi, meskipun naiknya akurasi pada saat di optimasi tidak terlalu meningkat, tetapi itu juga mempengaruhi.

Tabel 2. Nilai Akurasi Random Forest

Algoritma	Validasi	Akurasi	AUC
Random Forest	Cross Valdation X	94.60%	0.404
	Split Validation 0,5	92.21%	0.511
	Split Validation 0,6	94.34%	0.432
	Split Validation 0,7	94.54%	0.481
	Split Validation 0,8	94.97%	0.425
	Split Validation 0,9	94.94%	0.367

3.1. Pengujian Klasifikasi Random Forest dengan Split Validation

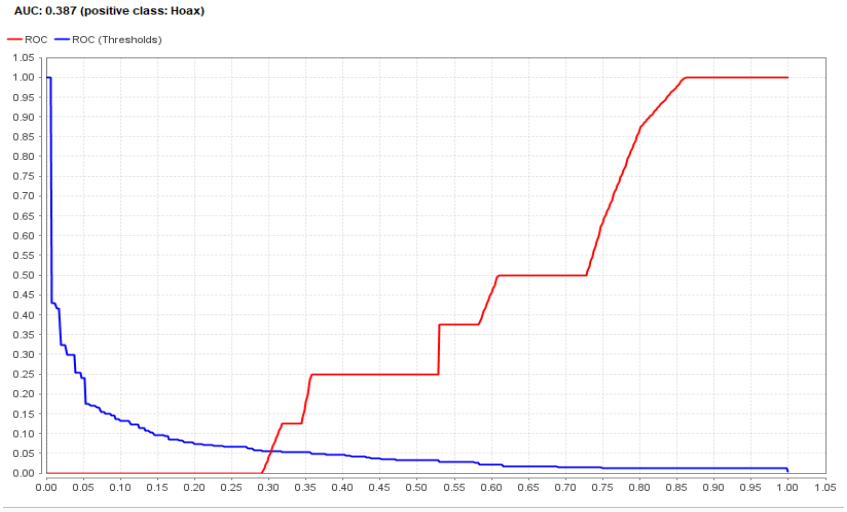
Dalam penelitian ini, seluruh tahapan uji coba algoritma klasifikasi pada dataset yang telah diolah menggunakan tools RapidMiner, pada tahap uji coba menggunakan klasifikasi Random Forest dengan metode Split Validation serta nilai 0,8 sebagai parameter, pada tahap ini hasil akurasi sebagai berikut:

accuracy: 94.97%

	true Fakta	true Hoax	class precision
pred. Fakta	151	8	94.97%
pred. Hoax	0	0	0.00%
class recall	100.00%	0.00%	

Gambar 1. Hasil Akurasi Split Validation

Pada gambar diatas dapat diketahui bahwa pada tahap pengujian menggunakan algoritma Random Forest dengan Split Validation 0,8 menghasilkan akurasi paling tinggi dengan jumlah 94,97% serta nilai precision nya dari masing-masing prediction true dan prediction false bernilai 100,00% dan 0,00% adapun recall dari masing-masing prediction bernilai 0,00% dan 100,00%. Adapun distribusi AUC (Area Under Curve) dari pengujian ini dapat dilihat sebagai berikut:



Gambar 2. Nilai AUC

DOI: https://doi.org/10.54082/jiki.65
P-ISSN: 2807-6664
E-ISSN: 2807-6591

Pada hasil uji coba diatas menunjukan bahwa grafik ROC dengan nllai AUC (Area Under Curve) dengan menggunakan model klasifikasi Naive Bayes serta metode Cross Validaion sebesar 0,387.

