Vol. 3, No. 1, Juni 2022, hal. 9-19

p-ISSN: 2723-567X
e-ISSN: 2723-5661

Jurnal Computer Science and Information Technology (CoSciTech)

http://ejurnal.umri.ac.id/index.php/coscitech/index

Filtering spam email menggunakan algoritma naïve bayes

Januar Al Amien[1], Harun Mukhtar[*2] ,M. Arif Rucyat[3]
Email: [1]januaralamien@umri.ac.id, [2]harunmukhtar@umri.ac.id, [3]muhammadarifrucyat@gmail.com

[1,2,3]Teknik Informatka, Fakultas Ilmu Komputer, Universitas Muhammadiyah Riau

Diterima: 18 Mei 2022 | Direvisi: - | Disetujui: 31 Mei 2022

Abstrak

Email atau Surat elektronik yang biasa disingkat surat-e adalah sarana kirim mengirim surat melalui jalur internet. Email spam dapat diartikan sebagai tindakan mendistribusikan pesan yang tidak diminta, seringkali dikirim secara massal menggunakan *email*. *Email*, dikirim untuk tujuan yang sah, dikenal sebagai Ham. Masalah penelitian ini ialah bagaimana memnyaring *email spam* yang masuk. Penelitian ini bertujuan untuk memfilter *spam email* secara otomatis menggunakan algoritma *naïve bayes* yang terhubung dengan *mail server*. *Naïve bayes* merupakan fungsi yang banyak digunakan oleh pengembang *spam filter* sebagai fungsi untuk *filtering email* karena sederhana dan mudah untuk diimplementasikan. Hasil penelitian menunjukkan bahwa *naïve bayes* dapat *memfilter email spam* dengan klasifikasi text. Saran penulis untuk pengembangan dapat menggunakan algoritma lain.

Kata kunci: *Email, Spam Filter, Naive Bayes, Mail Server, Zimbra*

Email spam filtering using nave bayes algorithm

Abstract

Email or e-mail commonly abbreviated e-mail is a means of sending mail through the Internet. Spam emails can be interpreted as the act of distributing un requested messages, often sent in bulk using email. The email, sent for a legitimate purpose, is known as Ham. The problem with this research is how to filter incoming spam emails. This research aims to filter email spam automatically using naïve bayes algorithm connected to mail server. Naïve bayes is a function widely used by spam filter developers as a function for email filtering because it is simple and easy to implement. The results showed that naïve bayes can filter spam emails with text classification. The author's advice for development can use other algorithms.

Keywords: *Email, Spam filter, Naive Bayes, Mail Server, Zimbra*

1. PENDAHULUAN

Email merupakan suatu entitas penting yang digunakan untuk berkomunikasi digital melalui internet, selain itu digunakan untuk transfer informasi berupa *file* bahkan dapat digunakan untuk media iklan. Pesan Elektronik menjadi primadona untuk berkomunikasi saat ini. Hanya terhubung dengan koneksi internet, berkirim pesan elektronik dapat dengan mudah dilakukan [1].

Spam, juga disebut sebagai *unsolicited commercial email* atau *unsolicited bulk email* telah menyebabkan beberapa masalah komunikasi dalam kehidupan sehari hari kita. Kerugian yang disebabkan karena spam antara lain spam menempati sumber daya yang besar (termasuk bandwidth jaringan, ruang penyimpanan.), contoh kasus *spam* bisa berupa iklan perjudian maupun pornografi. *Spam* atau *junk email* adalah penyalahgunaan dalam pengiriman berita elektronik untuk menampilkan berita, iklan, dan keperluan lainnya yang mengakibatkan ketidaknyamanan bagi para pengguna.[2]

Banyak pengguna *email* yang merasa terganggu dengan adanya *spam.* Dampak buruk yang paling utama dari *spam email* adalah waktu yang terbuang dengan percuma untuk menghapus spam. Permasalahan ini menjadi permasalahan yang penting untuk dipecahkan. Untuk mengatasi hal ini diperlukan suatu *filter spam* dengan algoritma tertentu yang dapat memisahkan antara *spam-email* dengan *non spam email* dan mendapatkan hasil dari pemfilteran tersebut. algoritma yang penulis gunakan dalam penelitian

ini adalah *Naïve Bayes*. Penggunaan email yang sangat intens ini menimbulkan dampak positif dan negatif karena 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 kita kenal sebagai spam atau junk mail (email sampah) yang mana *email* tersebut berisikan iklan, penipuan dan bahkan virus. [3]

Penelitian ini membahas tentang teknik mengatasi spam email dengan menggunakan *pemfilteran* berdasarkan *subject email* dan isi kandungan *email* berupa text dari dataset. Salah satu solusi untuk mengatasi permasalahan *spam email* tersebut adalah dengan teknik penyaringan *spam email*. Hal inilah yang mendasari dilakukannya penelitian tentang *filtering* spam email menggunakan algoritma *Naïve Bayes* sehingga dapat diketahui hasil yang diberikan oleh algoritma tersebut.
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 sememakin popular sejak awal 1990-an dan merupakan masalah yang dihadapi oleh sebagian besar pengguna *email*.[4]
Zimbra merupakan aplikasi *mail server* berlisensi bebas dimana memiliki fitur-fitur yang lengkap dan juga kemudahan untuk instalasi maupun management *mail server*, meskipun masalah keamanan *mail server* menjadi faktor yang utama yang harus diperhatikan oleh system administrator. Zimbra adalah sebuah produk *groupware* yang dibuat oleh Zimbra, Inc yang berlokasi di Palo Alto, California, Amerika Serikat. [5]
Naïve Bayes merupakan suatu metode klasifikasi yang menggunakan perhitungan probabilitas Teori *Naïve Bayes* diadopsi dari nama penemunya yaitu Thomas Bayes sekitar tahun 1950. *Naïve Bayes* menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari *dataset* yang diberikan. Keuntungan menggunakan *Naïve Bayes*, metode ini hanya membutuhkan jumlah data pelatihan (*Training Data*) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. [2]

$$P\ (H|X) = \frac{P(X|H).P(H)}{P(X)} \qquad (1)$$

Dimana:
X : Data dengan class yang belum diketahui
H : Hipotesis data merupakan satu class sesifik
P(H|X) : Probabilitas hipotesis H berdasar kondisi X (posterior probabilitas)
P(H) : Probabilitas hipotesis H (prior Probabilitas)
P(x|H) : Probabilitas X berdasarkan Kondisi pada hipotesis H
P(X) : Probabilitas X

