Ads Filtering Mengunakan Jaringan Syaraf Tiruan Perceptron, Naïve Bayes Classifier, dan Regresi Logistik

Achmad Fachrudin Rachimawan dan Brodjol Sutijo Utama
Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
e-mail: brodjol_su@satistika.its.ac.id

Abstrak— Email merupakan fasilitas yang mutlak diperlukan dalam berbagai bidang. Pentingnya email dan jumlahnya yang begitu banyak menyebabkan penyalahgunaan. Salah satu penyalahgunaan yang sering ditemui adalah email iklan yang dikirimkan oleh perusahaan penyedia konten internet saat pengguna mendaftar pada situs perusahaan tersebut. Terdapat metode agar email iklan dari perusahaan-perusahaan tersebut bisa secara otomatis dikenali yaitu klasifikasi. Data email berbentuk teks, sehingga jauh lebih rumit dan perlu proses untuk mempersiapkan data. Salah satu prosesnya adalah pembobotan ads atau adicity. Untuk metode klasifikasi yang digunakan adalah Naive Bayes Classifier (NBC) yang secara umum sering digunakan dalam data teks dan Perceptron yang diketahui keduanya merupakan metode yang cukup sederhana untuk menyelesaikan permasalahan yang kompleks. Kedua metode tersebut akan dibandingkan untuk mengetahui hasil klasifikasi yang paling baik. Hasil penelitian menunjukkan bahwa NBC lebih unggul dibanding Perceptron, dan pada NBC False Positive Ratio lebih mudah untuk dikontrol.

Kata Kunci: email, iklan, klasifikasi, naïve bayes, perceptron

I. PENDAHULUAN

Perkembangan teknologi saat ini telah tumbuh dengan luar biasa pesat, khususnya di bidang teknologi komunikasi dan informasi. Dengan keberadaan internet, segala informasi dan berita dapat diterima dan diakses oleh setiap orang. Bahkan dengan internet, setiap orang dapat mengirim dan menerima pesan dari satu orang ke orang lainnya dengan mudah menggunakan sebuah pesan elektronik, ataupun dengan mengunakan media sosial. Pesan elektronik yang lebih dikenal sebagai email merupakan fasilitas yang saat ini menjadi sarana yang mutlak diperlukan dalam berbagai bidang, mulai dari bidang industri, pendidikan, kesehatan, dll. Tetapi tidak semua orang menggunakan email dengan baik dan benar, bahkan dapat menyebabkan kerugian bagi orang lain. Hal ini dikarenakan fasilitas email yang murah dan mudah digunakan oleh setiap orang, sehingga mengakibatkan banyaknya penyalahgunaan pada penggunaan email itu sendiri, atau yang biasa disebut dengan email spam atau *bulkmail* yang biasanya berisi beragam tujuan, diantaranya adalah penipuan (berkedok amal, undian lottere), pencucian uang atau *money loundring* (menawarkan transaksi pekerjaan yang berhubungan dengan transaksi bank), atau bahkan menyebarkan virus. [1].

Ads atau iklan tidaklah berbahaya, dibandingkan dengan spam, *Ads* tidak mengandung unsur penipuan ataupun virus, hanya saja *Ads* dirasa sudah sangat *annoying* atau mengganggu [2].

Ads umumnya tidak dikenali oleh teknik spam *blocking* karena IP para pengirim *Ads* bukanlah termasuk ke dalam *Global Balcklist*. Ada beragam metode dalam teknik *learning* (machine) yang dapat digunakan untuk memilah *Ads* seperti *Random Forest*, SVM (Support Vector Machine), KNN (K Nearest Neighbors), dan masih banyak lagi. Dan dari beragam teknik tersebut yang digunakan dalam penelitian ini Adalah jaringan syaraf tiruan *Perceptron*, *Naive Bayes Classifier* dan Regresi Logistik. Digunakannya ketiga metode tersebut karena secara umum sering digunakan dalam data teks.