3.2. Pengujian Klasifikasi Random Forest dengan Cross Validation

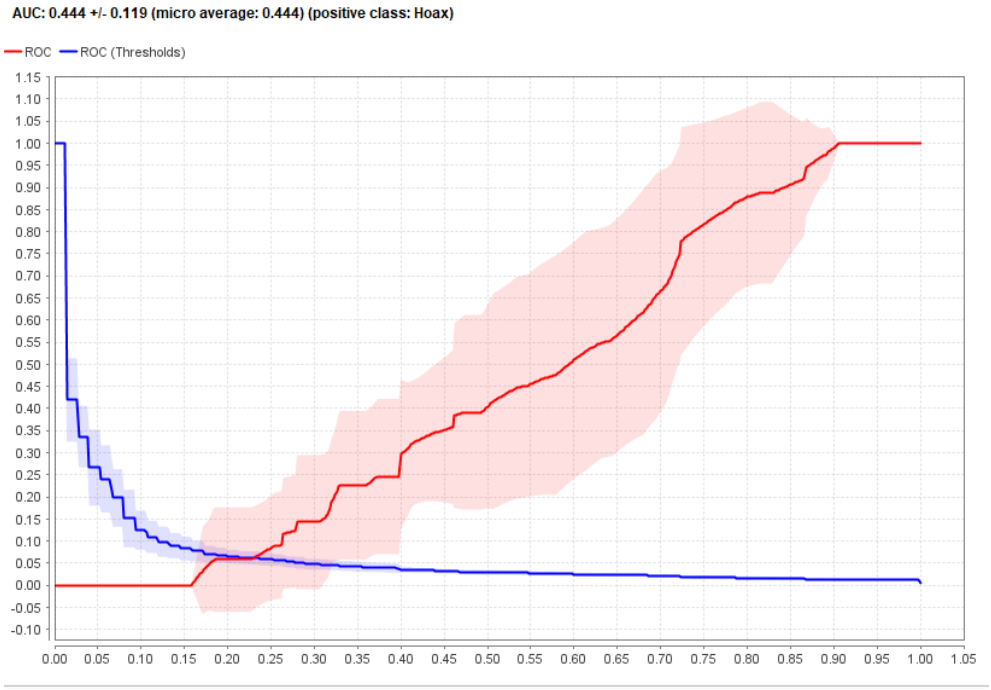
Selanjutnya uji coba dengan menggunakan model algoritma Random Forest. Pada tahap ini dataset diuji menggunakan metode Cross Validation dengan nilai X-Fold yang digunakan ialah X-Fold dengan hasil akurasi sebagai berikut:

accuracy: 94.60% +/- 0.84% (micro average: 94.60%)

	true Fakta	true Hoax	class precision
pred. Fakta	753	41	94.84%
pred. Hoax	2	0	0.00%
class recall	99.74%	0.00%	

Gambar 3. Hasil Akurasi Cross Validation

Pada gambar tersebut dilihat bahwa hasil dari uji coba menggunakan model algoritma Random Forest menghasilkan nilai akurasi rendah bila dibandingkan dengan uji coba sebelumnya. Pada pengujian menggunakan model algoritma Random Forest ini jumlah akurasi yang dihasilkan bernilai 94.60%. Dengan nilai precision nya dari masing-masing prediction true dan prediction false bernilai 94,84% dan 0,00% adapun recall dari masing-masing prediction bernilai 99,74% dan 0,00%. Adapun distribusi AUC (Area Under Curve) dari pengujian ini dapat dilihat sebagai berikut:



Gambar 4. Hasil AUC

Dari hasil uji coba tersebut menunjukan bahwa grafik ROC dengan nllai AUC (Area Under Curve) dengan menggunakan model klasifikasi Random Forest serta metode Cross Validaion sebesar 0,444.

DOI: https://doi.org/10.54082/jiki.65
P-ISSN: 2807-6664
E-ISSN: 2807-6591

4. KESIMPULAN

Penelitian ini telah berhasil melakukan pengukuran tingkat akurasi dari algoritma Random Forest pada dataset "Ginjal Akut" dengan menerapkan fitur validasi yaitu Cross Validation dengan tingkat akurasi sebesar 94,47% dengan AUC 0.404. Makalah ini membuat kontribusi untuk bidang data sains dan analisis sentimen. Dalam makalah ini kami membandingkan berbagai teknik klasifikasi dan membandingkan akurasi. Dalam domain analisis sentimen untuk Analisis Sentimen sangat banyak penelitian yang telah dilakukan. Pada penelitian sebelumnya telah melakukan analisis tingkat kata dari tweet tanpa mempertahankan urutan kata. Namun dalam penelitian ini kami telah melakukan klasifikasi dan optimasi untuk mengetahui akurasi algoritma Random Forest.