Text Mining

Text Mining merupakan teknik yang digunakan untuk menangani permasalahan klasifikasi, *clustering*, *information extraction* dan *information retrieval*. Pada dasarnya proses kerja dari *Text Mining* banyak mengadopsi dari penelitian data mining namun yang menjadi perbedaan adalah pola yang digunakan oleh *Text Mining* diambil dari sekumpulan bahasa alami yang tidak terstruktur sedangkan dalam data mining pola yang diambil dari database yang terstruktur. [6]

2. METODE PENELITIAN

2.1. Studi literature

Sebelum melakukan penelitian, peneliti pokok permasalahan serta materi materi pendukung. Diantaranya dataset penelitian terkait serta metode yang akan di implementasikan serta metode pengukur untuk menguji metode yang digunakan dalam implementasi disini yaitu metode naïve bayes sedangkan untuk kinerja diukur nilai akurasi dan presisinya.

2.2. Pengumpulan data

Pengumpulan data merupakan tahap awal yang digunakan sebagai masukan. Dataset lingspam yang di peroleh dari data public Kaggle: Machine Learning and Data Science Community.

a. Preprocessing

Preprocessing. Tahapan ini diantaranya adalah *cleaning, case folding, tokenizing* dan *filtering*. Tujuan dari tahapan ini adalah untuk mempersiapkan data agar menjadi data yang siap untuk dianalisis.[7]

b. Pelabelan

Dalam proses pembuatan dataset untuk sistem *filtering*, diperlukan mekanisme bagaimana agar dataset yang telah dikumpulkan memiliki label kelas yang benar. Pada kenyataanya *dataset* yang sudah dilabeli sangat sedikit dan sulit dicari. Proses pelabelan dataset akan mudah jika datanya berjumlah sedikit dan tidak terlalu besar, namun akan sangat membutuhkan waktu yang sangat

lama bahkan tidak mungkin dikerjakan sendiri jika dataset berjumlah sangat besar. Proses pelabelan data dapat dilakukan secara manual sendirian atau dikerjakan bersama-sama oleh beberapa bahkan puluhan hingga ratusan orang menggunakan teknik *crowdsourced labelling*.[8]

c. *Stemming*

Stemming yaitu proses mengubah suatu kata bentukan menjadi kata dasar. Proses *stemming* sangat tergantung kepada bahasa dari kata yang akan di stemm. Hal ini dikarenakan, dalam melakukan proses stemming harus mengaplikasikan aturan morfologikal dari suatu Bahasa. [9]

2.3. Imlementasi Naïve Bayes

Metode yang digunakan dalam penelitian ini yaitu naïve bayes. Metode *naïve bayes* sangat baik digunakan untuk pemfilteran, selain itu metode ini digunakan untuk memprediksi suatu kejadian pada masa yang akan datang, dengan cara membandingkan data atau evidence (bukti) yang ada pada masa lampau.

2.4. Pengujian *Naïve Bayes*

Pengujian metode dilakukan dengan data spam email yang diperoleh dari *dataset* yang telah dipisahkan kedalam class spam dan bukan spam.

2.5 Hasil

Hasil yang ingin didapat yaitu, *terfilternya email spam* dari metode *naïve bayes*. Selanjutnya akan menentukan apakah metode ini layak atau tidak jika dikembangkan serta jika ada penelitian lain yang ingin membandingkan dengan metode lain.

2.6. Dataset

Dataset adalah data yang telah di ambil dari internet, dataset ini berupa file teks format csv. Data ini akan dibersihkan dalam tahapan *cleaning* dan dilanjutkan dengan tahapan *preprocessing* dan klasifikasi. *Dataset* pada penelitian ini adalah *dataset* lingspam yang diambil dari data *public Kaggle: Machine Learning and Data Science Community*. Link:https://www.kaggle.com/mandygu/lingspam-dataset.

3. HASIL DAN PEMBAHASAN

Bab ini mengurai tentang proses skenario pembuatan *server mail*, perancangan topologi, perancangan aplikasi *mail client*, pengujian *model filtering Naïve Bayes*, hingga *filtering model* yang nantinya akan membantu dalam menyelesaikan laporan penelitian.

3.1 Analisis Sistem

Analisis sistem yang berjalan merupakan gambaran tentang Sistem yang saat ini sedang berjalan, yaitu Sistem *Mail Server* dengan menggunakan Sistem Operasi ubuntu Ver 16 dan *Mail Transfer Agent Zimbra Mail Server* Ver 8.8.12.

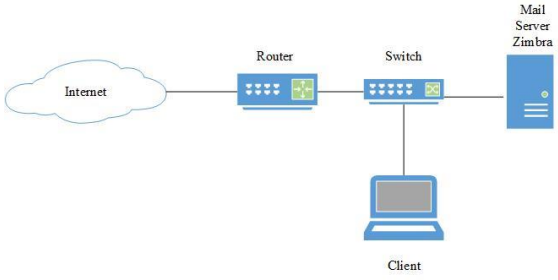
Sistem *filtering* pada *mail client* dibuat seperti halnya mail client pada umumnya. Metode *naive bayes* yang telah dibangun sebelumnya, akan diletakkan pada program dari *mail client* itu sendiri.

