

Klasifikasi Algoritma Naïve Bayes dan SVM Berbasis PSO Dalam Memprediksi Spam Email Pada Hotline-Sapto

Meriohengki¹, Mochamad Wahyudi²

STMIK Nusa Mandiri
Program Studi Ilmu Komputer
ehengki@outlook.com

Universitas Bina Sarana Informatika
Program Studi Teknologi Informasi
wahyudi@bsi.ac.id

Abstract - Accreditation can be interpreted as an effort by the government to standardize and guarantee the quality of college alumni so that the quality of verification between universities is not too varied and in accordance with work needs. SAPTO or Online Higher Education Accreditation System is a system organized by BAN-PT for the online higher education accreditation process. Developed to improve the efficiency and quality of higher education accreditation processes. At Sapto, the University acts as an entity that submits accreditation proposals for both the Higher Education Accreditation and the Study Program Accreditation. BAN-PT has approved a complaint service related to technical issues in SAPTO that can be addressed to the hotline-sapto email account. BAN-PT has a question and answer service through the e-mail hotline sapto which can be used by universities to facilitate related to the accreditation process with sapto. Submission and questions about the accreditation process with Sapto are still responded more quickly by staff who work as public relations BAN-PT. This relates to direct question and answer questions and also a lot of time spent reading or sending spam messages as well as unwanted irrelevant questions. Email technology is also used a lot not for positive purposes so as to benefit from spam email. On this occasion the research that will be conducted is the classification of spam emails from the hotline-sapto account and preprocessing and the calculation of its accuracy, AUC with various data mining classification methods, including the Naïve Bayes algorithm, Support Vector Machine (SVM), this method is used to predict spam emails with that is the purpose of the algorithm chosen is the most accurate algorithm that can predict spam emails. From the test results obtained by the calculation of the SVM method with PSO get an accuracy value of 85.25% with AUC of 0.892

Keywords: Text Mining, Accreditation, sapto, algoritma Support Vector Machine (SVM), Naive Bayes, PSO, spam, classification

PENDAHULUAN

Akreditasi dapat diartikan sebuah upaya pemerintah untuk menstandarisasi dan menjamin mutu alumni perguruan tinggi sehingga kualitas lulusan antara perguruan tinggi tidak terlalu bervariasi dan sesuai kebutuhan kerja. Sistem Akreditasi Perguruan Tinggi Online (SAPTO) adalah sistem yang diselenggarakan BAN-PT untuk proses akreditasi perguruan tinggi secara online. Dikembangkan untuk meningkatkan efisiensi dan kualitas proses akreditasi perguruan tinggi. Data terakhir per 19 Juli 2019 yang dikeluarkan Kementerian Riset dan Teknologi) ada 124 Perguruan Tinggi Negeri (PTN) dan 3.166 Perguruan Tinggi Swasta (PTS) yang berada di bawah Dikti (PT umum). Terdapat 110 Perguruan Tinggi Agama Negeri (PTAN) dan 1066 Perguruan Tinggi Agama Swasta (PTAS) yang berada dibawah kementerian agama. Yang berada di bawah Perguruan Tinggi Kedinasan (PTK) ada 186 Perguruan Tinggi Kedinasan Negeri. Menjawab tantangan diatas pemerintah dalam hal ini BAN-PT (Badan Akreditasi Nasional Perguruan Tinggi) harus

dapat mengoptimasikan pemanfaatan kemajuan teknologi informasi dalam bentuk pelayanan kepada masyarakat khususnya perguruan tinggi. Tidak sedikit perguruan tinggi yang masih bingung untuk menyampaikan pertanyaan dan keluhan tentang teknis menggunakan sapto. Sebagai contoh, masih banyak perguruan tinggi bingung cara mendapat akun sapto, cara unggah dan unduh dokumen akreditasi serta pertanyaan yang tidak relevan seperti legalisir dan status akreditasi. Saat ini BAN-PT memiliki layanan tanya jawab melalui e-mail hotline sapto yang dapat digunakan oleh perguruan tinggi untuk menyampaikan permasalahan seputar proses akreditasi dengan sapto. Penyampaian atau pertanyaan seputar proses akreditasi dengan sapto ini masih ditanggapi lambat oleh staf yang berfungsi sebagai humas BAN-PT. Hal ini dikarenakan selain melayani tanya jawab langsung dan banyak juga waktu mereka habis untuk membaca atau menghapus pesan spam serta pertanyaan-pertanyaan yang tidak relevan yang tidak diinginkan. Teknologi e-mail banyak pula dimanfaatkan bukan untuk tujuan positif sehingga memicu peningkatan spam e-

