IJCIT (Indonesian Journal on Computer and Information Technology)
Vol.3, No.2, November 2018, pp. 216~221
p-ISSN: 2527-449X, e-ISSN**:** 2549-7421

Diterima Agustus 23, 2018; Revisi Oktober 30 , 2018; Disetujui November 2, 2018

Email Spam Filtering Dengan Algoritma Random Forest

Muhamad Abdul Ghani[1], Agus Subekti[2]

[1]STMIK Nusa Mandiri Jakarta
Ganiabdul691@gmail.com

[2]Lembaga Ilmu Pengetahuan Indonesia
agus.subekti@lipi.go.id

Abstrak

Teknologi berbasis internet sudah menjadi kebutuhan primer. Berdasarkan hasil survey Badan Pusat Statistik bekerjasam dengan APJII, kegiatan pengiriman dan penerimaan email sudah mengalahkan posisi media sosial dengan mencapai 95.75%. Penggunaan email yang sangat intens dapat menimbulkan dampak positif dan negatif. Karena selain selain sebagai alat komunikasi, pada kenyataannya tidak semua orang menggunakan email dengan baik dan bahkan ada banyak sekali penyalahgunaan email sehingga berpotensi untuk merugikan orang lain. *Email* yang disalahgunakan ini biasa dikenal sebagai spam atau *junkmail* (email sampah) yang mana email tersebut berisikan iklan, penipuan dan bahkan virus. Dalam penelitian ini dilakukan perbandingan beberapa metode klasifikasi data mining diantaranya yaitu Algoritma *Naïve Bayes, SVM, J48,* dan *Random Forest* dalam memprediksi spam email dengan tujuan agar algoritma terpilih merupakan yang paling akurat. Dari hasil pengujian menggunakan dengan mengukur kinerja dari keempat algoritma tersebut menggunakan *Confusion Matrix* dan *ROC ,* diketahui bahwa algoritma *Random Forest* memiliki nilai accuracy paling tinggi, yaitu 94,22 % dan AUC 0,98 diikuti oleh algoritma J48 dengan accuracy sebesar 92,70% dan AUC 0,95, SVM dengan nilai accuracy 86,48% dan AUC 0,84 dan terendah yaitu metode naive bayes dengan nilai accuracy sebesar 78,87% dan AUC 0,79.
Kata kunci: algoritma naive bayes, email spam, J48, random forest, support vector machine

Abstract

Internet-based technology has become a primary need. Based on the results of a survey by the Central Bureau of Statistics in cooperation with APJII, email sending and receiving activities have outperformed the position of social media by reaching 95.75%. The use of e-mail that is very intense can have positive and negative impacts. Because other than as a means of communication, in reality not everyone uses email well and there is even a lot of email abuse that has the potential to harm others. This misused email is commonly known as spam or junkmail (junk e-mail) which contains e-mail, fraud and even viruses. In this study a comparison of several data mining classification methods including the Naïve Bayes, SVM, J48, and Random Forest algorithms in predicting spam e-mail with the aim that the selected algorithm is the most accurate. From the test results using measuring the performance of the four algorithms using Confusion Matrix and ROC, it is known that the Random Forest algorithm has the highest accuracy value, which is 94.22% and AUC 0.98 followed by the J48 algorithm with accuracy of 92.70% and AUC 0.95, SVM with 86.48% accuracy value and 0.84 AUC and the lowest is the naive bayes method with accuracy value of 78.87% and AUC 0.79.
Keyword*: J48, naive bayes algorithm, random forest, spam email, support vector machine*

1. Pendahuluan

Saat ini teknologi berbasis internet sudah menjadi kebutuhan primer. Menurut laporan terbaru Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), lebih dari 50% atau sekitar 143 juta orang penduduk Indonesia telah terhubung jaringan internet sepanjang 2017 (Bohang, 2018). Dalam penggunaan internet di Indonesia, surat elektronik (*email*) mengalahkan posisi media sosial (*social media*), hal tersebut berdasarkan hasil survei Badan Pusat Statistik bekerjasama dengan APJII, dengan hasil survey kegiatan pengiriman dan penerimaan *e-mail* mencapai 95,75%, sedangkan akses layanan media sosial mencapai 61,23%. (Sinaga, 2014). Penggunaan email yang sangat intens ini