3.2 Topologi Jaringan

Topologi jaringan ini menggunakan *virtual server Ubuntu* 16 dan di instalkan ke dalamnya *zimbra mail server*, untuk pengetesan menggunakan laptop *client.*

Server : Zimbra Mail Server, IP: 192.168.43.64

Client : Laptop IP. 192.168.43.29

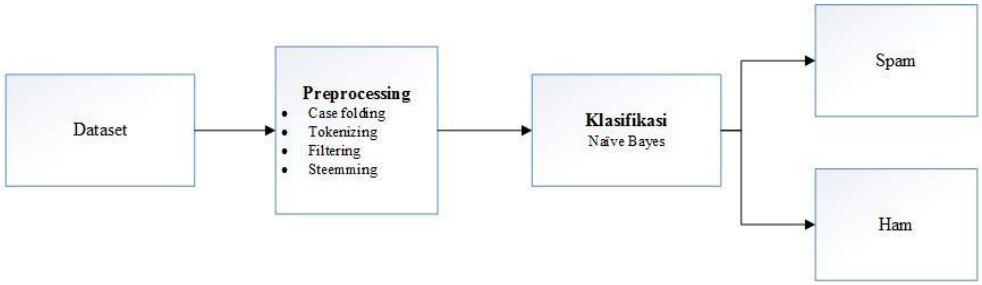


Gambar 1. Topologi *Jaringan*

Pada gambar 1 kedudukan server email sama dengan jaringan *LAN*, client tersebut dapat saling berkomunikasi dengan baik dengan *server email* dalam jaringan yang sama. Pada topologi ini, server email dapat diakses dari luar jaringan *(internet)* dengan syarat dapat diajukan ke mx *record domain* @arifrucyat.web.id yang diteruskan ke *server mail.* Agar *client* yang terdapat dalam jaringan *Lan* dapat berkirim *email.*

3.3 Skenario Proses Naïve Bayes Filtering Spam Email

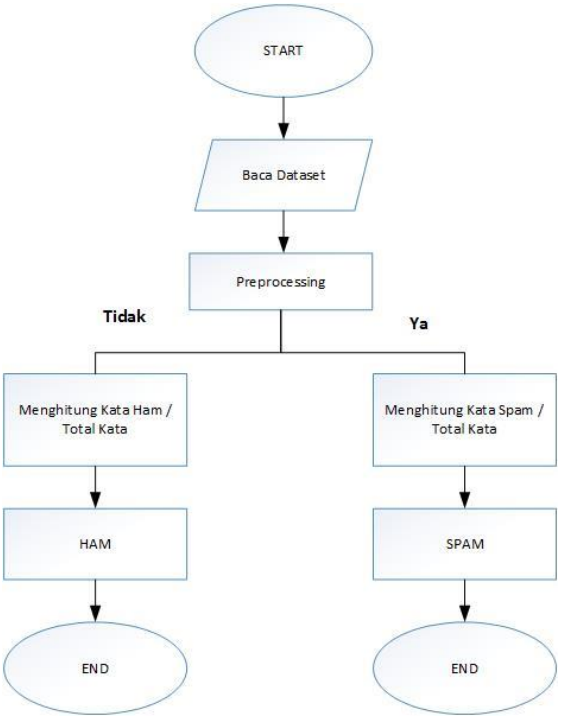
Proses algoritma *naïve bayes* dalam *memfilter spam email* ada beberapa tahapan yaitu sebagai berikut:



Gambar 2. Proses Analisis *Spam*

Pattern Discovery

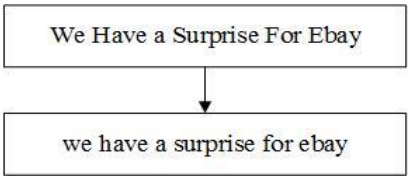
Pada tahapan ini digunakan algoritma *naive bayes* dalam tahap *pattern discovery* (pencarian pola). Tahapan proses *naïve bayes* dalam memfilter spam yaitu:



Gambar 3. *Pattern Discovery*

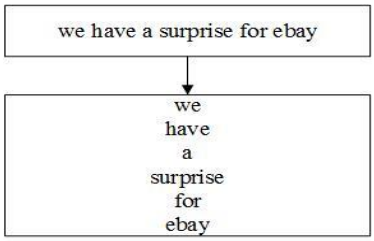
Tahapan *Case Folding*

Case folding merupakan proses yang mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf ‘a’ sampai dengan ‘z’ yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter (pembatas). Contoh penggunaan *case folding* sebagai berikut: [10]



Gambar 1. Proses *Case Folding*

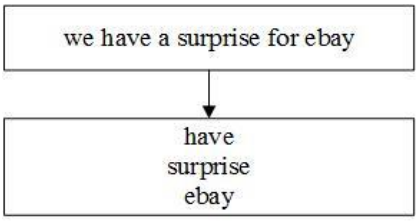
Tahapan *Tokenizing*
Tahap *tokenizing / parsing* adalah tahap pemotongan kata berdasarkan tiap kata yang menyusunnya. Selain itu, spasi digunakan untuk memisahkan antar kata.



Gambar 2. Proses *Tokenizing*

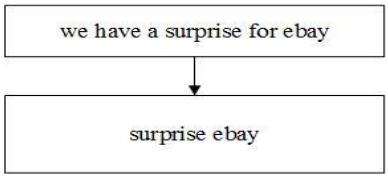
Tahapan *Filtering*

Tahap *filtering / stopword removal* adalah tahap mengambil kata - kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata penting). *Stopword* adalah kata-kata 16 yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. *Stopwords* adalah kata umum yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna. Umumnya dimanfaatkan dalam *task information retrieval,* termasuk oleh *Google.*



Gambar 3. Proses *Filtering*

Tahapan Steeming
Tahap stemming adalah tahap mencari root kata dari tiap kata hasil filtering. Pada tahap ini dilakukan proses pengembalian berbagai bentukan kata kedalam suatu representasi yang sama. Pencarian kata dasar dilakukan dengan menghilangkan semua imbuhan dari kata, baik itu awalan, sisipan, maupun akhiran. Disini saya menggunakan library Sastrawi untuk proses stemming.