Perceptron melatih jaringan untuk mendapatkan keseimbangan antara kemampuan jaringan untuk mengenali pola yang digunakan selama pelatihan serta kemapuan jaringan untuk memberikan respon yang benar terhadap pola masukan yang serupa (tetapi tidak sama) dengan pola yang dipakai selama pelatihan [3]. Penelitian oleh Owen & Richard [4] dengan menggunakan jaringan syaraf tiruan *perceptron* untuk klasifikasi email spam menghasilkan 5,02% error pada iterasi sebanyak 1000 kali.

Metode kedua yang digunakan dalam penelitian ini adalah *Naïve Bayes Classifier* (NBC), NBC telah banyak digunakan dalam penelitian mengenai *text mining* dan *spam filtering*, beberapa kelebihan NBC diantaranya adalah sederhana tapi memiliki akurasi yang tinggi [5]. Penelitian berkaitan dengan metode NBC telah dilakukan diantaranya oleh Durajati, C & Gumelar, A, B [6] menggunakan NBC, menghasilkan ketepatan klasifikasi sebesar 87% dan menyimpulkan bahwa semakin banyak data training semakin baik. Nugroho, P [7] menggunakan *naïve bayes classifier* untuk mengklasifikasikan email spam menghasilkan tingkat error sebesar 4,83%, dan juga menyimpulkan bahwa *naïve bayes classifier* mempunyai tingkat error yang besar jika terdapat selisih pada jumlah keyword yang ada di data training. Kurniawan, Effendi, & Sitompul [8], menggunakan NBC dengan *confix-stripping stemmer*, menyimpulkan bahwa hasil cukup baik untuk proses klasifikasi namun hanya memakai empat kategori untuk pengelompokan artikel berita Lestari [9], melakukan klasifikasi tipe kepribadian orang dengan menggunakan NBC dan menghasilkan ketepatan klasifikasi sebesar 92,5%. Dalam penelitian ini, kedua metode non-parametrik (Naïve Bayes dan Perceptron) akan dibandingkan dengan metode parametrik yaitu Regresi Logistik dan dicari metode yang menghasilkan tingkat galat paling kecil.

II. TINJAUAN PUSTAKA

A. *Advertisement*

Pengertian *Ad* atau iklan menurut bahasa adalah memperkenalkan suatu barang, produk dan mempromosikan barang atau pun jasa baik secara online mau pun offline yang disampaikan melalui media dan di biaya oleh pemrakarsa yang dikenal serta di tunjuk sebagian masyarakat melalui,radio, televisi, surat kabar, majalah dan lain-lain. Dan iklan juga di definisikan sebagai pesan yang menawarkan produk yang ditujukan kepada masyarakat lewat suatu media masa [10].

B. Text Mining

Istilah *data mining* adalah mencari pola dalam data. Demikian pula dengan *text mining* tentang mencari pola dalam teks. *Text mining* adalah proses menganalisis teks untuk mengekstrak informasi yang berguna untuk tujuan tertentu [11].

C. *Klasifikasi Text*

Klasifikasi teks merupakan proses menemukan pola baru yang belum terungkap sebelumnya. Klasifikasi teks dilakukan dengan memproses dan menganalisa data dalam jumlah besar. Dalam prosesnya, klasifikasi teks melibatkan struktur yang mungkin terdapat pada teks dan mengekstraks informasi yang relevan pada teks. Dalam menganalisis sebagian atau keseluruhan teks yang tidak terstruktur, klasifikasi teks mencoba mengasosiasikan sebagian atau keseluruhan satu bagian teks dengan yang lainnya berdasarkan aturan-aturan tertentu. [5].

D. *Praproses Text*

Karena pada setiap penelitian *text mining* melibatkan data atau informasi dari sumber data yang berbeda-beda, maka metode penanganan data dalam tahapan praproses pun menjadi berbeda-beda, sehingga metode yang akan digunakan diketahui setelah dilakukan praproses [12]. *Information gain* bisa dianggap masuk ke dalam praproses teks, digunakan untuk mengukur efektivitas suatu atribut dalam mengklasifikasikan data. Tahapan dalam *Information Gain* adalah dengan membuat:

a. *Wordlist* atau *dictionary*, dari seluruh jumlah email (5056) dikodekan menjadi angka-angka, dengan cara dibuat daftar seluruh kata yang pernah muncul pada email tersebut. Misal dari seluruh email ditemukan terdapat 200.000 kata, maka setiap kata dalam tiap email akan diganti dengan menggunakan angka-angka berdasar jumlah angka yang ditemukan.