menimbulkan dampak positif dan negatif. Karena selain sebagai alat komunikasi, pada kenyataannya tidak semua orang menggunakan email dengan baik dan bahkan ada banyak sekali penyalahgunaan email sehingga berpotensi untuk merugikan orang lain. *Email* yang disalahgunakan ini biasa dikenal sebagai spam atau *junkmail* (email sampah) yang mana email tersebut berisikan iklan, penipuan dan bahkan virus. (Pratiwi & Ulama, 2016). Hampir di semua aktifitas di internet dapat dengan mudah ditemukan spam. Keberadaan dan sifat spam yang dilakukan terus menerus dan menyampaikan hal yang kurang penting sangat menggangu dan dapat dibilang cukup meresahkan pengguna internet. (Imran, 2014) Misalnya, ketika pengguna menerima jumlah spam email yang cukup besar, banyak pengguna email harus menghabiskan waktu mereka untuk menghapus pesan yang tidak diinginkan tersebut. Bahkan karena itu, bisa jadi pesan yang penting terhapus. Saat ini belum diketahuin metode klasifikasi yang akurat dalam mengklasifikasikan email, apakah email yang diterima berupa spam atau email yang benar. Sehingga perlu diketahui bagaimana akurasi dari metode klasifikasi data mining yaitu Naïve Bayes, SVM, J48 dan *Random Forest*. Maksud dari penelitian ini adalah untuk melakukan analisis dan mendapatkan nilai akurat dari komparasi algoritma Naïve Bayes, SVM, J48 dan *Random Forest* dalam klasifiaksi *spam email.*

Email Spam

Email adalah singkatan dari *electronic mail* yang merupakan surat atau pesan dengan format digital. (Zakaria) Email banyak dapat diakses dengan mudah dengan berbagai gadget seperti komputer maupun ponsel smartphone.Email spam atau juga dikenal dengan email sampah adalah pesan massal yang tidak diminta, yang dikirim melalui email. Penggunaan spam telah memakin popular sejak awal 1990-an dan merupakan masalah yang dihadapi oleh sebagian besar pengguna email. Spammer biasanya mengirim email ke jutaan email, dengan harapan bahwa sejumlah kecil akan merespon atu berinteraksi dengan pesan tersebut. (Rouse, 2017).Seorang marketer asal Amerika Serikat, Gary Thuerk adalah orang pertama yang memberondong pesan tak diinginkan, ke 400 penerima pada tahun 1978. Thuerk, yang kala itu menjabat sebagai manajer pemasaran perusahaan Digital Equipment Corporation, mengirimkan promosi produk komputer perusahaannya. Alhasil, pesan promosi yang dikirim Thuerk menuai amarah dari para penerimanya Pesan email yang dikirim Thuerk itulah yang dinobatkan sebagai contoh email spam pertama di dunia. (Periwi, 2018).

Data mining

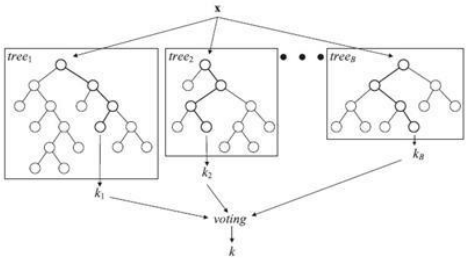
Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Interpretation/Evaluation Pola informasi yang dihasilkan dari proses data mining diterjemahkan menjadi bentuk yang lebih mudah dimengerti oleh pihak yang berkepentingan. Data mining merupakan sebuah proses, sehingga dalam melakukan proses tersebut harus sesuai dengan prosedur yaitu yang disebut dengan CRISP-DM (Cross-Industry Standard Process for Data Mining) yaitu sebagai keseluruhan proses, preprocessing data, pembentukan model, model evaluasi dan akhirnya penyebaran model (Larose, 2005). Enam fase tahapan Crisp menurut (Larose, 2005):

1. Fase pemahaman bisnis
2. Fase pemahaman data
3. Fase pengolahan data
4. Fase pemodelam
5. Fase evaluasi
6. Fase penyebaran