mail. E-mail yang disalahgunakan ini biasa dikenal sebagai spam atau *junkmail* (email sampah) yang mana email tersebut berisikan iklan, penipuan dan bahkan virus (Shiela Novelia Dharma Pratiwi, 2016). Sebelumnya sudah banyak penelitian yang sama telah dilakukan untuk memprediksi e-mail spam. seperti Penelitian yang dilakukan oleh (Muhamad Abdul Ghani, 2018) tentang email spam *filtering* dengan algoritma *random forest* penelitian ini melakukan 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 pengukuran kinerja menggunakan *Confusion Matrix* dan *ROC*, diketahui bahwa algoritma *Random Forest* memiliki nilai *accuracy* paling tinggi akurasi 94,22%. Lalu penelitian yang dilakukan oleh Abd. Samad Hasan Basari dkk dengan judul *Opinion Mining Of Movie Review Using Hybrid Method Of Support Vector Machine And Particle Swarm Optimization*. Penelitian ini menggunakan metode Support Vector Machine (SVM) sebagai metode klasifikasi, dan sebagai seleksi fitur menggunakan *Case Normalization*, *Tokenization*, *Stemming*, *Generate N-Grams*, dan *Particle Swarm Optimization (PSO)* akurasi yang didapat adalah 76,20 %. Begitu juga penelitian yang dilakukan oleh Dwi Andini Putri dkk dengan judul *Algorithm Application Support Vector Machine With Genetic Algorithm Optimization Technique For Selection Features For The Analysis Of Sentiment On Twitter*. Penelitian ini menggunakan data review pizza di twitter. Penelitian ini menggunakan metode Support Vector Machine (SVM) dan *Genetic Algorithm (GA)* untuk seleksi fitur dan akurasi didapat.

Pada kasus ini merupakan data teks yang dikumpulkan isi (body) dari e-mail, maka jenis metode Data Mining yang dapat digunakan adalah Text Mining. Berdasarkan pengetahuan tentang metode diatas maka penelitian ini adalah untuk melakukan pengukuran nilai akurasi dengan menggunakan confusion matrix dan AUC dengan metode SVM dan *Naïve Bayes* serta dipadukan dengan PSO (Particle Swarm Optimization) untuk mengetahui metode manakah yang paling optimal dalam pengklasifikasian e-mail spam.

Text Mining

Text Mining adalah penambangan data teks yang digunakan untuk mendapatkan desain yang dapat merangsang dan menarik dari berbagai sumber (Tandel, Jamadar, & Dudugu, 2019). Ekstraksi informasi, pencarian informasi, peringkasan, kategorisasi, dan pengelompokan adalah berbagai teknik dari text mining (Tandel et al., 2019). Area penerapan penambangan teks yang paling populer saat ini adalah 1. analisis Sentimen adalah pengolahan bahasa alami untuk mengetahui

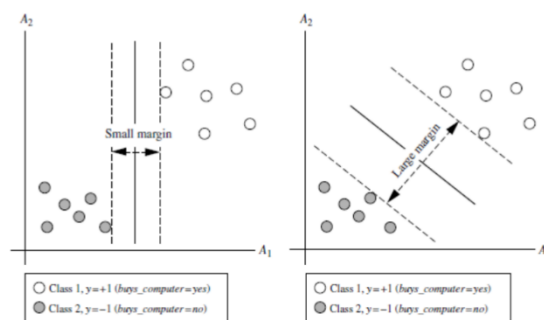
pendapat masyarakat tentang produk atau topik tertentu (Vinodhini & Chandrasekaran, 2012). 2. Topic Modeling biasa digunakan untuk melakukan tugas seperti mengeksplorasi korpus, mengklasifikasi dokumen dan mendapatkan informasi (Wang & Blei, 2011). 3. Chatbot adalah aplikasi komputer yang berinteraksi dengan pengguna yang menggunakan bahasa alami dengan cara yang mirip dengan meniru manusia (Argal et al., 2018). 4. Pengelompokan, biasanya pendekatan umum yang diterapkan untuk mempelajari parameter resolusi untuk tujuan pengelompokan grafik lokal dan global (Veldt, Gleich, & Wirth, 2019).