b. *Feature Vectors Content*, Memilih (misal) sebanyak 100 kata dari keseluruhan kata yang terdapat pada email. Tiap kata mengadung sebuah fitur yang di dalamnya terkandung bobot untuk tiap-tiap kata.

c. *Features (Words) Selection*, setelah diambil ketentuan tentang banyaknya *Vectors Content* yang diambil, maka selanjutkan akan dipilih kata-kata yang mewakili atau merepresentasikan email dengan bobot tertinggi yang kemudian dipersiapkan untuk *Vectors Content*. Dengan langkah,

1. Menghitung Bobot Ad (*Ad-icity)*

Dinotasikan:

Pr(w|A) = peluang bahwa sebuah kata muncul di email non iklan.

Pr(w|B) = peluang bahwa sebuah kata muncul di email iklan.

Ad-icity bisa ditulis:

$$Ad-icity(w) = \frac{\Pr(w \mid A)}{\Pr(w \mid A) + \Pr(w \mid \mathrm{B})} \qquad (1)$$

Tiap peluang dihitung berdasarkan proporsi yang relevan dalam data training.

2. *Selection of Word*, dalam proses ini dilakukan *rangking* untuk tiap kata, dan (misal dengan jumlah 100) kata dengan rank tertinggi kemudian dipilih. Praktek intuitifnya adalah untuk menetapkan peringkat tinggi untuk kata-kata yang *Ad-icity* yang jauh dari 0,5 (tinggi atau lebih rendah). Jika Ad-icity mendekati nilai 1, maka dapat dikatakan kata tersebut adalah indikator *Ads* yang baik, dan sebaliknya jika mendekati 0 maka dapat dikatakan kata tersebut adalah indikator normal yang baik.
   Namun akan ditemukan, bahwa hanya melihat dari *Adicity* tidaklah cukup, kata dengan kedua Peluang Pr(w|A) dan Pr(w|B) nya terlalu kecil tidak bisa dijadikan sebuah indikator yang baik. Walaupun *Ad-icity* kata tersebut jauh dari 0.5 (mendekati 0 atau 1). Maka perlu dilihat pula seberapa besar selisih absolut seperti berikut:

   |Pr(w|A) - Pr(w|B)|

Proses pemilihan kata akan berlangsung:

- Saring atau temukan kata dengan |*Ad-icity* – 0.5| < 0.05
- Saring atau temukan kata yang jarang, yaitu yang muncul di kedua kategori yang kurang dari threshold yang diberikan (1%).
- Untuk tiap kata yang tidak tersaring, hitung |Pr(w|A) - Pr(w|B)|
- Pilih kata dengan |Pr(w|A) - Pr(w|B)| terbesar.

d. Shuffling (pengacakan)

Rasio Ad-normal berubah-ubah (dinamis) dari waktu ke waktu, dan itu adalah suatu hal yang menarik, dari masalah ini ditemukan bahwa Naïve Bayes classifier menyesuaikan dengan cepat untuk rasio baru yang dilatih secara bertahap. Maka data yang digunakan diambil secara acak baik pada pelatihan dan uji set sebelum digunakan, sehingga rasio Ad-normal dari waktu ke waktu dapat dipelajari oleh algoritma yang telah dibentuk. [1]

E. Perceptron

Perceptron juga termasuk salah satu bentuk jaringan syaraf yang sederhana. Perceptron biasanya digunakan untuk mengklasifikasikan suatu tipe pola tertentu yang sering dikenal dengan pemisahan secara linear. Pada dasarnya, perceptron pada jaringan syaraf dengan satu lapisan memiliki bobot yang bisa diatur dan suatu nilai ambang (*threshold*). Algoritma yang digunakan oleh aturan perceptron ini akan mengatur parameter-parameter bebasnya melalui proses pembelajaran. Nilai *threshold* (θ) pada fungsi aktivasi adalah non negative. Fungsi aktivasi ini dibuat sedemikian rupa sehingga terjadi pembatasan antara daerah positif dan daerah negative.

