Infotek : Jurnal Informatika dan Teknologi
Vol. 6 No. 1, Januari 2023
Hal. 11-20

e-ISSN 2614-8773

DOI : 10.29408/jit.v6i1.7049 Link : https://dx.doi.org/10.29408/jit.v6i1.7049

Deteksi Spam Email Dengan Metode Naive Bayes Dan Particle Swarm Optimization (PSO)

Muhamad Abdul Ghani[1] , Hamdun Sulaiman[2]
[1,2]Program Studi Sistem Informasi, Universitas Bina Sarana Informatika
*hamdun.hsl@bsi.ac.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 pengolahan data dari email gmail dengan teks mining lalu menguji dengan beberapa metode klasifikasi data mining diantaranya yaitu Algoritma *Naïve Bayes, SVM, Random Forest* dan dipadukan dengan *Partical Swarm Optimization* 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 Naïve Bayes dengan *Partical Swarm Optimization (PSO)* memiliki nilai accuracy paling tinggi, yaitu 81.40 % dan AUC 0,78.

Kata kunci : Email Spam, Algoritma naive bayes, Support Vector Machine, Random forest, Teks Mining

Abstract

Internet-based technology has become a primary need. Based on the survey results from the Central Statistics Agency in collaboration with APJII, email sending and receiving activities have outperformed social media positions by reaching 95.75%. Very intense use of email can have both positive and negative effects. Because apart from being a communication tool, in reality not everyone uses email well and there are even so many misuses of email that have the potential to harm others. This misused email is commonly known as spam or junkmail (junk email) which contains advertisements, scams and even viruses. In this study, data processing from gmail emails with text mining was carried out and then tested with several data mining classification methods including the Naïve Bayes Algorithm, SVM, Random Forest and combined with Partical Swarm Optimization in predicting spam emails with the aim that the selected algorithm is the most accurate. From the test results by measuring the performance of the four algorithms using Confusion Matrix and ROC, it is known that the Naïve Bayes algorithm with Partical Swarm Optimization (PSO) has the highest accuracy value, namely 81.40% and AUC 0.78.

Keywords : Email Spam, Naive Bayes Algorithm, Support Vector Machine, Random forest, Text Mining

1. Pendahuluan

Internet telah membawa dampak perubahan yang sangat besar bagi masyarakat. Dimana segala kegiatan manusia telah berganti menjadi aktivitas digital di dunia internet. Sebagai bagian dari konvergensi telematika, dimana terdapat tiga unsur yaitu telekomunikasi, media dan

Infotek : Jurnal Informatika dan Teknologi
Vol. 6 No. 1, Januari 2023
Hal. 11-20

e-ISSN 2614-8773

DOI : 10.29408/jit.v6i1.7049 Link : https://dx.doi.org/10.29408/jit.v6i1.7049

informatika, internet telah menjadi bagian tak terpisahkan dalam kehidupan manusia [9].

Email telah menjadi salah satu alat komunikasi internet yang mudah dan cepat. Tetapi masih banyak masalah yang dihadapi oleh pengguna email. Masalah utama yang sering dihadapi adalah meningkatnya jumlah email yang tidak diharapkan atau yang biasa disebut spam. Pesan spam dapat berdampak pada penyalahgunaan koneksi internet dan sangat mengganggu pengguna. Pada umumnya, spam berisikan iklan, link situs yang tidak baik seperti pornografi dan virus yang dapat merusak komputer. Permasalahan tersebut, dapat diatasi dengan membuat sebuah anti spam. Anti spam ini berfungsi untuk mendeteksi email dan memberikan informasi kepada pengguna apabila terdapat suatu pesan yang memiliki potensial sebagai spam. Salah satu teknik untuk membuat anti spam adalah Bayesian Network (BN), yang dapat digunakan untuk memperkirakan kemungkinan bahwa pesan yang masuk adalah spam [12].

Metode Bayesian Network (BN) merupakan salah satu Probabilistic Graphical Model (PGM) yang dibangun dari teori probabilistik dan teori graf. Selain digunakan untuk mendeteksi sebuah pesan spam, metode BN juga dapat digunakan untuk mendiagnosa suatu penyakit [6].