Algoritma Random Forest

Algoritme Random Forest (RF) merupakan pengembangan dari metode Classification and Regression Tree (CART) dengan menerapkan metode bootstrap aggregating (bagging) dan *random feature selection* (Breiman 2001). Algoritme RF merupakan algoritme yang cocok digunakan untuk klasifikasi data yang besar dan pada algoritme RF tidak terdapat pruning atau pemangkasan variabel seperti pada algoritme decision tree. Metode RF menggabungkan banyak pohon (*tree*) tidak seperti single tree yang hanya terdiri dari satu pohon untuk membuat klasifikasi dan prediction class. Pada RF pembentukan tree dilakukan dengan cara melakukan training sampel data. Sampling with replacement adalah cara yang digunakan untuk mengambil sampel data. Pemilihan variabel yang digunakan untuk split diambil secara acak. Klasifikasi dijalankan setelah semua tree terbentuk. Penentuan klasifikasi pada

RF ini diambil berdasarkan vote dari masing-masing tree dan vote terbanyak yang menjadi pemenang. Arsitektur umum dari RF dapat dilihat pada Gambar dibawah ini.



Gambar 1. Arsitektur umum *Random Forest* (Verikas *et al.*2011)

Naïve Bayes

Naïve Bayes merupakan sebuah pengklasifikaian probablistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma menggunakan teorema Bayes dan mengamsusikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. Definisi lain yang dikemukakan oleh ilmuan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya. (Bustami, 2013). Naïve Bayes sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan (Pattekari,2012)

Persamaan dari teorema Naïve Bayes adalah:

$$P(H|X)=\frac{P(x|H).P(H)}{(X)}$$

Dimana:

X : Data dengan class yang belum diketahui

H : Hipotesis data yang merupakan suatu class spesifik

P(H|X) : Probabilitas hipotesisis H berdasarkan kondisi X (posteriori probabilitas)

P(H) : Probabilitas hipotesisis H (prior probabilitas))

P(X\H) : Probabilitas X berdasarkan kondisi pada hipotesis H

P(X) : Probabilitas X

Support Vector Machine

Support Vector Machine atau SVM adalah metode regresi atau pengklasifikasian data berdasarkan data-data sebelumnya dan permodelannya di supervisi terlebih dahulu. SVM termasuk kedalam jenis klasifikator yang biner, linier dan non probabilistik. SVM menggunakan decision boundary (batas keputusan) yang akan menentukan klasifikasi dari data-data pelatihan sehingga dapat dibentuk sebuah model linier atau hyperplane yang paling optimal untuk mengklasifikasikan data data tersebut.

Secara matematika, konsep dasar SVM yaitu: (Widiastuti)

$$\min \frac{1}{2}|w|^2$$

$$s.t\ y_i(x_i.w+b)-1 \geq 0$$

Dimana $(x_i.w+b)\geq 1$ untuk kelas 1, dan $(x_i.w+b) \leq -1$ untuk 2, x_i adalah data set, adalah output dari data x_i, dan w,b adalah parameter yang dicari nilainya. Formulasi optimasi SVM untuk kaus klasifikasi dua kelas dibedakan menjadi klassifikasi linier dan non-linier.

Algoritma J48

J48 adalah salah satu algoritma yang sama persis dengan C45 namun terdapat dalam software WEKA. Algoritma J48 membangun sebuah pohon keputusan berdasarkan pada seperangkat input datayang berlabel. Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan. Berikut ini tahapan algoritme J48 :

a. Menyiapkan data training
b. Menentukan akar dari pohon.
c. Menghitung nilai Gain melalui Persamaan

Tinjauan Studi

Beberapa penelitian terdahulu yang relevan dengan pengklasifikasian email spam menggunakan beberapa metode. Hal ini digunakan sebagai tolak ukur dari hasil penelitian yang telah dicapai. Berikut beberapa penelitian terdahulu terkait email spam:

Penelitian yang dilakukan oleh Sharma & Sahni. Penelitian ini melakukan perbandingan algoritma klasifikasi untuk analisis data email spam dengan empat algoritma yaitu ID3, J48, Simple CART dan ADTree. Hasil menunjukan algoritma J48 memiliki tingkat akurasi tinggi dari ke empat algortima terbut, yaitu 92,7624%.

Penelitian dilakukan oleh Rusland et al tentang analisis algoritma naïve bayes untuk filter email spam dengan beberapa dataset.

