Perbandingan Metode Random Forest dan Naïve Bayes dalam *Email Spam Filtering*

Maria Anita[1, a)], Bambang Susanto[2, b)], Lennox Larwuy[3, c)]

[1,2.3] Program Studi Matematika, Universitas Kristen Satya Wacana Salatiga,
Jalan Diponegoro 52-60 Salatiga, 50711

[a)]*email: mariaanita001.am@gmail.com*
[b)] *email: bambang.susanto@uksw.edu*
[c)] *email: lennox.larwuy@uksw.edu*

Abstrak

Email merupakan salah satu alat yang penting untuk berkomunikasi, mentransfer *file* serta dapat digunakan untuk media iklan melalui internet. Penggunaan *email* semakin meningkat maka banyak pihak yang mengirimkan *email* dengan pesan yang berisikan virus, penipuan, bahkan pornografi. *Email* seperti inilah yang disebut *spam email* dimana *email* yang tidak diinginkan oleh penerima yang dikirim secara massal. Banyak pengguna *email* merasa terganggu karena banyaknya waktu yang dihabiskan untuk menghapus satu per satu pesan *spam.* Dalam penelitian ini dilakukan perbandingan metode klasifikasi *Random Forest* dan Naïve Bayes dalam memprediksi *spam email* dengan tujuan melihat metode mana yang paling akurat. Data yang digunakan dalam penelitian ini adalah dataset *email* berjumlah 2605 data dengan dua variabel yaitu variabel *body* (yang menunjukkan isi dari *email)* dan variabel label (yang menunjukkan pelabelan) dimana 1 menunjukkan *spam* dan 0 menunjukkan bukan *spam.* Dari hasil pengujian menggunakan *confusion matrix* diketahui metode *random forest* memiliki nilai *accuracy* paling tinggi yaitu 98%, dan Naïve bayes 73%.

Kata kunci: Email, Klasifikasi, Naïve Bayes, Random Forest, Spam

Abstract

Email is an important tool not only for communicating and transferring files but also it can be used for advertising media over the Internet. Since the increase in email user numbers, many users send viruses, fraud, and even pornography contained emails. Those kinds of emails were called spam, where unexpected emails sent in bulk. Many email users are annoyed by the amount of time spent deleting individual spam messages. This study provides a comparison between the Random Forest and Naïve Bayes classification methods for email spam predicting. It aims for searching the most accurate method. The data used in this study is an email dataset totaling 2607 data with two variables, namely the body variable (which shows the contents of the email) and the label variable (which shows labeling) where 1 indicates spam and 0 indicates not spam. From the test result using the confusion matrix, it is known that the random forest method has the highest accuracy value, namely 98%, and Naïve Bayes 73%.

Keywords: Classification, Email, Naïve Bayes, Random Forest, Spam,

Pendahuluan

Pada era globalisasi sekarang ini, terdapat berbagi macam media informasi yang berkembang salah satunya adalah *email. Email* merupakan salah satu sarana dalam mengirimkan informasi dengan cepat melalui PC, dan gadget elektronik lainnya. Dengan meningkatnya pengguna *email* juga telah menyebabkan peningkatan jumlah *email spam.* Peningkatan pengguna *email* juga telah memikat beberapa pihak tertentu untuk membombardir *email* dengan pesan yang tidak diminta dengan tujuan tertentu, contohnya seperti penawaran barang dan jasa, promosi/iklan, penipuan serta pesan yang tidak layak. Pesan *spam* yang masuk ke kotak masuk sangat terasa karena dengan banyaknya pesan *spam* yang masuk ke *email* dapat menambah memori penyimpanan yang terpakai sehingga menyebabkan *gadget* kita semakin berat. Hal ini, dapat merepotkan pengguna *email* karena tidak dapat membedakan *email* mana yang benar-benar penting dengan *email spam.* Karena banyaknya pesan *spam* yang masuk maka, sangat tidak memungkinkan pengguna *email* menghapusnya secara satu per satu. Saat menerima *email spam* yang tidak diminta, banyak pengguna yang merasa sangat menjengkelkan.

Hasil dari penelitian sebelumya yang dilakukan oleh Renuka & Hamsapriya [1] menggunakan pengaplikasian *word stemming* akurasi dari metode Naïve Bayesian dapat ditingkatkan sampai 5-10% dimana metode naïve bayes lebih baik daripada metode *random forest.* Sedangkan penelitian menggunakan metode *random forest* yang dilakukan oleh Akinyelu & Adewumi [2] mendapatkan nilai presisi yang cukup tinggi dibanding metode naïve bayes dan tingkat kesalahan yang rendah. Oleh karena itu, pada penelitian ini penulis ingin mendapatkan hasil perbandingan klasifikasi mana yang lebih baik antara metode random forest atau naïve bayes yang digunakan berdasarkan tingkat akurasinya.