Dari permasalahan tersebut kita bisa menguji deteksi spam email menggunakan beberapa metode klasifikasi data mining, diantaranya yaitu algoritma Naïve Bayes, Support Vector Machine (SVM), Random Forest, dengan melakukan pengolahan data email dari akun gmail dan melakukan preprocessing lalu menghitung akurasinya. Ketiga metode tersebut digunakan dalam memprediksi spam email dengan tujuan agar algoritma terpilih merupakan algoritma yang paling akurat, sehingga dapat melakukam prediksi spam email. Dari ketiga metode tersebut akan dipadukan dengan Particle Swarm Optimization (PSO) dengan tujuan untuk meningkatkan akurasi dalam memprediksi email spam.

2. Tinjauan Pustaka

2.1. Penelitian Terkait

Berkaitan dengan penelitian deteksi email spam, tentunya sudah banyak yang meneliti perihal tersebut, terdapat 5 penelitian terkait:

1. Penelitian yang dilakukan oleh Nur Qodariyah Fitriyah, Hardian Oktavianto, Hasbullah dengan judul “Deteksi Spam Pada Email Berbasis Fitur Konten Menggunakan Naïve Bayes” yang terbit pada tahun 2019. Penelitian tersebut menghasilkan rata – rata akurasi sebesar 84.8% [1]
2. Penelitian selanjutnya dilakukan oleh Shiela Novelia Dharma Pratiwi, Brodjol Sutijo Suprih Ulama dengan judul “Klasifikasi Email Spam dengan Menggunakan Metode Support Vector

Infotek : Jurnal Informatika dan Teknologi
Vol. 6 No. 1, Januari 2023
Hal. 11-20

e-ISSN 2614-8773

DOI : 10.29408/jit.v6i1.7049 Link : https://dx.doi.org/10.29408/jit.v6i1.7049

Machine dan k-Nearest Neighbor" yang terbit pada tahun 2016. Penelitian tersebut menghasilkan ketepatan klasifikasi terbaik pada saat k=3 dengan hasil ketepatan klasifikasi sebesar 92.28% dengan error 7.72 % sedangkan kombinasi metode SVM menggunakan kernel linier dan RBF dengan 10-fold cv menghasilkan ketepatan klasifikasi terbaik dengan menggunakan SVM linier dengan ketepatan klasifikasi yang diberikan sebesar 96.6% dengan error 3.4% sehingga disimpulkan metode SVM lebih baik dibanding metode KNN [2]

3. Penelitian selanjutnya yaitu dari Aman Kumar Sharma dengan judul "A Comparative Study of Classification Algorithms for Spam Email Data Analysis" yang terbit pada tahun 2011. Yang telah melakukan suatu perbandingan algoritma klasifikasi untuk analisis data email spam. Dalam penelitiannya telah melakukan percobaan untuk menentukan akurasi klasifikasi dengan empat algoritma yaitu ID3, J48, Simple CART dan ADTree, dengan hasil kinerja akurasi klasifikasi tertinggi adalah J48 [3].
4. Penelitian selanjutnya dari Fahrur Rozi dan Rikie Kartadie dengan judul "Deteksi E-Mail Dan Spam Menggunakan Fuzzy Association Rule Mining" yang terbit pada tahun 2017. Berdasarkan pengujian yang dilakukan, penggunaan metode Fuzzy Association Rule Mining dalam pendeteksian SPAM dan non SPAM pada e-mail adalah baik, hal ini terlihat pada pengujian dimana nilai accuracy, precision, recall, dan F-Measure cukup tinggi yang mendekati nilai 1 [4].
5. Penelitian yang kelima dilakukan oleh Ratih Yulia Hayuningtyas dengan judul "Aplikasi Filtering of Spam Email Menggunakan Naïve Bayes" terbit pada tahun 2017. Metode klasifikasi yang digunakan adalah Naïve Bayes merupakan metode penyaringan yang paling popular. Evaluasi menggunakan confusion matrix yang menghasilkan akurasi sebesar 75,9% [5]

2.2. Landasan Teori

1. Email

Electronic mail merupakan salah satu dari sekian banyak layanan internet yang ada saat ini selain Netnews, Telnet, File Transfer Protokol (FTP) dan World Wide Web (www) dan masih banyak layanan yang lainnya. Layanan internet adalah berbagai program atau fasilitas yang disediakan oleh internet. Dari layanan internet tersebut yang paling banyak digunakan adalah layanan internet electronic mail. Electronic mail adalah layanan yang diberikan oleh internet yang berkembang sejak tahun 1960. Pada saat itu Internet belum terbentuk, yang ada hanyalah kumpulan 'mainframe' yang terbentuk sebagai jaringan. Mulai tahun 1980-an, electronic mail sudah bisa