Berdasarkan data yang diperoleh dari twitter tentang Ginjal Akut kecenderungan komentar menjadi disampaikan berisi berita fakta. Pengguna Twitter tahu tentang masalah ini. Dari 796 data pengujian, terdapat 755 data yang berisikan berita/postingan fakta, kemudian 41 postingan/berita hoax. Disini dapat disimpulkan bahwa pengguna media social di Indonesia mengetahui kasus Ginjal Akut. Berdasarkan hasil akurasi dari Algoritma Random Forest sebesar 94,47% dengan AUC 0.404

Ada berbagai teknik Pembelajaran Simbolik dan Mesin untuk mengidentifikasi sentimen dari teks. Pembelajaran mesin teknik lebih sederhana dan efisien daripada teknik simbolik. Teknik-teknik ini dapat diterapkan pada analisis sentimen twitter. Machine Learning adalah mesin yang dikembangkan untuk bisa belajar dengan sendirinya tanpa arahan dari penggunanya. Pembelajaran mesin dikembangkan berdasarkan disiplin ilmu lainnya seperti statistika, matematika dan data mining sehingga mesin dapat belajar dengan menganalisa data tanpa perlu di program ulang atau diperintah.

DAFTAR PUSTAKA

[1] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm," *Procedia Comput Sci*, vol. 161, pp. 765–772, 2019.

[2] I. F. Rozi, S. H. Pramono, and E. A. Dahlan, "Implementasi opinion mining (analisis sentimen) untuk ekstraksi data opini publik pada perguruan tinggi," *Jurnal EECCIS (Electrics, Electronics, Communications, Controls, Informatics, Systems)*, vol. 6, no. 1, pp. 37–43, 2012.

[3] A. Nayak and D. Natarajan, "Comparative study of naive Bayes, support vector machine and random forest classifiers in sentiment analysis of twitter feeds," *International Journal of Advance Studies in Computer Science and Engineering (IJASCSE)*, vol. 5, no. 1, p. 16, 2016.

[4] R. Weischedel *et al.*, "White paper on natural language processing," in *Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod, Massachusetts, October 15-18, 1989*, 1989.

[5] A. Nayak and D. Natarajan, "Comparative study of naive Bayes, support vector machine and random forest classifiers in sentiment analysis of twitter feeds," *International Journal of Advance Studies in Computer Science and Engineering (IJASCSE)*, vol. 5, no. 1, p. 16, 2016.

[6] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[7] W. Lin, Z. Wu, L. Lin, A. Wen, and J. Li, "An ensemble random forest algorithm for insurance big data analysis," *Ieee access*, vol. 5, pp. 16568–16575, 2017.

[8] M. Tsytsarau and T. Palpanas, "Survey on mining subjective data on the web," *Data Min Knowl Discov*, vol. 24, pp. 478–514, 2012.

[9] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[10] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. J. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the workshop on language in social media (LSM 2011)*, 2011, pp. 30–38.

DOI: https://doi.org/10.54082/jiki.65
P-ISSN: 2807-6664
E-ISSN: 2807-6591

[11] K. Singh, R. Nagpal, and R. Sehgal, "Exploratory data analysis and machine learning on titanic disaster dataset," in *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2020, pp. 320–326.

[12] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," *Int J Comp Sci*, vol. 1, no. 2, pp. 111–117, 2006.

[13] K. Choudhary and S. Wadhwa, "Glaucoma detection using cross validation algorithm," in *2014 Fourth International Conference on Advanced Computing & Communication Technologies*, 2014, pp. 478–482.

[14] K. Chaudhary, P. Maheshwari, and S. Wadhwa, "Glaucoma detection using cross validation algorithm: A comparitive evaluation on rapidminer," in *2014 IEEE Conference on Norbert Wiener in the 21st Century (21CW)*, 2014, pp. 1–5.

[15] W. Lin, Z. Wu, L. Lin, A. Wen, and J. Li, "An ensemble random forest algorithm for insurance big data analysis," *Ieee access*, vol. 5, pp. 16568–16575, 2017.

[16] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: Random forest," in *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3*, 2012, pp. 246–252.

[17] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *1998 IEEE international conference on evolutionary computation proceedings. IEEE world congress on computational intelligence (Cat. No. 98TH8360)*, 1998, pp. 69–73.

[18] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization," *Swarm intelligence*, vol. 1, no. 1, pp. 33–57, 2007.

[19] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE Trans Cybern*, vol. 43, no. 6, pp. 1656–1671, 2012.

[20] W. Richert, *Building machine learning systems with Python*. Packt Publishing Ltd, 2013.

[21] J. Ling, I. P. E. N. Kencana, and T. B. Oka, "Analisis Sentimen Menggunakan Metode Naïve Bayes Classifier Dengan Seleksi Fitur Chi Square," *E-Jurnal Matematika*, vol. 3, no. 3, pp. 92–99, 2014.