Support Vector Machine (SVM)

SVM merupakan metode *supervised learning* yang menganalisa data dan mengenali pola-pola yang digunakan untuk klasifikasi (Basari, Hussin, Ananta, & Zeniarja, 2013). Kelebihan yang dimiliki oleh SVM adalah dapat mengidentifikasi hyperplane terpisah yang memaksimalkan margin antara dua kelas yang berbeda (Chou, Cheng, Wu, & Pham, 2014) namun SVM memiliki kelemahan terhadap masalah penunjukan fitur dan parameter yang tidak sesuai (Basari, Hussin, Ananta, & Zeniarja, 2013). Algoritma ini bekerja dengan menggunakan pemetaan nonlinear untuk mengubah data training ke dimensi yang lebih tinggi. Dalam dimensi baru ini, akan mencari garis pemisah (hyperplane) optimal linear (yaitu, batas keputusan memisahkan tupel dari satu kelas dari yang lain).

Dengan pemetaan nonlinear yang tepat ke dimensi yang cukup tinggi, data dari dua kelas selalu dapat dipisahkan dengan garis (F. Anugrah 2017).

Sumber: (Han and Kamber, 2000)



Naïve Bayes (NB)

Definisi yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Bustami, 2014). *Naïve Bayes* adalah pengklasifikasi probabilistik yang didasarkan pada teori Bayesian dengan asumsi masing-masing fitur kelas tertentu tidak bergantung pada fitur lainnya (Sharmila & Geethanjali, 2016).

Persamaan dari teorema *Naïve Bayes* adalah:

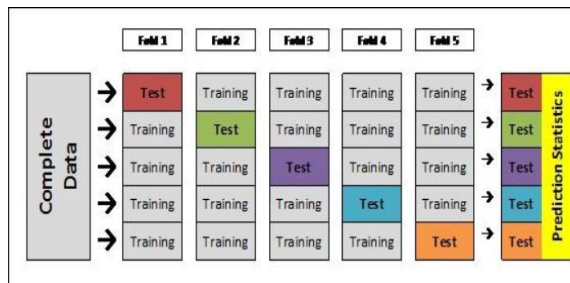
$$P(H|X) = \frac{P(H|X)P(H)}{P(X)}$$

Keterangan:

- X : Data dengan class yang belum diketahui
- H : Hipotesis data yang merupakan suatu class spesifik
- P(H|X) : Probabilitas hipotesis H berdasarkan kondisi X (posteriori probabilitas)
- P(H) : Probabilitas hipotesis H (prior probabilitas)
- P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H
- P(X) : Probabilitas X

k- Fold Cross Validation

k-Fold Cross Validation merupakan sebuah Teknik validasi dengan cara membagi data secara acak ke dalam k bagian dan masing-masing bagian akan dilakukan proses klasifikasi (Han, Kamber, & Pei, 2012). Dengan menggunakan *k-Fold Cross Validation* akan dilakukan percobaan sebanyak k. Percobaan dilakukan menggunakan satu data *testing* (data uji) k-1 bagian akan menjadi data *training* (data latih) sehingga untuk tiap percobaan akan didapat data *testing* yang berbeda-beda. Cross Validation yang memiliki akurasi paling baik menggunakan 10 *fold cross validation* (Witten, et al, 2011).



Gambar 1 tabel *k-fold cross validation*

Confusion Matrix

Confusion matrix merupakan alat untuk menganalisa serta mengukur seberapa baik pengklasifikasian dengan model yang digunakan. Evaluasi menggunakan confusion matrix akan menghasilkan nilai *accuracy*, *precision* dan *recall*. *Confusion matrix* berisi informasi aktual (*actual*) dan prediksi (*predicted*) pada model klasifikasi.