Infotek : Jurnal Informatika dan Teknologi
Vol. 6 No. 1, Januari 2023
Hal. 11-20

e-ISSN 2614-8773

DOI : 10.29408/jit.v6i1.7049 Link : https://dx.doi.org/10.29408/jit.v6i1.7049

dinikmati oleh khalayak umum. Electronic mail adalah salah satu proses pengiriman surat melalui internet dengan menggunakan [13].

2. Spam

Spam atau junk mail adalah penyalahgunaan dalam pengiriman berita elektronik untuk menampilkan berita iklan dan keperluan lainnya yang mengakibatkan ketidaknyamanan bagi para pengguna [8].

3. Text mining

Text mining adalah proses mengeksplorasi dan menganalisis sejumlah besar data teks tidak terstruktur yang dibantu oleh perangkat lunak yang dapat mengidentifikasi konsep, pola, topik, kata kunci, dan atribut lainnya dalam data [14]'

4. Data Mining

Data mining adalah proses yang memperkerjakan satu atau lebih teknik pembelajaran komputer (machine learning) untuk menganalisis dan mengekstraksi pengetahuan (knowledge) secara otomatis. Data mining merupakan proses iterative dan interaktif untuk menemukan pola atau model baru yang sempurna, bermanfaat dan dapat dimengerti dalam suatu database yang sangat besar (massive database). Data mining berisi pencarian trend atau pola yang diinginkan dalam database besar untuk membantu pengambil keputusan diwaktu yang akan datang, pola-pola ini dikenali perangkat tertentu yang dapat memberikan suatu analisa data yang berguna dan berwawasan yang kemudian dapat dipelajari dengan lebih teliti, yang mungkin saja menggunakan perangkat pendukung keputusan yang lain [15].

5. Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) merupakan algoritma pencarian berbasis populasi dan diinisialisasi dengan populasi solusi acak dan digunakan untuk memecahkan masalah optimasi. Particle swarm optimization (PSO) adalah teknik yang terinspirasi oleh proses alami burung yang berkelompok, dan juga dikenal sebagai segerombolan intelijen dengan mempelajari perilaku sosial atau kelompok hewan [10].

6. Random Forest

Random Forest (RF) adalah klasifikasi yang terdiri dari beberapa pohon keputusan. Setiap pohon keputusan dibuat menggunakan vektor acak. Secara umum vektor acak disisipkan dalam pembentukan pohon yaitu dengan memilih nilai F acak, seperti F atribut (fitur) masukan untuk dibagi pada setiap node di pohon keputusan yang akan dibentuk. Dengan memilih nilai acak F maka tidak harus memeriksa semua atribut yang ada dan melihat nilai F yang dipilih atribut [11].

7. Naïve Bayes

Bayesian classification adalah pengklasifikasian statistik yang dapat digunakan untuk memprediski probabilitas keanggotaan suatu class. Bayesian classification didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan decesion tree dan neural network.

Infotek : Jurnal Informatika dan Teknologi
Vol. 6 No. 1, Januari 2023
Hal. 11-20

e-ISSN 2614-8773

DOI : 10.29408/jit.v6i1.7049 Link : https://dx.doi.org/10.29408/jit.v6i1.7049

Bayesian classification terbukti memiliki akurasai dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar. Metode Bayes merupakan pendekatan statistic untuk melakukan inferensi induksi pada persoalan klasifikasi. Pertama kali dibahas terlebih dahulu tentang konsep dasar dan definisi pada Teorema Bayes, kemudian menggunkan teorema ini untuk melakukan klasifikasi dalam Data Mining [7].