Garis pemisah antara daerah positif dan daerah nol memiliki pertidaksamaan:

$$w_1x_1 + w_2x_2 + b > \theta \qquad (2)$$

Sedangkan garis pemisah antara daerah negative dengan daerah nol memiliki pertidaksamaan:

$$w_1x_1 + w_2x_2 + b < -\theta \quad (3)$$

Algortma *Perceptron* sama dengan *delta rule*, yang membedakan hanya pada *Perceptron* garis pemisah menggunkan *threshold* atau suatu nilai ambang batas, jika pada *delta rule* garis pemisah menggunakan nilai 0.

F. Delta Rule

Pada delta rule akan mengubah bobot yang menghubungkan antara jaringan input ke output (y_in) dengan nilai target (t). Hal ini untuk dilakukan untuk meminimalkan error selama pelatihan pola. Delta rule untuk memperbaiki bobot ke-I (untuk setiap pola) adalah:

$$\Delta w_i = \alpha\,(t - y\_in) * x_i \quad (4)$$

dengan:

- x = vector input
- y_in = input jaringan ke input output Y

$$y\_in = \sum_{i=1}^{n} x_i * w_i \quad (5)$$

- t = target (output)

Nilai w baru diperoleh dari nilai w lama ditambah dengan Δw,

$$w_i = w_i + \Delta w_i \quad (6)$$

G. Naïve Bayes Classifier

Teorema Bayes merupakan teorema yang mengacu konsep probabilitas bersyarat [3]. Secara umum teorema Bayes dapat dinotasikan pada persamaan berikut:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad (7)$$

Metode *naive bayes classification* (NBC), merupakan salah satu metode yang dapat mengklasifikasikan teks. Kelebihan NBC adalah algoritmanya sederhana tetapi memiliki akurasi yang tinggi. Dalam algoritma NBC setiap dokumen direpresentasikan dengan pasangan atribut "a_1, a_2, a_3,..., a_n" dimana a_1 adalah kata pertama, a_2 adalah kata kedua dan seterusnya. Sedangkan V adalah himpunan kategori email. Pada saat klasifikasi algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan (V_{MAP}). Adapun persamaan V_{MAP} adalah sebagai berikut:

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}}\ P(v_j)\prod_i P(a_i|v_j) \quad (8)$$

Nilai P(v_j) dihitung pada saat data *training*, didapat dengan rumus sebagai berikut:

$$P(v_j) = \frac{|doc\ j|}{|training|} \quad (9)$$

Dimana |*doc j*| merupakan jumlah dokumen (email) yang memiliki kategori *j* dalam *training*. Sedangkan |*training*| merupakan jumlah dokumen (email) dalam contoh yang digunakan untuk *training*. Untuk probabilitas kata a_i untuk setiap kategori P($a_i|v_j$), dihitung pada saat *training*.

$$P(a_i|v_j) = \frac{|n_i + 1|}{|n + kosakata|} \quad (10)$$

Dimana n_i adalah jumlah kemunculan kata a_i dalam email yang berkategori v_j, sedangkan *n* adalah banyaknya seluruh kata dalam email dengan kategori v_j dan |*kosakata*| adalah banyaknya kata dalam contoh pelatihan.

H. Regresi Logistik

Regresi logistik adalah salah satu model untuk menduga hubungan antara peubah respon kategori dengan satu atau lebih peubah prediktor yang kontinyu ataupun kategori. Peubah respon yang terdiri dari dua kategori yaitu "ya (sukses)" dan "tidak (gagal)", dan dinotasikan 1="sukses" dan 0="gagal", maka akan mengikuti sebaran Bernoulli [13], menyatakan model regresi logistik:

$$\pi(X_i) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip})} \quad (11)$$