Metode

Metode yang digunakan dalam penelitian ini adalah metode random forest dan Naïve Bayes dalam menyaring email spam. Kedua metode tersebut akan dibandingkan efektifitasnya dalam menyaring email spam dengan data yang tersedia. Oleh karena itu terlebih dahulu diperkenalkan beberapa teori sebagai berikut.

1. *Email*

Email merupakan sebuah metode untuk mengirim, menerima, dan menyimpan pesan melalui jaringan internet.

Menurut Hayuningtyas [3], *email* terdiri dari 3 komponen, yaitu:

1) *Envelope*
   Proses ini digunakan oleh *Mail Transport Agent* (MTA) untuk melihat rute atau jalur pesan. Biasanya pengguna tidak melihat bagian ini karena terjadi pada bagian MTA untuk pengiriman.
2) *Header*
   Header berisi informasi mengenai *email* tersebut, mulai dari *from* (alamat *email* yang diikuti nama pengirim), *to* (alamat *email* yang diikuti oleh nama penerima), *subject* (Rangkuman isi pesan), dan *Date* (waktu dan tanggal saat pesan dikirim).
3) *Body*
   Body merupakan isi pesan dari pengirim ke penerima dapat berupa *text*, maupun *file.*

2. *Spam*

Spam atau *junk mail* adalah penyalahgunaan dalam sistem pesan elektronik untuk mengirim promosi produk, atau jasa, pornografi, virus, dan hal lain yang tidak penting yang secara massal ke ribuan pengguna *email*. Spam pertama kali ada pada Mei tahun 1978 berisi iklan tentang *product DecSystem-20* yang dikirimkan oleh *Digital Equipment Corporation* (DEC).

Menurut Defiyanti & D. L. Crispina Pardede [4] ada beberapa pengklasifikasian berdasarkan karakteristik dari *spam,* yaitu:

1) Alamat pengirim yang tidak benar
2) Pemalsuan *header email* untuk menyembunyikan *email* yang sesungguhnya menjadi sulit untuk ditetapkan sebagai *spam* atau *nonspam*
3) Indentitas penerima tidak nyata
4) Alamat *email* yang berada pada "*To"* mempunyai variasi *email* penerima
5) Isi *subject* tidak berhubungan dengan isi *email*
6) Isi *email* mempunyai sifat keragu-raguan
7) *Unsubribe* tidak bekerja
8) Mengandung *script* tersembunyi.

3. *Klasifikasi*

Dalam analisis data klasifikasi merupakan proses untuk menemukan model atau fungsi yang akan menjelaskan atau membedakan konsep kelas data, yang bertujuan untuk dapat memperkirakan kelas dari suatu objek [5]. Ada dua tahapan penting dalam klasifikasi, yaitu:

1) Tahap *training* yaitu tahapan pembelajaran menggunakan data *training* dimana data yang digunakan oleh algoritma untuk membentuk sebuah model klasifikasi.
2) Tahap *testing* yaitu tahapan dimana menguji metode menggunakan data *testing.*

Menurut Faisal, Reza M; Nugrahadi [6] pengukuran kinerja algoritma yang umum yang dapat digunakan salah satunya yaitu:

4. *Confusion matrix*

Confusion matrix adalah salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi.

Tabel 1. *Confusion Matrix*

		Respon (pengamatan)	
		Ada : $Y = 1$	Tidak ada : $Y = 0$
Prediksi Respon	Ada $\hat{Y} = 1$	Positif Benar n_{TP}	Positif Palsu n_{FP}
	Tidak ada $\hat{Y} = 0$	Negatif Palsu n_{FN}	Negatif Benar n_{TN}

Dimana,

n_{TP} (*True Positives)* : banyaknya subjek yang terdeteksi positif benar, artinya baik hasil maupun yang sebenarnya sama-sama positif.

n_{FP} (*True Negatives)* : banyaknya subjek yang terdeteksi positif palsu, artinya hasilnya menunjukkan positif sedangkan seharusnya negatif.

n_{FN} *(False Positives)* : banyaknya subjek yang terdeteksi negatif palsu, artinya hasilnya menunjukkan negatif sedangkan seharusnya positif.