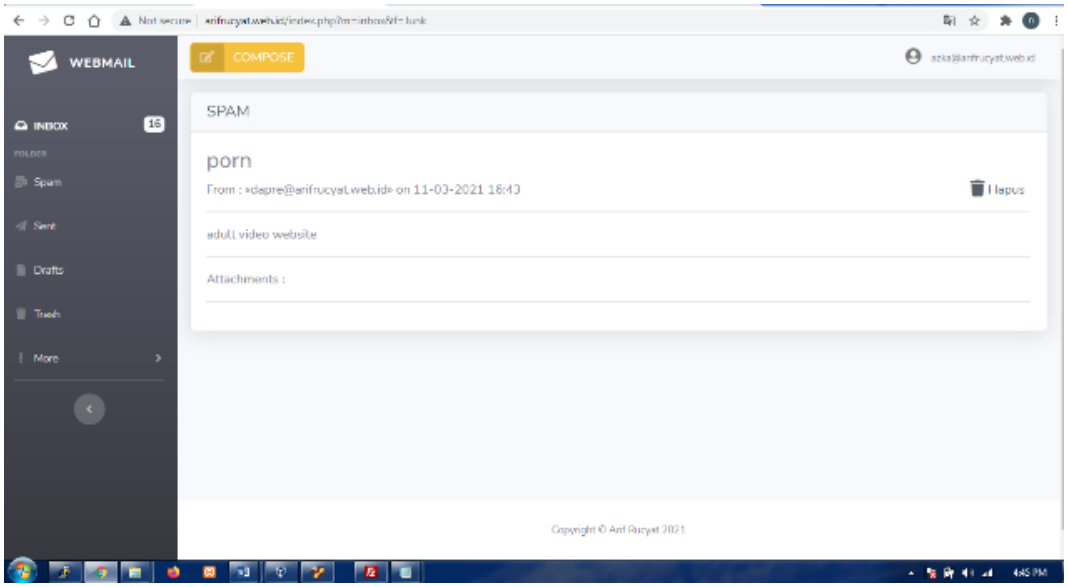
Gambar 4. Proses Steeming

3.4 Pengujian Filtering Spam

Pada tahap ini dilakukan pengujian dengan dataset yang telah di uplod ke *mail server zimbra.* Proses algoritma *Naïve Bayes* bekerja dalam menentukan *spam email* dengan belajar dari *dataset* yang telah ada, jadi algoritma naïve bayes ini dalam *memfilter spam email* dengan cara mendeteteksi kata kata yang dianggap *spam* dari *email* tersebut, apabila didalam *email* terdapat kata kata spam yang telah di masukkan kedalam dataset, maka *email* tersebut akan di anggap sebagai *spam.*

1. Pengujian Pertama

Pada pengujian *filter spam email* pertama di ilustrasikan dengan cara mengirim email yang mempunyai *subject "porn"*. Bagaimanakah hasil dari *system filtering* dalam *memfilter email* tersebut.
Pengirim: dapre@arifrucyat.web.id
Penerima : azka@arifrucyat.web.id



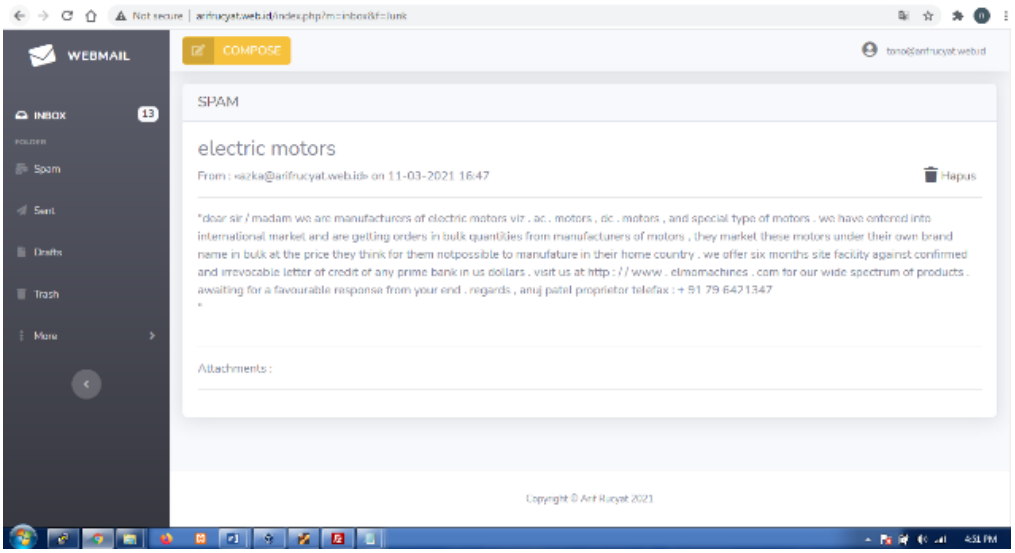
Gambar 8. Tampilan Isi *Spam* Pertama

Pada gambar 8 merupakan suatu *interface* dari *mail client* pada aplikasi untuk mengirimkan *email* dengan isi *subject "porno"*, dan isi *email "adult video website"*. Dan ternyata hasil yang di dapat setelah mengalami proses *filtering* adalah *email* tersebut merupakan *SPAM*. Hal ini dapat dibuktikan pada gambar 3.14, *email* berada pada folder *SPAM* berarti terdeteksi sebagai *SPAM*.

2. Pengujian Kedua

Pada pengujian *filter spam email* kedua di ilustrasikan dengan cara mengirim *email* yang mempunyai *subject "electric motors"*. Bagaimanakah hasil dari *system filtering* dalam *memfilter email* tersebut.