Dengan π(x) adalah peluang kejadian sukses dengan nilai probabilita $0 \le \pi(x) \le 1$ dan βj adalah nilai parameter dengan j = 1,2,.....,p. π(x) merupakan fungsi yang non linier, sehingga perlu dilakukan transformasi ke dalam bentuk logit untuk memperoleh fungsi yang linier agar dapat dilihat hubungan antara variabel bebas dan variabel tidak bebas. Dengan melakukan transformasi dari logit π(x), maka didapat persamaan yang lebih sederhana. Proses pendugaan parameter dari regresi logistik menggunakan metode MLE. Menurut Agresti [13], metode MLE memberikan nilai duga bagi β dengan cara memaksimumkan fungsi likelihood dan mensyaratkan bahwa data mengikuti sebaran Bernoulli. Fungsi likelihood untuk model regresi logistik dikotomus adalah:

$$\xi(\beta) = \prod_{i=1}^{n} f(\beta, y) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i} \quad (12)$$

Agar nilai fungsi mencapai maksimum maka turunan parsial pertama terhadap disamadengankan nol. Persamaan hasil turunan masih nonlinier, maka dibutuhkan metode iterasi Newton-Raphson [14]. Pengujian signifikansi parameter model regresi logistik dilakukan secara simultan dan secara parsial. Pengujian secara simultan dilandaskan pada hipotesis:

$H_0 : \beta_1 = \beta_2 = ... = \beta_j = 0$ (tidak ada pengaruh antara peubah prediktor terhadap peubah respon)

H_1: paling sedikit ada satu $\beta_j \neq 0$ (ada pengaruh antara peubah prediktor terhadap peubah respon)

dengan statistik uji G adalah:

$$-2\ln\left[\frac{L_0}{L_1}\right] \sim X^2(p) \quad (13)$$

Dengan X^2 adalah derajat bebas yang nilainya sama dengan banyaknya parameter, di mana H_0 akan ditolak jika nilai statistik uji G ≥ dengan tingkat kepercayaan (1-α)100. Sedangkan pengujian secara parsial dilandaskan pada hipotesis:

H0: $\beta_j = 0$ (tidak ada pengaruh antara masing-masing peubah prediktor terhadap peubah respon)

H1: $\beta_j \neq 0$ (ada pengaruh antara masing-masing peubah prediktor terhadap peubah respon)

Rumus statistik uji *Wald* adalah:

$$\left[\frac{\beta}{Se(\beta_j)}\right] \square Z \quad ; j = 0,1,2, ...,p \quad (14)$$

Hipotesis nol ditolak jika |W| > Zα/2 artinya peubah prediktor berpengaruh nyata terhadap peubah respon.

Hosmer dan Stanley [14] menyatakan bahwa peubah respon dengan dua kategori (biner) dengan ketentuan jika $\pi(x) \geq 0.5$ maka hasil prediksi adalah 1, jika $\pi(x) < 0.5$ maka hasil prediksi adalah 0. Klasifikasi menggunakan model peluang dengan persamaan sebagai berikut:

$$\text{logit } \pi(x_i) = \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} \quad (15)$$

Dengan π(x) adalah peluang kejadian sukses dan i adalah kategori atau kelas data (email).

I. Pengukuran Performa

Pengukuran performa dilakukan untuk melihat hasil yang didapatkan dari klasifikasi. Terdapat beberapa cara untuk mengukur performa, beberapa cara yang sering digunakan adalah dengan menghitung akurasi dan *False Positive Ratio* [18].

Akurasi merupakan persentase dari total dokumen yang teridentifikasi secara tepat dalam proses klasifikasi.

$$akurasi = \frac{jumlah\ klasifikasi\ benar}{jumlah\ dokumen\ uji\ coba} \times 100\% \quad (16)$$

False positive ratio adalah persetase dari jumlah email Ad yang gagal dikenali dibandingkan dengan jumlah email normal.

$$FPR = \frac{\text{jumlah False Positive}}{\text{jumlah email normal (bukan Ads)}} \times 100\% \quad (17)$$

III. METODELOGI PENELITIAN

A. Sumber Data

Data yang digunakan merupakan email berbahasa inggris berjumlah 5056 yang merupakan inbox dari sebuah perusahaan selama kurun waktu 3 bulan http://spamassassin.apache.org/publiccorpus/.
Email dikategorikan menjadi dua, yaitu email yang mengandung iklan atau *Ads* dan email yang tidak mengandung iklan. Dengan proporsi jumlah email iklan sebanyak 1549 dan email yang tidak mengandung iklan sebanyak 3507. Email diproses menggunakan software R dan sublime Text 2 (Unregistered Version).