n_{TN} *(False Negatives)* : banyaknya subjek yang terdeteksi negatif benar, artinya baik hasil maupun yang sebenarnya sama-sama menunjukkan negatif

Menurut Anggana [7] *confusion matrix* seperti diperlihatkan pada tabel 1 sering digunakan untuk menghitung tingkat akurasi. Setelah mendapatkan hasil dari *confusion matrix,* juga dapat dilakukan perhitungan untuk:

1) *Recall,* merupakan perbandingan subjek yang memiliki hasil positif benar dengan jumlah hasil positif benar dan negatif palsu.

$$recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (1)$$

2) *Precision,* merupakan perbandingan subjek yang memiliki hasil positif benar dengan jumlah hasil positif benar dan positif palsu.

$$precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (2)$$

3) *Accuracy*, merupakan perbandingan dari subjek yang diindentifikasi benar dengan jumlah semua subjek yang digunakan.

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

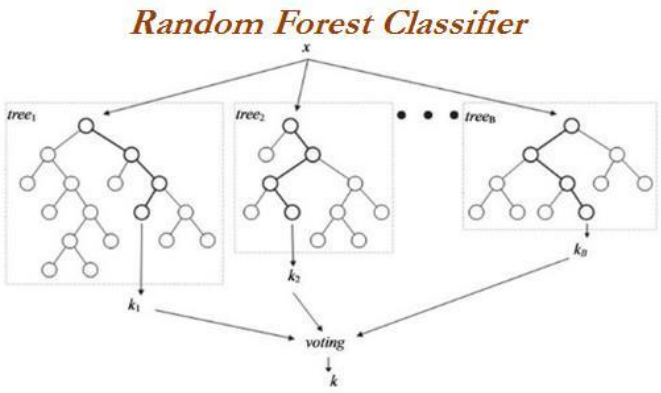
$$f1\ score = \frac{2*precision*recall}{precision+recall} \quad (4)$$

4) Laju error, merupakan perbandingan dari subjek yang diidentifikasi salah dengan jumlah dari seluruh objek.

$$error = \frac{\mathrm{FP+FN}}{\mathrm{TP+FP+FN+TN}} \quad (5)$$

5. *Random Forest*

Random Forest adalah salah satu algoritma *Supervised Learning* yang dikeluarkan oleh Breiman pada tahun 2001, dan biasanya digunakan untuk menyelesaikan masalah yang berhubungan dengan klasifikasi, regresi, dan lainnya [8][9]. Menurut Samudra [10] *Random Forest* merupakan pengembangan dari metode *Classification and Regression Tree* (CART) yang menerapkan metode *bagging* atau *bootstrap* dan *random failure selection. Bagging* merupakan metode yang dapat memperbaiki hasil dari algoritma klasifikasi. Cara kerja *random forest* dalam klasifikasi ditunjukkan seperti gambar di bawah ini.



Gambar 1. *Random Forest Classifier*

Berdasarkan gambar 1, setiap pohon ditanam secara terpisah, kemudian diambil secara acak dari setiap sampel. Kemudian, di akhir dilakukan voting untuk menentukan kelas yang terakhir. Karena menggunakan *ensemble* dari *decision tree, random forest* tidak dapat menentukan signifikansi dari setiap variabel, hanya dapat menunjuukan tingkat kepentingan variabel saja.

6. *Naïve bayes*

Menurut Syarli & Muin [11] Naïve Bayes merupakan salah satu dari algoritma pembelajaran yang efektif dan efisien untuk *machine learning* dan data mining. Persamaan dari Naïve Bayes sendiri adalah:

$$P(H|X) = \frac{P(H|X)\,P(H)}{P(X)} \tag{6}$$

dimana,

X : Data dengan *class* yang belum diketahui
H : Hipotesis data yang merupakan suatu *class* yang spesifik
$P(H|X)$` : Probabilitas hipotesis H berdasarkan kondisi X
$P(H)$: Probabilitas Hipotesis H (prior hipotesis)
$P(X/H)$: Probabilitas X berdasarkan kondisi pada hipotesis H
$P(X)$: Probabilitas X

Hasil dan Diskusi

Data yang digunakan dalam penelitian ini adalah *dataset email* yang berasal dari *Kaggle (https://www.kaggle.com/nitishabharathi/email-spam-dataset)* data sudah dalam bentuk file bertipe csv. Data ini mengandung 2605 data dengan 2 variabel yaitu variabel *body* (yang menunjukkan isi dari *email)* dan variabel label (yang menunjukkan pelabelan) dimana 1 menunjukkan *spam* dan 0 menunjukkan bukan *spam*. Data dipisahkan menjadi data *training* sebanyak 75% dari data asli dan data *testing* 25% dari data asli.