		Actual	
		True	False
Predicted	T	True Positive (TP)	False Negative (FN)
	F	False Positive (FP)	True Negative (TN)

Gambar 2: *Confusion Matrix*

Sumber: (Gorunescu, 2011)

Formulasi perhitungan Confusion Matrix:

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

$$Sensitivity = \frac{TP}{TP + FN}$$

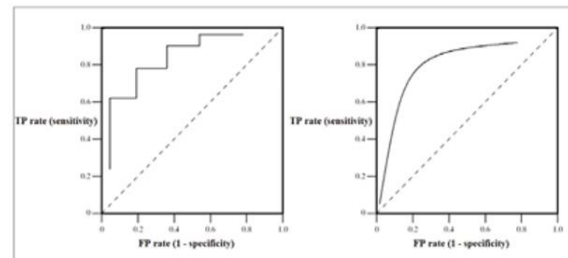
$$Specificity = \frac{TN}{TN + FP}$$

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

Curve ROC

Fungsi kurva ROC adalah untuk menunjukkan akurasi serta membandingkan klasifikasi secara visual atau grafik. Secara teknik, kurva ROC juga disebut grafik ROC, dua dimensi grafik yaitu TP rate diletakan pada sumbu Y sedangkan FP rate diletakan pada sumbu X. Grafik ROC menggambarkan *trade-off* antara manfaat (*true positives*) dan biaya (*false positives*). Berikut tampilan dua jenis kurva ROC (*discrete dan continuous*).



Gambar 3: Kurva ROC

Sumber: Gorunescu, 2011

Pada gambar 3 garis diagonal membagi ruang ROC, yaitu:

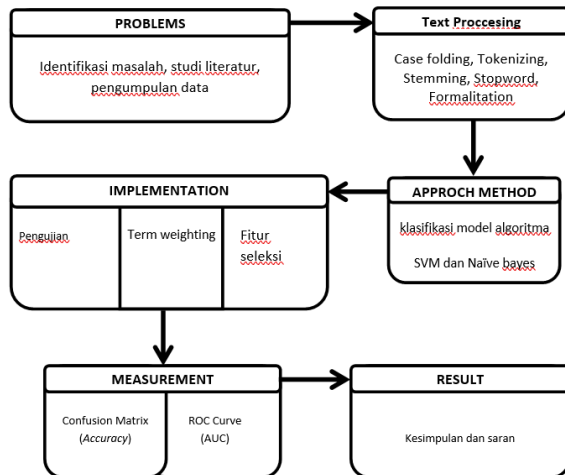
1. (a) poin diatas garis diagonal merupakan hasil klasifikasi yang baik
2. (b) poin dibawah garis diagonal merupakan hasil klasifikasi yang buruk

Dapat disimpulkan bahwa, satu poin pada kurva ROC adalah lebih baik dari pada yang lainnya jika arah garis melintang dari kiri bawah ke kanan atas didalam grafik. Tingkat akurasi dapat di diagnosa sebagai berikut (Gorunescu, 2011):

Nilai AUC	Klasifikasi
0.90-1.00	Sangat Baik
0.80-0.90	Baik
0.70-0.80	Cukup
0.60-0.70	Buruk
0.50-0.60	Salah

METODOLOGI PENELITIAN

Pada Gambar 4 dapat dilihat diagram alur terkait tahapan yang akan dilakukan pada penelitian ini



Gambar 4 alur metode penelitian

Pengumpulan Data

Pada penelitian ini menggunakan beberapa data yang harus disiapkan terlebih dahulu. Data tersebut terdiri dari data utama berupa content email yang didapat dari akun email hotlinesapto dan data pendukung seperti data kata dasar Bahasa Indonesia. Data content email akun hotlinesapto dimulai pada tahun 2017. Data disalin dalam bentuk spreadsheet dan terkumpul sebanyak 1700 data. Sedangkan data pendukung didapatkan dari buku, jurnal dan publikasi lainnya.