B. Langkah Analisis

1. Menyiapkan data email, membaginya menjadi email berisi iklan dan tidak berisi iklan.
   a) Email yang diambil untuk tahap pelatihan dan pengujian merupakan sampel dari masing-masing kategori email.
   b) Data sampel tersebut dibagi menjadi data *training* dan data *testing* dengan proporsi 50:50, 60:40, 70:30, 80:20, dan 90:10.

2. Praproses Teks

Information gain bisa dianggap masuk ke dalam praproses teks, digunakan untuk mengukur efektivitas suatu atribut dalam mengklasifikasikan data. Tahapan dalam *Information Gain* adalah dengan membuat:

a. *Wordlist* atau *dictionary*, dari seluruh jumlah email (5000) dikodekan menjadi angka-angka, dengan cara dibuat daftar seluruh kata yang pernah muncul pada email tersebut. Misal dari seluruh email ditemukan terdapat 200.000 kata, maka setiap kata dalam tiap email akan diganti dengan menggunakan angka-angka berdasar jumlah angka yang ditemukan. Dengan kata lain tiap kata akan menjadi variabel predictor untuk respon jenis email.
b. *Feature Vectors Content*, Memilih sebanyak 100 kata dari keseluruhan kata yang terdapat pada email. Tiap kata mengadung sebuah fitur yang di dalamnya terkandung bobot untuk tiap-tiap kata.
c. *Features (Words) Selection*, setelah diambil ketentuan tentang banyaknya *Vectors Content* yang diambil, maka selanjutkan akan dipilih kata-kata yang mewakili atau merepresentasikan email dengan bobot tertinggi yang kemudian dipersiapkan untuk *Vectors Content*.
d. Melakukan pengacakan pada data yang digunakan sebai input baik pada data *training* maupun data *testing* agar klasifikasi semakin baik.

3. Klasifikasi email menggunakan jaringan syaraf tiruan *perceptron*.
   Algoritma pelatihan perceptron adalah sebagai berikut:
   a. Inisialisasi semua bobot dan bias (umumnya wi = b 0)
   b. Tentukan laju pemahaman (=α). Untuk penyederhanaan biasanya α diberi nilai = 1 Selama ada elemen vector masukan yang respon unit keluarannya tidak sama dengan target, lakukan:
      1) Set aktivasi unit masukan xi = si (i = 1,…n)
      2) Hitung respon unit keluaran: net = Σ xiwi + b
      3) Perbaiki bobot pola yang mengandung kesalahan (y ≠ t)

   Ada beberapa hal yang perlu diperhatikan dalam algoritma tersebut:
   a) Iterasi dilakukan terus hingga semua pola memiliki keluaran jaringan yang sama
   b) dengan targetnya (jaringan sudah memahami pola). Iterasi tidak berhenti setelah semua pola dimasukan.
   c) Perubahan bobot hanya dilakukan pada pola yang mengandung kesalahan (keluaran jaringan ≠ target). Perubahan tersebut merupakan hasil kali unit masukan dengan target laju pemahaman. Perubahan bobot hanya akan terjadi kalau unit masukan ≠ 0.
4. Klasifikasi teks menggunakan NBC dengan tahapan
   a) Membagi data menjadi *testing* dan *training*, pada data *training* telah diketahui jenis dari kategori email.
   b) Menghitung probabilitas dari V_j, dimana V_j merupakan kagetori email, yaitu j_1 = Non-*Ads*, j_2 = *Ads.*

c) Menghitung probabilitas kata w_k pada kategori v_j.
d) Model probabilitas NBC disimpan dan digunakan untuk tahap data *testing*.
e) Menghitung probabilitas tertinggi dari semua kategori yang diujikan (V_{MAP}).
f) Mencari nilai V_{MAP} paling maksimum dan memasukkan email tersebut pada kategori dengan V_{MAP} maksimum.
g) Menghitung nilai akurasi dari model yang terbentuk.