Pengirim : azka@arifrucyat.web.id
Penerima : tono@arifrucyat.web.id

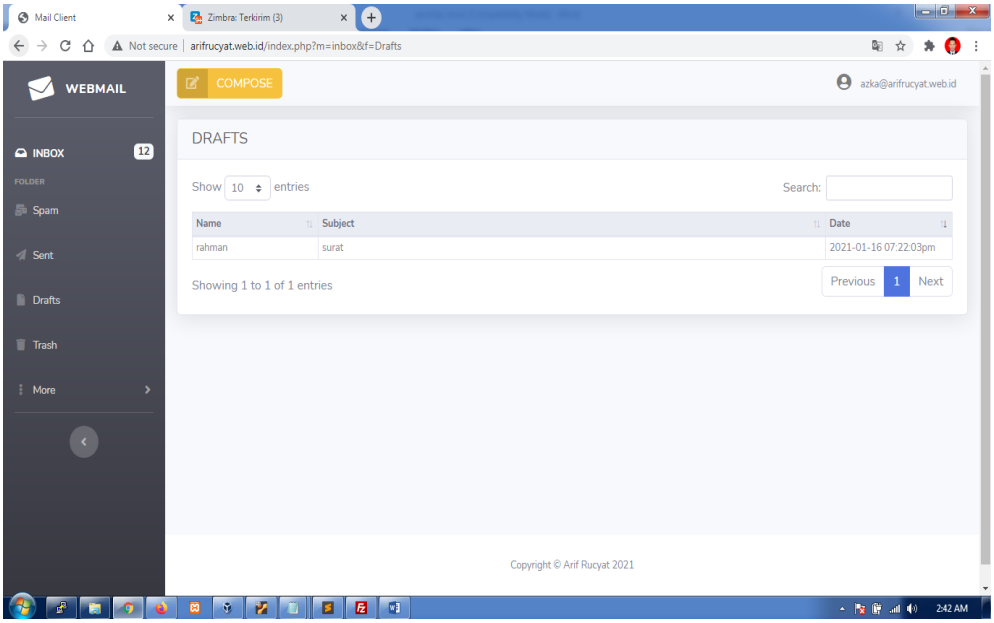


Gambar 9. Tampilan Isi *Spam* Kedua

Pada gambar 9 merupakan suatu interface dari mail client pada aplikasi untuk mengirimkan email dengan isi subject "electric motors", dan isi email ""dear sir / madam we are manufacturers of electric motors viz . ac . motors , dc . motors , and special type of motors . we have entered into international market and are getting orders in bulk quantities from manufacturers of motors , they market these motors under their own brand name in bulk at the price they think for them notpossible to manufature in their home country . we offer six months site facility against confirmed and irrevocable letter of credit of any prime bank in us dollars . visit us at http : / / www . elmomachines . com for our wide spectrum of products . awaiting for a favourable response from your end . regards , anuj patel proprietor telefax : + 91 79 6421347". Dan ternyata hasil yang di dapat setelah mengalami proses filtering adalah email tersebut merupakan SPAM. Hal ini dapat dibuktikan pada gambar 9, email berada pada folder SPAM/Junk berarti terdeteksi sebagai SPAM.

1. Tampilan *Draft*

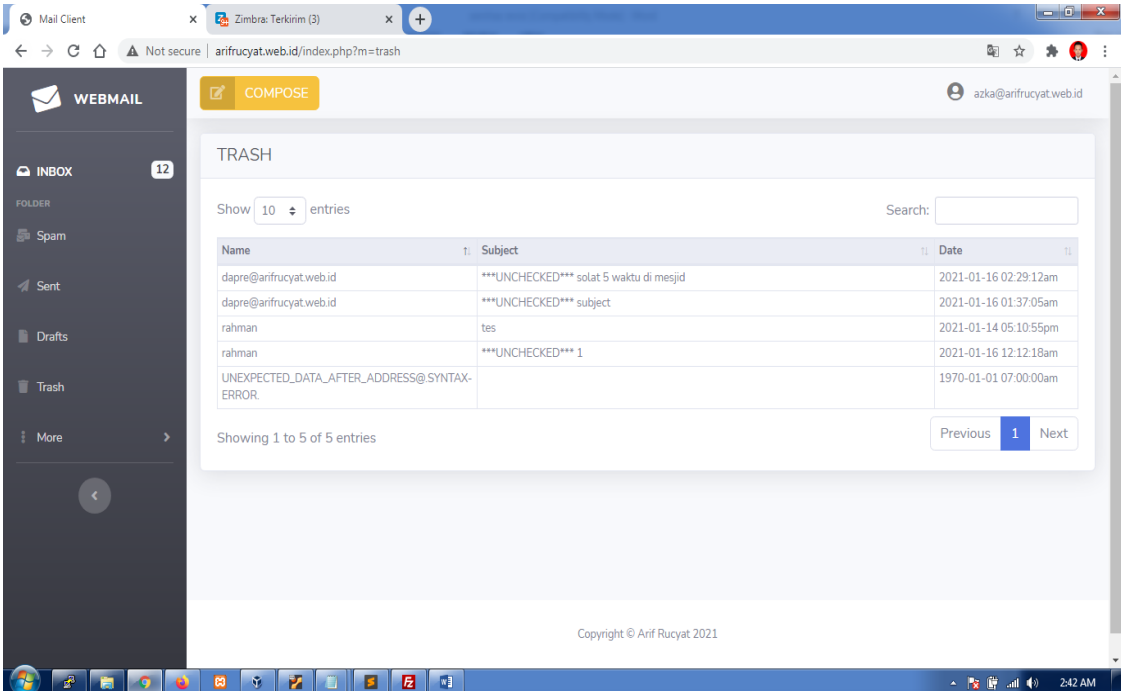
Berikut ini adalah halaman *draft* yang berisikan *email* yang belum ingin dikirim namun sengaja disimpan.



Gambar 10. Tampilan Draft

2. Tampilan Trash

Berikut ini adalah halaman *Trash* yang berisikan *email email* yang telah masuk lalu dihapus makan akan masuk ke halaman *trash*.