Pengelompokan Data

Tahap pengelompokan email digunakan untuk mengelompokan dokumen yang telah terkumpul kedalam kelas spam atau non spam sesuai dengan sentimen yang terkandung pada email. Pengelompokan ini dilakukan oleh bagian yang biasa menjawab email tersebut dan dilakukan secara manual berdasarkan jenis seperti pada gambar berikut;

Class	Text
Not Spam	Mohon informasi mengapa pada saat saya ikuti link untuk Ganti passwordnya selalu gagal atau ditolak.Terima kasih sebelumnya..
Not Spam	Salam Hormat, Mohon petunjuk untuk mendapatkan akun sapto
Spam	Selamat siang, Saya ingin bertanya terkait perguruan tinggi swasta tempat kuliah saya yang sudah tidak ada. Bagaimana cara saya mendapatscan sertifikat akreditasi PTS tersebut?
Spam	I'm glad to get in touch with you here Our factory is located in shandong province, the hometown of Confucius and mencius.The company mainly produces granules, emulsions, suspensions and microcapsules of various pesticides. If you need, please contact us, and hope to cooperate with your company.

Gambar 5 pesan email berdasarkan kelas

Term Weighting

Data yang sudah melalui preprocessing selanjutnya harus berbentuk numerik. Dalam penelitian ini mengubah teks yang sudah berupa token kedalam numerik menggunakan pembobotan *Term Frequency Invert Document Frequency* (TF-IDF). Pada penelitian ini algoritma pembobotan yang digunakan adalah menggunakan TF-IDF. Proses perhitungan TF-IDF memanfaatkan tools rapidminer 9.3

$$w_{x,y} = tf_{x,y} \times \log \left(\frac{N}{df_x} \right)$$

Dimana:

- tf_{x,y} = Jumlah kemunculan x dalam y
- df_x = Jumlah dokumen yang mengandung x
- N = Jumlah total dokumen

HASIL DAN PEMBAHASAN

Business Understanding

Pemahaman mengenai objek penelitian dilakukan dengan menggali informasi melalui akun hotline-sapto dan melihat folder kotak masuk periode September 2018 sampai dengan Agustus 2019. Konsentrasi yang dilakukan pada fase ini mencari isi (content email). Pada tahap ini juga dilakukan pemahaman untuk mencari metode klasifikasi yang terbaik agar dapat membantu pada saat proses pengolahan data yang akan dilakukan dengan cara membandingkan hasil dari algoritma yang digunakan dan untuk meningkatkan akurasi performa dari metode klasifikasi.

Data Understanding

Tahap data understanding ini, dilakukan penyalinan data content atau body email secara manual yang diperoleh dari akun hotline-sapto. Data awal ini diekstraksi ke dalam bentuk spreadsheet dokumen. Jumlah data awal yang diperoleh sebanyak 1756 data. Selanjutnya proses cleansing dilakukan sehingga didapat dataset berjumlah 1700. Dataset tersebut akan dibagi menjadi 10 bagian dan akan dilakukan pengulangan sebanyak 10 pengulangan. Pengambilan data tersebut dilakukan secara acak agar semua data dapat menjadi data training juga menjadi data testing.

Text Preprocessing

Pada tahap ini data yang sudah dikumpulkan dan dikelompokan sebelumnya perlu melalui tahap preprocessing terlebih dahulu agar teks atau dokumen yang akan di klasifikasikan sesuai dengan keperluan penelitian. Didalam penelitian ini ada beberapa tahap preprocessing yang akan dilakukan dimana tahap pertama adalah menghapus format atau bentuk URL dan email contoh simbol "@", "#".

Tahap selanjutnya terdiri dari:

1. **Formalisasi**
 Menghapus karakter-karakter yang salah dalam penulisannya seperti "ackreditasi", "admistrasi" dan juga masih ditemukan beberapa kata mengandung bahasa informat dan belum sesuai dengan EYD.
2. **Case folding**
 Merubah huruf kapital di setiap dokumen teks menjadi huruf kecil agar mengurangi keragaman karakter serta menghapus karakter selain huruf alphabet a-z.
3. **Tokenizing**
 Tahapan dimana semua kalimat dipecah menjadi token dengan menggunakan white space serta menghilangkan tanda baca jika terdapat tanda baca ataupun simbol selain huruf.