5. Klasifikasi email menggunakan Regresi Logistik dengan tahapan
   a) Membagi data menjadi *testing* dan *training*, pada data training telah diketahui jenis dari kategori email.
   b) Melakukan uji independensi dengan menggunakan data training.
   c) Membuat model regresi logistik antara variabel bebas dan variabel terikat kemudian melakukan pengujian serentak dan parsial terhadap model yang diperoleh.
   d) Mengintepretasi model logistic biner dan juga odds ratio yang diperoleh.
   e) Menghitung ketepatan klasifikasi model regresi logistik.
6. Membandingkan performansi metode jaringan syaraf *perceptron*, NBC, dan Regresi Logistik berdasarkan tingkat akurasi ketepatan klasifikasi, *False positive ratio* dan waktu yang digunkana untuk *running*.

IV. ANALISIS DAN PEMBAHASAN

A. Praproses Data

Data email yang telah dikumpulkan kemudian dilakukan praproses yaitu menentukan *wordlist* dan *Feature Vector Content*, Setelah itu akan dipilih 100 varibel dengan bobot tertinggi.

Tabel 1
Hasil Analisa Statistika Deskriptif

Word Vector (Word Selection)	
Kata	**Kata**
Ramp	!
Ads	Cc
Illu	http
2000	Hpl
enron	your
:	:
:	:

Selanjutnnya kata dengan jumlah huruf kurang dari 2 akan ditambahkan suatu pengenal agar dapat terbaca oleh *software R* saat dilakukan pemanggilan *word vector*. Hasil yang didapat akan tampak seperti berikut,

Tabel 2
Penambahan Suatu Pengenal Pada Kata

Kata		Hasil
!	-->	char_!
Cc	-->	som_cc

Berikut merupakan frekwensi kata yang muncul pada tiap kategori Email:

Tabel 3
Frekwensi Kemunculan Kata pada Email

Non-Ads	Jumlah	Iklan (Ads)	Jumlah
2000	4290	2000	78
ads	2	Ads	1900
char_!	1083	char_!	2459
enron	6293	Enron	0
hpl	2416	Hpl	0
http	233	http	983
Illu	0	Illu	1201
ramp	27311	Ramp	2
som_cc	1719	som_cc	12
your	1867	your	1952
:	:	:	:
:	:	:	:

Dalam uji *Naive Bayes Classifier* (NBC) maupun Jaringan Syaraf Tiruan *Perceptron* ini akan dibagi menjadi dua yaitu data *training* dan data *testing* dengan partisi sebesar 50, 60, 70, 80, 90 pada data training.

B. Klasifikasi Email Iklan menggunakan Metode Naive Bayes Classifier

Naive Bayes merupakan suatu metode klasifikasi dengan menggunakan probabilitas keanggotaan dalam suatu kelas. Pada penelitian ini, *Naive Bayes* digunakan untuk mengklasifikasikan apakah suatu email yang masuk mengandung iklan ataukah tidak. Berikut adalah hasil klasifikasi email iklan menggunakan metode Naïve Bayes.

Tabel 4.
Performa NBC pada Tiap Partisi

	Training		Testing	
	Error		Error	
Dataset	**Train**	**Akurasi**	**Test**	**Akurasi**
50: 50	0.057	0.943	*0.054	0.946
60: 40	0.076	0.924	0.079	0.921
70: 30	0.069	0.931	0.095	0.905
80: 20	*0.052	0.948	0.066	0.934
90: 10	0.075	0.925	0.085	0.915
Dataset	**falsepos train**		**falsepos test**	
50: 50	0.070		*0.074	
60: 40	0.105		0.106	
70: 30	0.094		0.127	
80: 20	*0.066		0.089	
90: 10	0.101		0.113	

dapat diketahui bila akurasi terbaik pada data *training* diperoleh pada partisi 80:20, dan 50:50 pada data *testing*, dan partisi yang sama untuk *false positive ratio.*

1. Intervensi Probabilitas Prior

nilai *false positive ratio* pada data *training* dan *testing* masih tinggi dan melebihi error *rate*, hal ini harus ditanggulangi mengingat *false positive* (email bukan iklan yang terklasifikasi sebagai email iklan) pada penelitian berkaitan dengan email selalu diupayakan untuk mencapai angka 0 dengan rasio 0%. Maka dilakukan intervensi probabilitas prior.