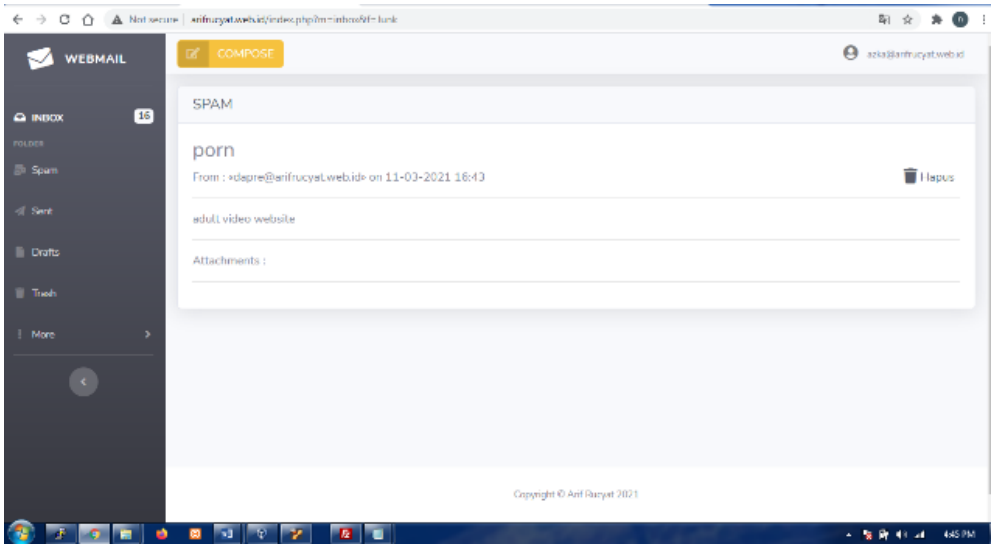
Gambar 11. Tampilan *Trash*

3.5 Pengujian Filtering Spam

Pada tahap ini dilakukan pengujian dengan dataset yang telah di uplod ke *mail server zimbra.* Proses algoritma *Naïve Bayes* bekerja dalam menentukan *spam email* dengan belajar dari *dataset* yang telah ada, jadi algoritma naïve bayes ini dalam *memfilter spam email* dengan cara mendeteteksi kata kata yang dianggap *spam* dari *email* tersebut, apabila didalam *email* terdapat kata kata spam yang telah di masukkan kedalam dataset, maka *email* tersebut akan di anggap sebagai *spam.*

1. Pengujian Pertama

Pengujian *filter spam email* pertama di ilustrasikan dengan cara mengirim email yang mempunyai *subject "porn"*. Bagaimanakah hasil dari *system filtering* dalam *memfilter email* tersebut.
Pengirim: dapre@arifrucyat.web.id
Penerima : azka@arifrucyat.web.id

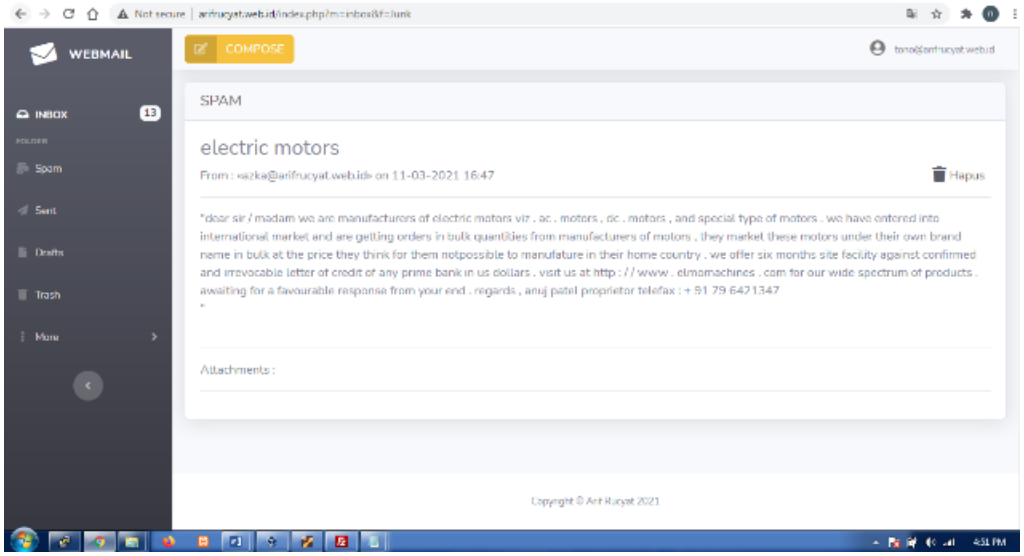


Gambar 12. Tampilan Isi *Spam* Pertama

Gambar 12 merupakan suatu *interface* dari *mail client* pada aplikasi untuk mengirimkan *email* dengan isi *subject "porno"*, dan isi *email "adult video website"*. Dan ternyata hasil yang di dapat setelah mengalami proses *filtering* adalah *email* tersebut merupakan *SPAM*. Hal ini dapat dibuktikan pada gambar 12, *email* berada pada folder *SPAM* berarti terdeteksi sebagai *SPAM*.

2. Pengujian Kedua

Pada pengujian *filter spam email* kedua di ilustrasikan dengan cara mengirim *email* yang mempunyai *subject "electric motors"*. Bagaimanakah hasil dari *system filtering* dalam *memfilter email* tersebut.
Pengirim : azka@arifrucyat.web.id
Penerima : tono@arifrucyat.web.id



Gambar 13. Tampilan Isi *Spam* Kedua

Gambar 13 merupakan suatu interface dari mail client pada aplikasi untuk mengirimkan email dengan isi subject "electric motors", dan isi email ""dear sir / madam we are manufacturers of electric motors viz . ac . motors , dc . motors , and special type of motors . we have entered into international market and are getting orders in bulk quantities from manufacturers of motors, they market these motors under their own brand name in bulk at the price they think for them notpossible to manufature in their home country . we offer six months site facility against confirmed and irrevocable letter of credit of any prime bank in us dollars . visit us at http : / / www . elmomachines . com for our wide spectrum of products . awaiting for a favourable response from your end . regards , anuj patel proprietor telefax : + 91 79 6421347". Dan ternyata hasil yang di dapat setelah mengalami proses filtering adalah email tersebut merupakan SPAM. Hal ini dapat dibuktikan pada gambar 3.15, email berada pada folder SPAM/Junk berarti terdeteksi sebagai SPAM.