Input	tokenizing
Mohon informasi mengapa pada saat saya ikuti link ganti passwordnya selalu gagal atau ditolak.	mohon informasi mengapa pada saat saya ikuti link ganti passwordnya selalu gagal atau ditolak

Gambar proses Tokenizing

4. **Stopword**
 Tahap *stopword* adalah merupakan proses menghilangkan kata-kata umum yang sering muncul dan yang tidak berkontribusi banyak pada isi dokumen.

Input	Stopword
mohon informasi mengapa pada saat saya ikuti link ganti passwordnya selalu gagal atau ditolak	mohon Informasi saat ikuti link ganti passwordnya selalu gagal ditolak

5. **Stemming**
 Pada tahap ini sebelum dilakukan proses stemming, peneliti harus terlebih dahulu mempunyai kamus kata dasar bahasa Indonesia. Penelitian ini menggunakan sastrawi stemmer yaitu sebuah library stemmer bahasa indonesia yang sudah dipublikasi di internet.

Input	Stemming
mohon Informasi saat ikuti link ganti passwordnya selalu gagal tolak	mohon Informasi saat ikut ganti password selalu gagal tolak

Model proses preprocessing yang dilakukan dengan Rapidminer 9.3 adalah sebagai berikut:

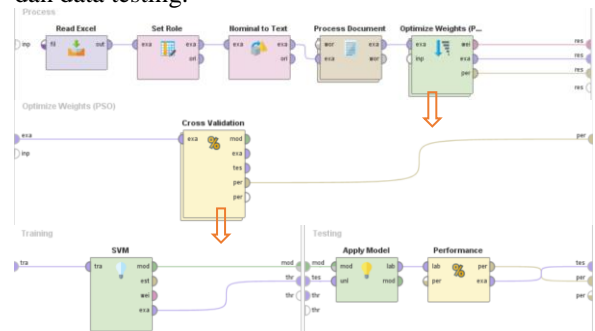


Tahapan Pemodelan

Pada fase ini secara langsung melibatkan teknik data mining yaitu dengan cara melakukan pemilihan teknik data mining dan menetapkan algoritma yang akan digunakan. Tools yang digunakan pada fase pemodelan ini adalah Rapidminer versi 9.3.

Pengujian Model SVM berbasis PSO

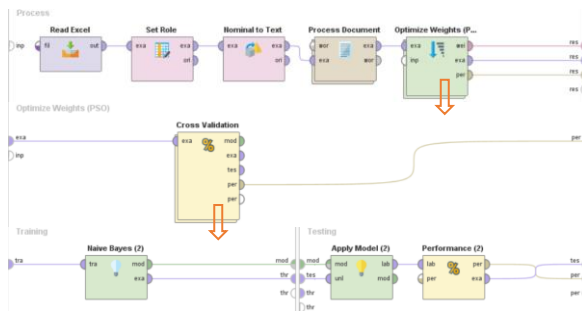
Pada pengujian ini, data digunakan adalah data bersih yang telah melalui preprocessing. Data tersebut diambil dari file process documents, hal ini dilakukan karena dataset disimpan dalam bentuk text document. Process validasi terdiri dari data training dan data testing.



Gambar 6: Algoritma Support Vector Machine dan PSO

Pengujian Model Naïve Bayes berbasis PSO

Model yang diusulkan dalam *Naïve Bayes* berbasis *Partical Swarm Optimization*



Gambar 7: Algoritma Naïve Bayes dan PSO

Evaluasi dan Validasi

Tahapan 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 Kurva ROC dan Confusion Matrix dari masing-masing algoritma.