Pengolahan data untuk Random Forest dan Naïve Bayes dilakukan menggunakan software Rstudio menggunakan *library* dengan beberapa *command* diantaranya:

```
>library(plyr)
>library(tidyr)
>library(dplyr)
>library(randomForest)
```

```
>library(wordcloud)
>library (tm)
#input data
>library(readr)
>data<-read_csv("E:/MARIA/Skripsi/Spam.csv")
>View(data)
>data=data[,-1]
```

variabel label diubah kedalam bentuk *factor level* agar dapat dianalisis**.** Menggunakan *command* sebagai berikut

```
>dataset$Class=data$Label
>str(dataset$Class)
>set.seed(222)
>split=sample(2,nrow(dataset),prob= c(0.75,0.25),replace = TRUE)
>train_set = dataset[split == 1,]
>test_set = dataset[split == 2,]
>prop.table(table(train_set$Class))
```

Disini kita peroleh hasil *email* label 1 (Spam) memiliki proporsi 0.1662188 dan untuk email label 0 (Ham) memiliki proporsi 0.8337812. Setelah mendapatkan proporsi dari tiap label kita lakukan analisis untuk kedua metode yang kita gunakan dengan *command* sebagai berikut:

```
#analisis Random Forest dan Naïve Bayes
>rf_classifier = randomForest(x = train_set[-1210], y = train_set$Class, ntree = 500)
>rf_classifier
>library(e1071)
>control <- trainControl(method="repeatedcv", number=10, repeats=3)
>system.time( classifier_nb <- naiveBayes(train_set, train_set$Class, laplace = 1,
>trControl = control,tuneLength = 7) )
>confusionMatrix(nb_pred,test_set$Class)
```

Hasil uji yang diperoleh disajikan sebagai berikut:

1. Hasil Random Forest

```
rf_pred   0   1
      0 538   7
      1   0  94

               Accuracy : 0.989
                 95% CI : (0.9776, 0.9956)
    No Information Rate : 0.8419
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9576

 Mcnemar's Test P-Value : 0.02334

            Sensitivity : 1.0000
            Specificity : 0.9307
         Pos Pred Value : 0.9872
         Neg Pred Value : 1.0000
             Prevalence : 0.8419
         Detection Rate : 0.8419
   Detection Prevalence : 0.8529
      Balanced Accuracy : 0.9653

       'Positive' Class : 0
```

Gambar 2. Hasil *confusion matrix* dari *Random Forest*

Berdasarkan confusion matrix di Gambar 2, dapat diketahui bahwa jumlah *email* yang terdeteksi sebagai *spam* dan benar *spam* sebanyak 94. *email* yang terdeteksi sebagai *spam* namun sebenarnya bukan *spam* sebanyak 7, Untuk yang terdeteksi bukan *spam* namun sebenarnya *spam* sebanyak 0, sedangkan *email* yang benar-benar bukan *spam* adalah 538Dilakukan juga hasil perhitungan berdasarkan *confusion matrix* pada gambar 2

$$recall = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{538}{538 + 7} = 0.987$$

$$precision = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{538}{538 + 0} = 1$$

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} = \frac{538 + 94}{538 + 7 + 0 + 94} = \frac{632}{639} = 0.989$$

$$f1\ score = \frac{2 * precision * recall}{precision + recall} = \frac{1.96}{1.98} = 0.989$$

$$error = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} = \frac{7}{538 + 7 + 0 + 94} = \frac{7}{639} = 0.0109$$

2. Hasil Naïve Bayes

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 370   0
         1 168 101

               Accuracy : 0.7371
                 95% CI : (0.7011, 0.7708)
    No Information Rate : 0.8419
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4105

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6877
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.3755
             Prevalence : 0.8419
         Detection Rate : 0.5790
   Detection Prevalence : 0.5790
      Balanced Accuracy : 0.8439