No	Pengujian	Pengirim	Penerima	Status Pesan	Penggunaan *Mailbox* Akun	Kategori

	Dataset					
1	*Document* 473	dapre@arifrucyat.web.id	azka@arifrucyat.web.id	Terkirim	zulkifli@arifrucyat.web.id 13.09 KB	*Spam (Pornography)*
2	*Document* 103	azka@arifrucyat.web.id	tono@arifrucyat.web.id	Terkirim	tono@arifrucyat.web.id 44.44 KB	*Spam (Promotion/ Advertisement)*
3	*Document* 174	zulkifli@arifrucyat.web.id	dapre@arifrucyat.web.id	Terkirim	azka@arifrucyat.web.id 60.96 KB	*Spam (Promotion/ Advertisement)*
4	*Document* 160	zulkifli@arifrucyat.web.id	willy@arifrucyat.web.id	Terkirim	dapre@arifrucyat.web.id 85.40 KB	*Spam (Promotion/ Advertisement)*
5	*Document* 100	willy@arifrucyat.web.id	kevin@arifrucyat.web.id	Terkirim	kevin@arifrucyat.web.id 3.73 KB	*Spam (Pornography)*

Jadi Naïve Bayes bekerja, setiap email yang masuk akan diperiksa berdasarkan kata-kata yang termasuk dalam karakteristik yang telah ditentukan berdasarkan dataset. Kemudian dari kata-kata tersebut dihitung probabilitas suatu email apakah tergolong spam atau ham. Suatu email tergolong spam jika kata-kata terdapat didalam dataset spam , maka email tersebut tidak dapat masuk kedalam client inbox, tetapi akan masuk ke folder spam/junk, akan tetapi jika tidak email tersebut masuk dalam kategori ham maka akan masuk ke client inbox.

3. Pengujian Dataset

Proses klasifikasi data pada penelitian kali ini menggunakan *tools Anavonda Navigator / jupyter lab. Dataset* memiliki 3 atribut yaitu *subject, message*, dan *label*. Dataset memiliki 2412 ham email dan 481 spam *email*, total keseluruhan data yaitu 2893 data. Disini penulis akan mencari *accuracy, presisi, dan recall* dari *dataset* menggunakan algoritma *naïve bayes.* Secara keseluruhan proses klasifikasi terdiri data *retrieve* dataset, *operator split* data, *operator* model, *operator apply*, dan *performance*. Pada aplikasi *jupyter lab* akan menampilkan hasil dari *accuracy, presisi, dan recall* serta hasil *example set* sesuai dengan hasil *training* terhadap pemodelan klasifikasinya.

	Precision	**Recall**	**Fi-score**	**Support**
0	1.00	0.98	0.99	785
1	0.92	0.99	0.96	170
Accuracy			0.98	955
Macro |Avg	0.96	0.99	0.97	955
Weighted Avg	0.99	0.98	0.98	955

Smote

SMOTE (*Synthetic Minority Over-Sampling Technique*) menghasilkan hasil yang baik dan efektif untuk menangani class imbalance yang mengalami overfitting pada proses teknik over-sampling untuk kelas minoritas (positif). *SMOTE* menciptakan sebuah contoh dari kelas minoritas sintetis yang beroperasi di ruang fitur daripada ruang data. Dengan menduplikasi contoh kelas minoritas, teknik *SMOTE* menghasilkan contoh sintetis baru dengan melakukan ekstrapolasi sampel minoritas yang ada dengan sampel acak yang diperoleh dari nilai k tetangga terdekat.[11]

Confusion Matrix

Deteksi *email spam* dapat dievaluasi dengan cara yang berbeda ukuran kinerja. *Confusion Matrix* sedang digunakan memvisualisasikan deteksi *email* untuk model. *Confusion Matrix* dapat di definisikan seperti dibawah ini: [12]

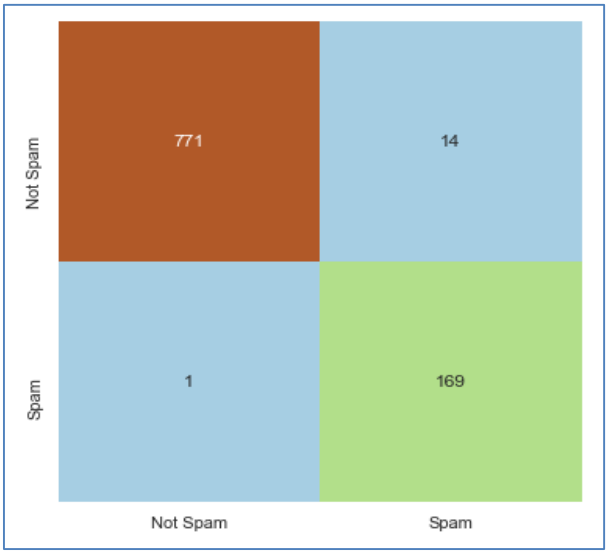
$$TN = True\ Negative – Ham\ email\ predicted\ as\ ham \quad (2)$$