8. Support Vector Machine

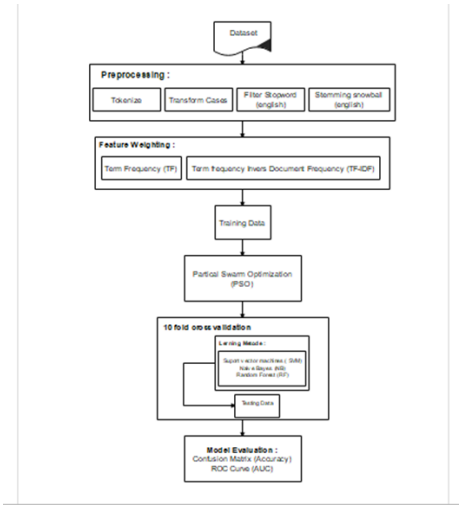
Pada tahun 1992 Support Vector Machine (SVM) diperkenalkan oleh Vapnik, Boser dan Guyon. SVM merupakan salah satu teknik yang relatif muda dibandingkan dengan teknik lainnya, tetapi memiliki performansi yang lebih baik berdasarkan metode dengan keterlibatan beberapa kernel yang menyangkut beberapa bidang untuk menunjukkan peningkatan. SVM merupakan sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi-fungsi linier dalam sebuah ruang ciri (feature space)berdimensi tinggi, dilatih dengan algoritma pembelajaran yang didasarkan pada teori optimasi dengan mengimplementasikan bias pembelajaran yang berasal dari teori pembelajaran statistik [16].

3. Metode Penelitian

Pada penelitian ini menggunakan metode experiment, yaitu penelitian yang melibatkan penyelidikan kepada beberapa variable menggunakan tes tertentu yang dikendalikan sendiri oleh peneliti untuk melakukan pengklasifikasian dengan pengumpulan data, pengolahan tahap awal, metode yang diusulkan, Experimen dan pngujian model trakhir evaluasi dan validasi.

Penelitian ini menggunakan data primer berupa data spam email dari email Gmail , data di sajikan dalam bentuk text sebanyak 281 data yang terdiri dari data spam sebanyak 86 dan data non-spam sebanyak 195 di ambil tahun 2021, perangkat lunak yang digunaka untuk menganalisa dengan aplikasi Rapid Miner 9.0.

Pada penelitian ini data training dan data testing akan dipisah dengan menggunakan 10 fold cross validation. Dataset tersebut akan dibagi menjadi 10 bagian dan akan dilakukan pengulangan sebanyak 10 pengulangan. Contoh pada iterasi ke-3, jika bagian ketiga dijadikan sebagai data testing maka sisa bagian lainnya akan digunakan sebagai data training. Pengambilan data tersebut dilakukan secara acak agar semua data dapat menjadi data training juga menjadi data testing.



Infotek : Jurnal Informatika dan Teknologi
Vol. 6 No. 1, Januari 2023
Hal. 11-20

e-ISSN 2614-8773

DOI : 10.29408/jit.v6i1.7049 Link : https://dx.doi.org/10.29408/jit.v6i1.7049

Gambar 1. Bagan langkah penelitian

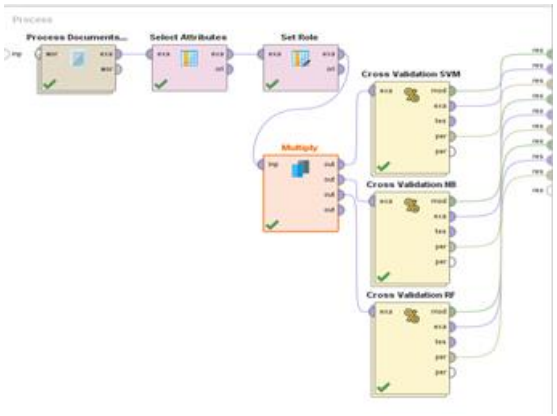
4. Hasil dan Pembahasan

Pada bagian ini di jelaskan hasil tujuan penelitian ini melakukan analisis dan komparasi untuk memperoleh hasil yang paling tinggi akurasinya dalam memadukan metode klasifikasi algoritma Support Vector machine, Naive Bayes, Random Forest dengan Partical Swarm Optimization (PSO). Sebelum data dikomparasikan ada tahapan pengolahan data terlebih dahulu untuk mendapatkan data yang bersih dan siap untuk digunakan dalam penelitian. Dalam *text mining* tahapan awal yang akan dilakukan adalah tahap *preprocessing*.