       'Positive' Class : 0
```

Gambar 3. Hasil *Confusion Matrix* dari Naïve Bayes

Berdasarkan confusion matrix pada Gambar 3, dapat diketahui bahwa jumlah *email* yang terdeteksi sebagai *spam* dan benar *spam* sebanyak 101. *email* yang terdeteksi sebagai *spam* namun sebenarnya bukan *spam* sebanyak 0, Untuk yang terdeteksi bukan *spam* namun sebenarnya *spam* sebanyak 168, sedangkan *email* yang benar-benar bukan *spam* adalah 370. Dilakukan juga hasil perhitungan berdasarkan *confusion matrix* Gambar 3

$$recall = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{370}{370 + 0} = 1$$

$$precision = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{370}{370 + 168} = 0.6877$$

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} = \frac{370 + 101}{370 + 0 + 101 + 168 + 101} = \frac{471}{639} = 0.737$$

$$f1\ score = \frac{2 * precision * recall}{precision + recall} = \frac{1.3754}{1.6877} = 0.81$$

$$error = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} = \frac{168}{370 + 0 + 101 + 168} = \frac{168}{639} = 0.263$$

Hasil Perbandingan 2 Metode

Tabel 2. Hasil Perbandingan Confusion Matrix

	Random Forest	Naïve Bayes
Presicion	1	0.6877
Recall	0.98	1
F1-score	*0.989*	0.81
accuracy	*0.989*	0.737

Berdasarkan Tabel 2, tingkat akurasi dari metode *random forest* memang lebih tinggi dibandingkan dengan naïve bayes, baik dari segi *precision, recall, dan f1-score nya.*

Kesimpulan

Berdasarkan uraian hasil dari dan pembahasan penelitian ini, dapat disimpulkan bahwa klasifikasi *random forest* lebih baik dibandingkan menggunakan metode naïve bayes. Hal ini ditunjukkan dari nilai akurasi dari metode *random forest.* Random forest mampu melakukan 98.9% benar dalam akurasi data *spam email.* Sedangkan Naïve Bayes mampu melakukan 73,7% benar dalam akurasi data *spam email.* Jika dilihat selisih dari metode naïve bayes dan *random forest* cukup jauh yaitu 26%.

Referensi

[1] D. K. Renuka and Dr. T. Hamsapriya, “Email classification for Spam Detection using Word Stemming,” *Int J Comput Appl*, vol. 1, no. 5, pp. 58–60, 2010, doi: 10.5120/125-241.

[2] A. A. Akinyelu and A. O. Adewumi, “Classification of phishing email using random forest machine learning technique,” *J Appl Math*, vol. 2014, no. May, 2014, doi: 10.1155/2014/425731.

[3] 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.

[4] S. Defiyanti and D. L. Crispina Pardede, “Perbandingan kinerja algoritma id3 dan c4.5 dalam klasifikasi spam-mail,” *ReCALL*, 2008.

[5] I. Nurandini and A. F. Huda, “Klastering Dokumen dengan Menambahkan Metadata Menggunakan Algoritma COATES,” *Kubik: Jurnal Publikasi Ilmiah Matematika*, vol. 2, no. 2, pp. 39–44, Nov. 2017, doi: 10.15575/kubik.v2i2.1859.

[6] D. Faisal, Reza M; Nugrahadi, *Belajar Data Science*, no. February. Banjarbaru, Kalimantan Selatan, Indonesia: Scripta Cendekia, 2019.

[7] H. D. Anggana, “Penerapan Model Klasifikasi Regresi Logistik, Support Vector Machine , Classification and Regression Tree Terhadap Data Kejadian Difteri Di Provinsi Jawa Barat,” *Euclid*, vol. 5, no. 2, p. 20, 2018, doi: 10.33603/e.v5i2.1121.

[8] G. Louppe, “Understanding Random Forests: From Theory to Practice,” no. July, 2014.

[9] D. F. Durrah, R. Cahyandari, and A. S. Awalluddin, “Model Regresi Data Panel Terbaik untuk Faktor Penentu Laba Neto Perusahaan Asuransi Umum Syariah di Indonesia,” *Kubik: Jurnal Publikasi Ilmiah Matematika*, vol. 5, no. 1, pp. 28–34, Oct. 2020, doi: 10.15575/kubik.v5i1.8488.

[10] A. Y. Samudra, “Pendekatan Random Forest untuk Model Peramalan Harga Tembakau Rajangan Di Kabupaten Temanggung,” vol. 8, no. 5, p. 55, 2019.

[11] Syarli and A. A. Muin, “Metode Naive Bayes Untuk Prediksi Kelulusan,” *Jurnal Ilmiah Ilmu Komputer*, vol. 2, no. 1, pp. 22–26, 2016.