$$TP = True\ Positive – Spam\ email\ predicted\ as\ spam \quad (3)$$

$$FP = False\ Positive – Spam\ email\ predicted\ as\ ham \quad (4)$$

$$FN = False\ Negative – Ham\ email\ predicted\ as\ spam \quad (5)$$

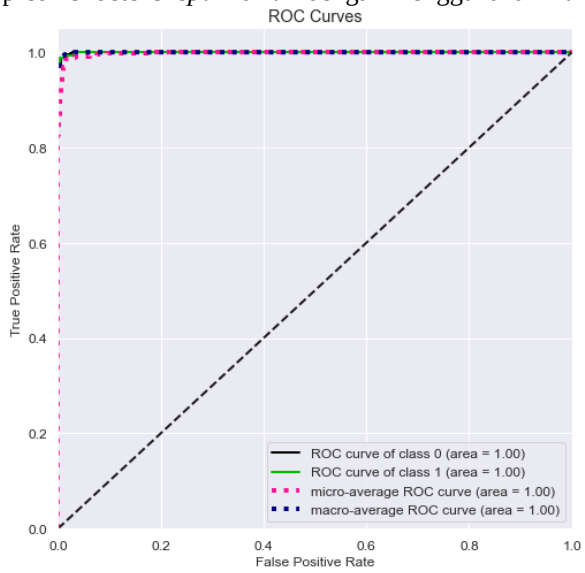
Predicted Class	***Ham***	***Spam***
Ham	*TN*	*FP*
Spam	*FN*	*TP*



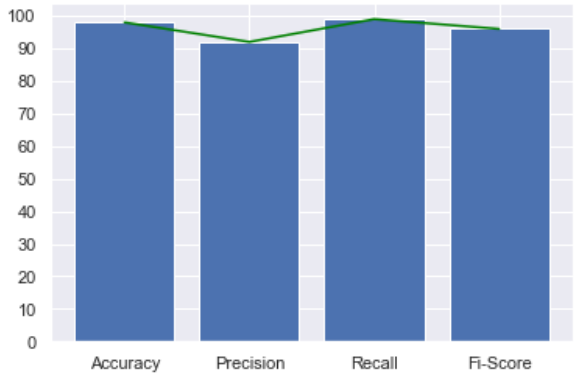
Gambar 14. Confusion Matrix

Grafik

Dibawah ini merupakan grafik hasil prediksi deteksi *spam email* dengan menggunakan *Naïve Bayes*.



Gambar 15. Grafik



Gambar 16. Grafik Bar

Gambar 16 menunjukkan tingkat accuracy terbaik dalam deteksi spam email sebesar 98.00%, Precision 92.00%, Recall 99.00%, Fi-Score 96.00%.

4. KESIMPULAN

Berdasarkan penelitian yang dilakukan mengenai imlementasi filtering spam email menggunakan algoritma naïve bayes maka dapat ditarik kesimpulan sebagai berikut

1. System *filtering email* spam dengan metode naïve bayes sangat efektif untuk *filtering text*, karna naïve bayes dapat memfilter email spam berdasarkan kata-kata yang ada didataset.
2. Pengujian dataset dengan pemodelan naïve bayes mendapatkan *accuracy* sebesar 98.00%.

DAFTAR PUSTAKA

[1] M. Al-Tahrawi, M. Abualhaj, and S. Al-Khatib, "Polynomial neural networks versus other spam email filters: An empirical study," *TEM J.*, vol. 9, no. 1, pp. 136–143, 2020, doi: 10.18421/TEM91-19.

[2] R. Y. Hayuningtyas, "Aplikasi Filtering of Spam Email Menggunakan Naïve Bayes," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 2, no. 1, pp. 53–60, 2017.

[3] S. N. D. Pratiwi and B. S. S. Ulama, "Klasifikasi Email Spam dengan Menggunakan Metode Support Vector Machine dan k-Nearest Neighbor," *J. Sains dan Seni ITS*, vol. 5, no. 2, pp. 344–349, 2016, doi: 10.12962/j23373520.v5i2.16685.

[4] M. A. Ghani and A. Subekti, "Email Spam Filtering Dengan Algoritma Random Forest," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. Vol.3, No., no. 2, p. 216~221, 2018.

[5] H. Mukhtar, Daniel Adi Putra Sitorus, and Yulia Fatma, "Analisa Dan Implementasi Security Mail Server," *J. Fasilkom*, vol. 10, no. 1, pp. 25–32, 2020, doi: 10.37859/jf.v10i1.1906.

[6] T. Kurniawan, "Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter Terhadap Media Mainstream Menggunakan Naïve Bayes Classifier Dan Support Vector Machine Media Mainstream Menggunakan Naïve Machine," p. 1, 2017.

[7] D. Juang, "Analisis Spam dengan Menggunakan Naïve Bayes," *J. Teknovasi*, vol. 03, no. 1998, pp. 51–57, 2016.

[8] A. Rachmat and Y. Lukito, "SENTIPOL: Dataset Sentimen Komentar Pada Kampanye PEMILU Presiden Indonesia 2014 dari Facebook Page," *Konf. Nas. Teknol. Inf. dan Komun. 2017*, no. December, pp. 218–228, 2016.

[9] M. A. Rofiqi, A. C. Fauzan, A. P. Agustin, and A. A. Saputra, "Implementasi Term-Frequency Inverse Document Frequency (TF- IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query," *J. Comput. Sci. Appl. informatics*, vol. 1, no. 2, pp. 58–64, 2019.

[10] R. Sistem, O. Nilai, K. Algoritma, and K. Spam, "Optimasi Nilai K pada Algoritma KNN untuk Klasifikasi Spam dan Ham Email," *J. resti*, vol. 1, no. 10, pp. 377–383, 2021.

[11] S. A. Putri, "Integrasi Teknik Smote Bagging Dengan Information," *J. Ilmu Pengetah. dan Teknol. Komput.*, vol. 2, no. 2, pp. 22–31, 2017.

[12] S. Gibson, B. Issac, L. Zhang, and S. M. Jacob, "Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms," *IEEE Access*, vol. 8, pp. 187914–187932, 2020, doi: 10.1109/access.2020.3030751.