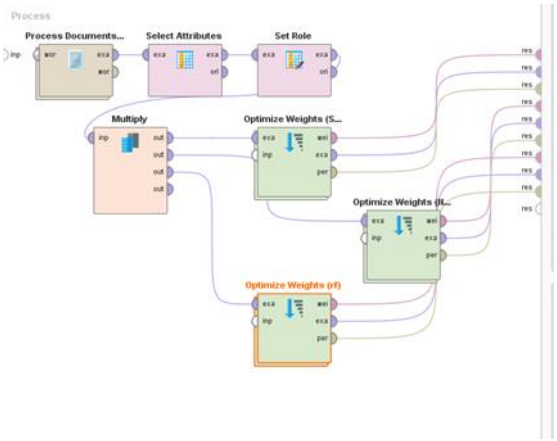
Gambar 2. Desain Model Preprocessing

Pada Gambar 1 diatas menjelaskan tentang alur proses preprocessing. Adapun tahapan-tahapan yang digunakan dalam proses *preprocessing* adalah , *Tokenization :linguistic token, Tranform Cases, Filtering stopword (english), Stemming englih (snowball).* Setelah data di preprosessing selanjutnya baru pemodelan pada tahapam iniini secara langsung melibatkan teknik data mining yaitu dengan cara melakukan pemilihan teknik data mining dan menetapkan algoritma yang akan digunakan. Tool yang digunakan pada fase pemodelan ini adalah Rapidminer versi 9.0.

Adapun hasil dalam pengujian model yang dilakukan adalah mengklasifikasi spam dan non spam menggunakan algoritma Support Vector Machine, Naïve Bayes,Random Forest dan penambahan Particle Swarm Optimization.



Gambar 3. Desain Awal SVM ,NB dan RF



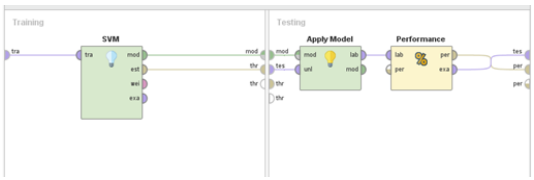
Gambar 4. Desain SVM+PSO ,NB+PSO dan RF + PSO

Gambar 2 dan 3 itu sebagai Desain Awal mode *Naive Bayes Classifier* ,*Support Vector Machine* dan *Random Forest* yang dipaduan dengan PSO.Selanjutnya dilakukan proses pengujian model metode Support Vector Machine.

Infotek : Jurnal Informatika dan Teknologi
Vol. 6 No. 1, Januari 2023
Hal. 11-20

e-ISSN 2614-8773

DOI : 10.29408/jit.v6i1.7049 Link : https://dx.doi.org/10.29408/jit.v6i1.7049

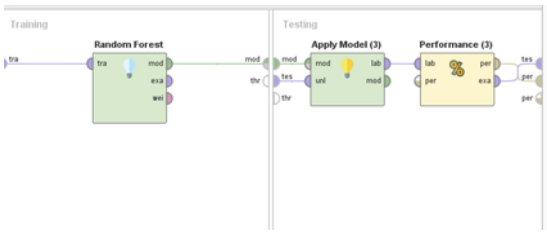


Gambar 5. Desain Algoritma SVM di dalam Cross Validation

Gambar 4 Menjelaskan desain proses di dalam operator *cross validation SVM* pada gambar 2. Pada pengujian ini, data digunakan adalah data bersih yang telah melalui *preprocessing.* Data tersebut diambil dari file process documents, hal ini dilakukan karena dataset disimpan dalam bentuk text document. Process validasi terdiri dari *data training* dan *data testing.* Selanjutnya dilakukan proses pengujian model metode Naive Bayes Classifier.

Gambar 6. Desain Algoritma Naive Bayes di dalam Cross Validation

Gambar 5 Menjelaskan desain proses di dalam operator *cross validation NB* pada gambar 2. Pada pengujian ini, data digunakan adalah data bersih yang telah melalui *preprocessing*. Proses validasi terdiri dari *data training* dan *data testing.* Subproses *training* digunakan untuk melatih model. Model yang dilatih kemudian diproses pada subproses *testig.* Selanjutnya dilakukan proses pengujian model metode *Random Forest.*



Gambar 7. Desain Algoritma *Random Forest* di dalam Cross Validation

Gambar 6 Menjelaskan desain proses di dalam operator *cross validation RF* pada gambar 2. Pada pengujian ini, data digunakan adalah data bersih yang telah melalui *preprocessing*. Proses validasi terdiri dari *data training* dan *data testing.* Subproses *training* digunakan untuk melatih model. Model yang dilatih kemudian diproses pada subproses *testing.*

Selanjut tahap evaluasi bertujuan untuk menentukan nilai kegunaan dari model yang telah berhasil dibuat pada langkah sebelumnya. Untuk evaluasi digunakan *10-fold cross validation*. Dari hasil pemodelan yang telah dilakukan sebelumnya Berikut ini akan dijelaskan dan *Confusion Matrix* dari masing-masing algoritma .