Penelitian yang dilakukan oleh Pratiwi dan Ulama tentang klasifikasi email spam menggunakan metode Support Vector Machine (SVM) dan *k-Nearest Neighbor.*

Tabel 1. Perbandingan penelitian terkait

Peneliti	Tahun	Metode	Hasil
Aman Kumar Sahrma, Suruchi Sahni	2011	Komparasi algoritma klasifikasi ID3, J48, Simple CART, dan ADtre	Tingkat akurasi paling tinggi didapatkan oleh algoritma J48 dengan akurasi 92,7624 %
Sheila Novelia Dharma Pratiwi, Brodjol Sutijo Ulama	2016	Penerapan metode algortima SVM (*Suoppr t Vector Machine), K-Nearest Neighbor*	Akurasi yang tertinggi di dapatkan oleh svm dengan tingkat akurasi 96,6%
Nurul Fitriah Rusland, Norfaradilla Wahid, Shahreen Kasim, Hanayanti Hafit	2017	Penerapan naïve Bayes untuk klasifikasi spam email	Tingkat akurasi yang di peroleh oleh naïve bayes 72,57 %

2. Metode Penelitian

Eksperimen dan pengujian model

Proses eksperimen yang penulis lakukan inin menggunakan Weka 3.8 untuk pengujian model dilakukan menggunakan dataset Spambase dari UCI Machine Learning Repository. Tahapan pengujian untuk klasifikasi Spam email sebagai berikut:

a. Menyiapkan dataset untuk eksperimen yang sudah diketahui classsnya.
b. Mendesain arsitekstur algoritma Naïve Bayes, SVM, J48 dan *Random Forest.*
c. Melakukan traing dan testing terhadap algortima Naïve Bayes, SVM, J48 dan *Random Forest* dan mencatat hasil *accuracy* dan AUC.

Evaluasi dan validasi Hasil

Validasi dilakukan 10 fold cross validation. Untuk 10 fold cross validation data eksperimen akan dibagi menjadi 10 bagian. Satu bagian untuk data testing Sembilan bagian lainnya untuk data training. Sedangkan pengukuran akurasi diukur dengan *confusion matrix* dan kurva ROC (*Receiver Operating Characteristics)* untuk mengukur nilai AUC.

Tabel 2. Confusion Matrix

Classification	*Predicted Class*		
		Class = Yes	*Class = No*
Observed Class	*Class = Yes*	*(True Positive-TP)*	*(False Negative -FN)*
	Class = No	*(False Positive-FP)*	*(True Negative -TN)*

3. Hasil dan Pembahasan

Pada bagian ini dijelaskan hasil tujuan penelitian ini melakukan anaslisis dan komparasi untuk memperoleh hasil yang paling akurat dari komparasi Naïve Bayes, SVM, J48 dan *Random Forest* untuk klasifikasi spam email.

3.1. Hasil Ekperimen menggunakan Naïve Bayes

Hasil yang diperoleh dengan menggunakan algoritma Naïve Bayes adalah nilai accuracy 78,8742% . seperti pada tabel 3 sebanyak 1765 data diprediksi sesuai yaitu spa dan sebanyak 924 data diprediksi spam tetapi ternyata non spam, 1864 data siprediksi sesuai yaitu non spam dan 48 data diprediksi non-spam tetapi ternyata spam

Tabel 3. Confusion matrix algoritma naïve bayes

	True Non Spam	True Spam
Pred. Non Spam	1864	924
Pred. Spam	48	1765

Nilai accuracy dari confusion matrix tersebut adalah sbagai berikut

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{1864+1765}{1864+1765+48+924} = 78{,}8742\ \%$$

3.2. Hasil Ekperimen menggunakan SVM

Hasil yang diperoleh dengan menggunakan algoritma Naïve Bayes adalah nilai accuracy 86,4812% . seperti pada tabel 3 sebanyak 1340 data diprediksi sesuai yaitu spa dan sebanyak 149 data diprediksi spam tetapi ternyata non spam, 2639 data siprediksi sesuai yaitu non spam dan 473 data diprediksi non-spam tetapi ternyata spam

Tabel 4. Confusion matrix algoritma SVM

	True Non Spam	True Spam
Pred. Non Spam	2639	149
Pred. Spam	473	1340

Nilai accuracy dari confusion matrix tersebut adalah sbagai berikut

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{2639+1340}{2639+1340+473+149} = 86{,}4812\ \%$$