Nilai Akurasi SVM berbasis PSO

Tabel 7: Akurasi SVM berbasis PSO

accuracy: 84.59% +/- 4.70% (micro average: 84.59%)

	true Spam	true Not Spam	class precision
pred. Spam	167	103	61.85%
pred. Not Spam	159	1271	88.88%
class recall	51.23%	92.50%	

Perhitungan Confusion Matrix;

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{167+1271}{167+1271+159+103} = 0,845$$

$$Sensitivity = \frac{TP}{TP+FN} = \frac{167}{167+103} = 0,61$$

$$Specificity = \frac{TN}{TN+FP} = \frac{1271}{1271+159} = 0,88$$

$$PPV = \frac{TP}{TP+FP} = \frac{167}{167+159} = 0,51$$

$$NPV = \frac{TN}{TN+FN} = \frac{1271}{1271+103} = 0,92$$

Akurasi yang diperoleh yaitu 84.59 % dimana dari 167 data diklasifikasikan positif sesuai dengan prediksi dan 103 data diprediksi positif tetapi ternyata hasil prediksi negatif, Kemudian dari 1271 data diklasifikasikan negatif diprediksi sesuai, dan 159 data diprediksi negatif ternyata hasil prediksinya positif.

Nilai Akurasi Naïve Bayes berbasis PSO

Tabel 8: Akurasi Naïve Bayes berbasis PSO

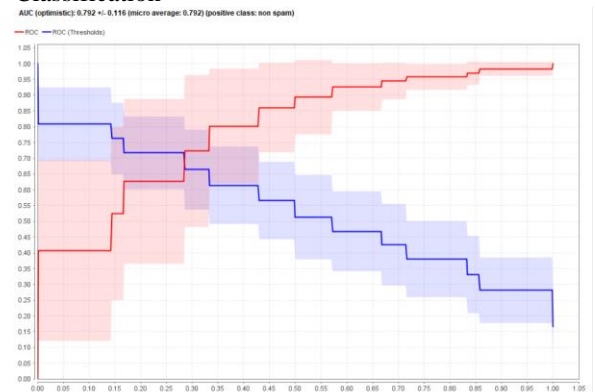
accuracy: 81.24% +/- 1.97% (micro average: 81.24%)

	true Spam	true Not Spam	class precision
pred. Spam	202	195	50.88%
pred. Not Spam	124	1179	90.48%
class recall	61.96%	85.81%	

Akurasi yang diperoleh yaitu 81.24% dimana dari 202 data diklasifikasikan positif sesuai dengan prediksi dan 195 data diprediksi positif tetapi ternyata hasil prediksi negatif, Kemudian dari 1179 data diklasifikasikan negatif diprediksi sesuai, dan 124 data diprediksi negatif ternyata hasil prediksinya positif.

Kurva ROC SVM berbasis PSO

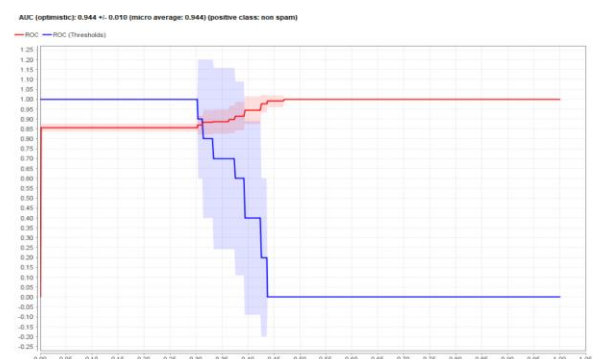
Kurva ROC dengan nilai AUC (Area Under Curve) yang dihasilkan dari gambar 4.13 diatas sebesar 0.892 dimana diagnosa hasilnya adalah Good Classification



Gambar 4.13 Kurva ROC SVM Berbasis PSO

Kurva ROC Naïve Bayes berbasis PSO

Kurva ROC dengan nilai AUC (Area Under Curve) yang dihasilkan dari gambar 4.14 diatas sebesar 0.948 dimana diagnosa hasilnya adalah Good Classification



Gambar 4.14 Kurva ROC Naïve Bayes Berbasis PSO

Perbandingan Accuracy, Precision, Recall dan AUC

Hasil komparasi akurasi dan AUC algoritma yang telah digunakan sebagai berikut:

Tabel 9: Hasil komparasi

Algoritma	Accuracy	Precisison	Recall	AUC
SVM	84.59%	89.05%	92.51%	0.792
NB	80.59%	89.59%	96.02%	0.942
SVM+ PSO	85.25%	85.19%	99.05%	0.892
NB+ PSO	81.24%	90.49%	85.81%	0.948

KESIMPULAN

Dalam penelitian ini dilakukan pengujian model dengan menggunakan Support Vector Machines (SVM) dan Support Vector Machines (SVM) berbasis Particle Swarm Optimization (PSO) serta menggunakan Naïve Bayes dan Naïve bayes berbasis Particle Swarm Optimization (PSO) dengan menggunakan data body email hotline sapto yang di olah dengan tools rapidminier 9.3 dan untuk text preprocessing menggunakan gataframework. Telah diketahui dengan model di atas didapat hasil pengujian nilai accuracy, precision dan recall dan AUC dari setiap model. Support Vector Machines (SVM) berbasis Particle Swarm Optimization (PSO) didapat nilai accuracy 85.25% dengan nilai AUC 0.892. Untuk pengujian Support Vector Machines (SVM) nilai accuracy 84.59% dengan nilai AUC 0.792. Sedangkan pengujian model Naïve Bayes berbasis Particle Swarm Optimization (PSO) didapat nilai accuracy 81.24% dengan nilai AUC 0.892 dan untuk pengujian model Naïve Bayes didapat accuracy 80.59% dengan nilai AUC 0.942. Meski demikian dapat disimpulkan pengujian email spam pada hotline sapto telah diketahui dengan menggunakan model Support Vector Machines (SVM) berbasis Particle Swarm Optimization (PSO) lebih baik daripada model yang lainnya. Model Support Vector Machines (SVM) berbasis Particle Swarm Optimization (PSO) dapat memberikan pemecahan untuk permasalahan klasifikasi spam email pada hotline sapto.

REFERENSI

Argal, A., Gupta, S., Modi, A., Pandey, P., Shim, S., & Choo, C. (2018). Intelligent travel chatbot for predictive recommendation in echo platform. *2018 IEEE 8th Annual Computing and Communication Workshop and Conference, CCWC 2018*. <https://doi.org/10.1109/CCWC.2018.8301732>

Basari, A. S. H., Hussin, B., Ananta, I. G. P., & Zeniarja, J. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. *Procedia Engineering*. <https://doi.org/10.1016/j.proeng.2013.02.059>

Bustami. (2014). Penerapan Algoritma Naive Bayes. *Jurnal Informatika*.

Chou, J. S., Cheng, M. Y., Wu, Y. W., & Pham, A. D. (2014). Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification. *Expert Systems with Applications*. <https://doi.org/10.1016/j.eswa.2013.12.035>

Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques*.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques. In Data Mining: Concepts and Techniques*. <https://doi.org/10.1016/C2009-0-61819-5>

Muhamad Abdul Ghani, A. S. (2018). *Email Spam Filtering Dengan Algoritma Random Forest*.

Sharmila, A., & Geethanjali, P. (2016). DWT Based Detection of Epileptic Seizure from EEG Signals Using Naive Bayes and k-NN Classifiers. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2016.2585661>

Shiela Novelia Dharma Pratiwi, B. S. S. U. (2016). *Klasifikasi EmailSpam dengan Menggunakan Metode Support Vector Machinedan k-Nearest Neighbor*. 344–349.

Tandel, S. S., Jamadar, A., & Dudugu, S. (2019). A Survey on Text Mining Techniques. *2019 5th International Conference on Advanced Computing and Communication Systems, ICACCS 2019*. <https://doi.org/10.1109/ICACCS.2019.8728547>

Veldt, N., Gleich, D. F., & Wirth, A. (2019). Learning resolution parameters for graph clustering. *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*. <https://doi.org/10.1145/3308558.3313471>

Vinodhini, G., & Chandrasekaran, R. (2012). International Journal of Advanced Research in Computer Science and Software Engineering Sentiment Analysis and Opinion Mining: A Survey. *International Journal of Advanced Research in Computer Science and Software Engineering*.

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. <https://doi.org/10.1145/2020408.2020480>