Tabel 1. Confusion Matrix SVM

accuracy: 74.51% +/- 9.16% (mikro average: 74.42%)			
	true spam	true non spam	class precision
pred. negative	69	27	71.88%
pred. positive	17	59	77.63%
class recall	80.23%	68.60%	

$$Acc\ (Accuracy) = \frac{TP + TN}{TP+TN+FP+FN} = \frac{69 + 59}{69 + 59 + 17 + 27} = \frac{128}{172} = 0.7442$$

Akurasi yang diperoleh yaitu 74.51 % dari 86 data spam dan 86 data non spam . Data spam yang sesuai prediksi yaitu 69 data. Data spa, yang termasuk ke dalam prediksi positif yaitu 17 data.

Infotek : Jurnal Informatika dan Teknologi
Vol. 6 No. 1, Januari 2023
Hal. 11-20

e-ISSN 2614-8773

DOI : 10.29408/jit.v6i1.7049 Link : https://dx.doi.org/10.29408/jit.v6i1.7049

Data non spam yang termasuk kedalam prediksi negatif yaitu 27 data dan data non spam yang sesuai prediksi yaitu 59 data.

Tabel 2. Confusion Matrix NB

accuracy: 71.60% +/- 8.38% (mikro: 71.51%)			
	true spam	true non spam	class precision
pred. negative	67	30	69.07%
pred. positive	19	56	74.67%
class recall	77.91%	65.12%	

$$Acc\ (Accuracy) = \frac{TP + TN}{TP+TN+FP+FN} = \frac{67 + 56}{67 + 56 + 19 + 30} = \frac{123}{172} = 0.7151$$

Akurasi yang diperoleh yaitu 71.51 % dari 86 data spam dan 86 data non spam . Data spam yang sesuai prediksi yaitu 67 data. Data spam yang termasuk ke dalam prediksi positif yaitu 19 data. Data non spam yang termasuk kedalam prediksi negatif yaitu 30 data dan data non spam yang sesuai prediksi yaitu 56 data.

Tabel 3. Confusion Matrix RF

accuracy: 60.52% +/- 6.85% (mikro: 60.47%)			
	true spam	true non spam	class precision
pred. negative	66	48	57.89%
pred. positive	20	38	65.52%
class recall	76.74%	44.19%	

$$Acc\ (Accuracy) = \frac{TP + TN}{TP+TN+FP+FN} = \frac{66 + 38}{66 + 38 + 20 + 48} = \frac{104}{172} = 0.6047$$

Akurasi yang diperoleh yaitu 60.47 % dari 86 data spam dan 86 data non spam . Data spam yang sesuai prediksi yaitu 66 data. Data spam yang termasuk ke dalam prediksi positif yaitu 20 data. Data non spam yang termasuk kedalam prediksi negatif yaitu 4 8 data dan data non spam yang sesuai prediksi yaitu 38 data.

Tabel 4. Confusion Matrix SVM + PSO

accuracy: 77.84% +/- 7.88% (mikro average: 77.91%)			
	true spam	true non spam	class precision
pred. negative	77	29	72.64%
pred. positive	9	57	86.36%
class recall	89.53%	66.28%	

$$Acc\ (Accuracy) = \frac{TP + TN}{TP+TN+FP+FN} = \frac{77 + 57}{77 + 57 + 9 + 29} = \frac{134}{172} = 0.7791$$

Akurasi yang diperoleh yaitu 77.91% dari 86 data spam dan 86 data non spam . Data spam yang sesuai prediksi yaitu 77 data. Data spam yang termasuk ke dalam prediksi positif yaitu 9 data. Data non spam yang termasuk kedalam prediksi negatif yaitu 2 9 data dan data non spam yang sesuai prediksi yaitu 57 data.