3.3. Hasil Ekperimen menggunakan J48

Hasil yang diperoleh dengan menggunakan algoritma Naïve Bayes adalah nilai accuracy 92,4972% . seperti pada tabel 3 sebanyak 1631 data diprediksi sesuai yaitu spam dan sebanyak 154 data diprediksi spam tetapi ternyata non spam, 2634 data siprediksi sesuai yaitu non spam dan 182 data diprediksi non-spam tetapi ternyata spam

Tabel 5. Confusion matrix algoritma J48

	True Non Spam	True Spam
Pred. Non Spam	2634	154
Pred. Spam	182	1631

Nilai accuracy dari confusion matrix tersebut adalah sbagai berikut

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{2634+1631}{2634+1631+182+154} = 92{,}4972\%$$

3.4. Hasil Ekperimen menggunakan Random Forest

Hasil yang diperoleh dengan menggunakan algoritma Naïve Bayes adalah nilai accuracy 94,2186% . seperti pada tabel 3 sebanyak 1673 data diprediksi sesuai yaitu spam dan sebanyak 126 data diprediksi spam tetapi ternyata non spam, 2662 data siprediksi sesuai yaitu non spam dan 140 data diprediksi non-spam tetapi ternyata spam

Tabel 6. Confusion matrix algoritma Random Forest

	True Non Spam	True Spam
Pred. Non Spam	2662	126
Pred. Spam	140	1673

Nilai accuracy dari confusion matrix tersebut adalah sbagai berikut

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{2662+1673}{2662+1673+140+126} = 94{,}2186\ \%$$

4. Kesimpulan

Dalam penelitian ini dilakukan pengujian model dengan membandingkan empat metode *data mining* yaitu algoritma *Naïve Bayes, SVM, J48, dan Random Forest,* hasil dari evaluasi dan validasi, diketahui bahwa *Random Forest* memiliki akurasi yang paling tinggi diantara metode yang dikomparasikan sebesar 94,2186% dan AUC sebesar 97,7 %, diikuti oleh algoritma *J48,SVM* dan *Naïve Bayes* yan yang memiliki akurasi yang paling rendah. Dengan demikian hasil evaluasi menggunakan curva ROC yaitu, algoritma klasifikasi Random Forest bernilai 97,7% dengan tingkat diagnose *excellent classification*.Dapat disimpulkan bahwa penggunaan metode Random Forest merupakan metode yang cukup baik dalam memprediksi spam email.

Pada bagian ini, penulis memberikan saran- saran (1). Menggunakan metode lain seperti AdaBost, (2). Melakukan pengembangan dengan *feature selection* yang lain seperti *genetic algorithm* , PSO dan metode *feature selection* lainnya untuk menyeleksi atribut yang berpengaruh kuat, (3). Penelitian ini dapat dikembangkan dengan membandingkan algortima *data mining* lainya misalkan saja dengan metode *Support vector machine, Knearest Neighbor, CART* dan lainya atau dapat mengoptimalkan parameter dengan *Particle Swarm Optimization, Genetic Algoritmh* dan lainya

Referensi

Dang, V., & Croft, W. B. (n.d.). Feature Selection for Document Ranking using Best First Search and Coordinate Ascent, 2–5.

Mongkareng, D., Setiawan, N. A., & Permanasari, A. E. (2017). Implementasi Data Mining dengan Seleksi Fitur untuk Klasifikasi Serangan pada Intrusion Detection System (IDS), (gambar 2), 314–321.

Novelia, S., Pratiwi, D., Sutijo, B., & Ulama, S. (2016). Klasifikasi Email Spam dengan Menggunakan Metode Support Vector Machine dan k- Nearest, *5*(2), 344–349.

Parveen, P., & Halse, P. G. (2016). Spam Mail Detection using Classification, *5*(6),347–349. https://doi.org/10.17148/IJARCCE.2016.5674

Sharma, A. K. (2011). A Comparative Study of Classification Algorithms for Spam Email Data Analysis, *3*(5), 1890–1895.

Tree-j, A. D. (2017). Algoritma decision tree-j48, k-nearest, dan zero-r pada kinerja akademik, 12–18.