Tabel 5. Confusion Matrix NB + PSO

accuracy: 81.34% +/- 5.25% (mikro: 81.40%)			
	true spam	true non spam	class precision
pred. negative	72	18	69.07%
pred. positive	14	68	74.67%
class recall	83.72%	65.12%	

$$Acc\ (Accuracy) = \frac{TP + TN}{TP+TN+FP+FN} = \frac{72 + 68}{72 + 68 + 14 + 18} = \frac{140}{172} = 0.8140$$

Akurasi yang diperoleh yaitu 81.40 % dari 86 data spam dan 86 data non spam . Data spam yang sesuai prediksi yaitu 72 data. Data spam yang termasuk ke dalam prediksi positif yaitu 14 data. Data non spam yang termasuk kedalam prediksi negatif yaitu 18 data dan data non spam yang sesuai prediksi yaitu 68 data.

Tabel 6. Confusion Matrix RF + PSO

accuracy: 70.82% +/- 8.56% (mikro: 70.93%)			
	true spam	true non spam	class precision
pred. negative	70	34	67.31%
pred. positive	16	52	76.47%
class recall	81.40%	60.47%	

$$Acc\ (Accuracy) = \frac{TP + TN}{TP+TN+FP+FN} = \frac{70 + 52}{70 + 52 + 16 + 34} = \frac{122}{172} = 0.7093$$

Akurasi yang diperoleh yaitu 70.82 % dari 86 data spam dan 86 data non spam . Data spam yang sesuai prediksi yaitu 70 data. Data spam yang termasuk ke dalam prediksi positif yaitu 16 data.

Infotek : Jurnal Informatika dan Teknologi
Vol. 6 No. 1, Januari 2023
Hal. 11-20

e-ISSN 2614-8773

DOI : 10.29408/jit.v6i1.7049 Link : https://dx.doi.org/10.29408/jit.v6i1.7049

Data non spam yang termasuk kedalam prediksi negatif yaitu 34 data dan data non spam yang sesuai prediksi yaitu 52 data.

Adapun perbandingan hasil komparasi akurasi dan AUC Algoritma telah digunakan sebagai berikut:

Tabel 7. Perbandingan Nilai Akurasi dan Area Under Curve (AUC)

Algoritma	*Accuracy*	AUC
SVM	74.42%	0.791
NB	71.51%	0.664
RF	60.47%	0.708
SVM + PSO	77.91%	0.834
NB+ PSO	81.40%	0.787
RF+ PSO	70.93%	0.792

Dalam penelitian ini, hasil perhitungan metode Naïve bayes dengan PSO mendapatkan nilai akurasi 81.40 %. berdasarkan Tabel 4.7, dapat disimpulkan bahwa akurasi algoritma Naïve bayes dengan PSO mendapatkan akurasi yang lebih tinggi dibandingkan algoritma yang lainya. Model klasifikasi teks yang digunakan dapat memudahkan untuk mengetahui email spam dan non spam.

Berdasarkan data yang diolah menggunakan *tool Rapidminer*, data email akan terpisah menjadi kata-kata yang memiliki bobot pada setiap kata-katanya. Katakata tersebut akan digunakan untuk melihat kata –kata yang berhubungan dengan sentimen yang sering muncul dan memiliki bobot tertinggi dan dapat digunakan untuk mengetahui spam dan non spam.

5. Kesimpulan

Dalam penelitian ini dilakukan pengklasifikasian teks mining dan pengujian model dengan membandingkan metode algoritma Naïve Bayes, SVM, Random Forest dan penambahan Partical Swarm Optimization, hasil dari evaluasi dan validasi, diketahui bahwa Naïve Bayes dengan PSO yang memiliki akurasi yang paling tinggi diantara metode yang dikomparasikan sebesar 81.40 % dan AUC sebesar 78,7 %, Dapat disimpulkan bahwa penggunaan metode Naïve Bayes dengan PSO merupakan metode yang cukup baik dalam memprediksi spam email gmail.

6. Daftar Pustaka

[1]. N. Q. Fitriyah, H. Oktavianto and H. , "Deteksi Spam Pada Email Berbasis Fitur Konten Menggunakan Naïve Bayes," JUSTINDO (Jurnal Sistem & Teknologi Informasi Indonesia), vol. 5, no. 1, pp. 1-7, 2020.

[2]. S. N. D. Pratiwi and B. S. S. Ulama, "Klasifikasi Email Spam dengan Menggunakan Metode Support Vector Machine dan k-Nearest Neighbor," JURNAL SAINS DAN SENI ITS, vol. 5, no. 2, pp. D-344 - D-349, 2016.

[3]. A. K. Sharma and S. Sahni , "A Comparative Study of Classification Algorithms for Spam Email Data Analysis," International Journal on Computer Science and Engineering (IJCSE), vol. 3, no. 5, pp. 1890 - 1895, 2011.

[4]. F. Rozi and R. Kartadie, "Deteksi E-Mail Dan Spam Menggunakan Fuzzy Association Rule Mining," JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), vol. 2, no. 2, pp. 94 - 98, 2017.

Infotek : Jurnal Informatika dan Teknologi
Vol. 6 No. 1, Januari 2023
Hal. 11-20

e-ISSN 2614-8773

DOI : 10.29408/jit.v6i1.7049 Link : https://dx.doi.org/10.29408/jit.v6i1.7049

[5]. R. Y. Hayuningtyas, "Aplikasi Filtering of Spam Email Menggunakan Naïve Bayes," IJCIT (Indonesian Journal on Computer and Information Technology) , vol. 2, no. 1, pp. 53 - 60, 2017.

[6]. P. Sucháneka, F. Mareckib and R. Buckic, "Self-learning bayesian networks in diagnosis," Procedia Computer Science 35, p. 1426 – 1435 , 2014.

[7]. H. Annur, "Klasifikasi Masyarakat Miskin Menggunakan Metode Naïve Bayes," ILKOM Jurnal Ilmiah , vol. 10, no. 2, pp. 160 -165 , 2018.

[8]. A. Wibisono , S. D. Rizkiono and A. Wanto, "Filtering Spam Email Menggunakan Metode Naive Bayes," TELEFORTECH : Journal of Telematics and Information Technology, vol. 1, no. 1, pp. 9 - 17, 2020.

[9]. E. N. Putra, "Pengiriman E-Mail Spam Sebagai Kejahatan Cyber Di Indonesia," Jurnal Cakrawala Hukum, vol. 7, no. 2, pp. 169 - 182, 2016.

[10]. Y. K. Putra and M. Sadali, "Perbandingan Algoritma Naive Bayes dan Naive Bayes Berbasis PSO untuk Analisis Kredit pada PT. BPR Syariah Paokmotong," Infotek : Jurnal Informatika dan Teknologi, vol. 2, no. 2, pp. 61 - 69, 2019.

[11]. F. Mu'Alim and R. Hidaya, "mplementasi Metode Random Forest Untuk Penjurusan Siswa Di Madrasah Aliyah Negeri Sintang," Jurnal JUPITER, vol. 14, no. 1, pp. 116 - 125, 2022.

[12]. D. Andrian, M. Fachrurrozi and N. Yusliani, "Deteksi Spam Email Menggunakan Bayesian Network," ANNUAL RESEARCH SEMINAR 2016 , vol. 2, no. 1, p. 209 – 211 , 2016.

[13]. A. Mawarsih, "Pengaruh Electronic Mail Sebagai Media Komunikasi Terkadap Mengerjakan Tugas Kuliah Mahasiswa," Ejournal Ilmu Komunikasi, vol. 2, no. 1, pp. 334 - 348 , 2014.

[14]. F. Fathonah and A. Herliana, "Penerapan Text Mining Analisis Sentimen Mengenai Vaksin Covid – 19 Menggunakan Metode Naïve," Jurnal Sains dan Informatika, vol. 7, no. 2, p. 155 – 164 , 2021.

[15]. E. D. Sikumbang, "Penerapan Data Mining Penjualan Sepatu Menggunakan Metode Algoritma Apriori," Jurnal Teknik Komputer , vol. 4, no. 1, p. 156 – 161 , 2018.

[16]. I. Fathurrahman and F. , "Klasifikasi Penentuan Penerima Program Keluarga Harapan (PKH) Menggunakan Algoritma Support Vector Machine (Svm) Pada Kantor Dinas Sosial Lombok Timur," Infotek : Jurnal Informatika dan Teknologi, vol. 3, no. 1, pp. 27 - 31, 2020.