Tabel 5.
Performa NBC dengan Epsilon Range 0.2

	Training		Testing	
Dataset	**Error**	**Akurasi**	**Error**	**Akurasi**
50: 50	0.095	0.905	0.103	0.897
60: 40	0.103	0.897	0.098	0.902
70: 30	0.099	0.901	0.107	0.893
80: 20	0.105	0.895	0.086	*0.914
90: 10	*0.087	0.913	0.107	0.893
Dataset	**False Positive Ratio**		**False Positive Ratio**	
50: 50	0.0083		0.0064	
60: 40	0.0105		*0.0099	
70: 30	*0.0008		0.0140	
80: 20	0.0093		0.0141	
90: 10	0.0140		0.0141	

Berdasarkan Tabel 5 diketahui nilai error terendah untuk NBC dengan epsilon range sebesar 0.2 terletak pada partisi 90: 10 pada data *training* dan partisi 80: 20 pada data

testing. Sedangkan untuk *false positive ratio* terletak pada partisi 70: 30 pada data *training* dan 60: 40 pada data *testing. False Positive Ratio* terendah mencapai angka 0.08%, sudah sangat kecil hingga mendekati nol.

C. Perceptron

Algororitma pada perceptron merubah dan meng-*update* bobot hingga bobot tidak mengalami perubahan lagi Hasil klasifikasi *training testing* untuk tiap partisi data dengan menggunakan perceptron dengan bobot awal 0 bisa dilihat pada tabel berikut ini:

Tabel 6.
Performa Perceptron dengan Bobot Awal 0

Dataset	error train	Akurasi	error test	Akurasi
50: 50	0.213	0.787	0.108	0.892
60: 40	0.374	0.627	0.254	0.746
70: 30	0.244	0.756	0.085	0.915
80: 20	0.288	0.712	0.086	0.914
90: 10	*0.197	0.803	*0.072	0.928

Dataset	error test w	Akurasi	falsepos
50: 50	0.163	0.837	0.115
60: 40	0.208	0.792	0.146
70: 30	0.139	0.861	0.098
80: 20	*0.136	0.864	*0.097
90: 10	0.139	0.861	0.098

Berdasarkan Tabel 6 diketahui bahwa nilai akurasi pada error tanpa meng-*update* bobot pada data *testing* lebih tinggi ketimbang melakukan *update* bobot pada tiap iterasi pelatihan dalam data *testing*.

1. *Global Optimization pada Perceptron*

Permasalahan optimasi terletak pada tujuan untuk mendapatkan atau menemukan solusi terbaik dari semua solusi yang mungkin. Untuk menemukan kemungkinan lain dalam memperoleh hasil klasifikasi, dalam penelitian ini akan digunakan nilai initial bobot atau bobot awal yang berbeda-beda.

Tabel 7
Performa Perceptron dengan Bobot Awal yang Berbeda-Beda (Partisi Data 70: 30)

	seed 1	seed 2	seed 3	seed 4	seed 5
error train	0.2442	*0.1762	0.1801	0.2155	0.3517
error test	0.0851	*0.0760	0.0863	0.0863	0.2751
error test w	0.1392	*0.1224	0.1392	0.1302	0.3241
falsepos	0.0980	*0.0862	0.0980	0.0917	0.2278
	seed 6	**seed 7**	**seed 8**	**seed 9**	**seed 10**
error train	0.3445	0.2793	0.3119	0.3373	0.3445
error test	0.2809	0.2893	0.2887	0.2822	0.2822
error test w	0.3254	0.1656	0.2532	0.3177	0.3125
falsepos	0.2287	0.1162	0.1779	0.2232	0.2196
	seed 11	**seed 12**	**seed 13**	**seed 14**	
error train	0.3599	0.3616	0.3870	0.4887	
error test	0.2738	0.2738	0.2622	0.6211	
error test w	0.3151	0.3138	0.2957	0.3840	
falsepos	0.2214	0.2205	0.2078	0.2704	

. Dari tabel 7 di atas diketahui bahwa error terkecil terletak pada run ke 2

D. Regresi Logistik

Metode ketiga yang digunakan untuk klasifikasi email iklan adalah regesi logistik. Pada bab regresi logistik ini akan dihitung ketepatan klasifikasi seperti pada kedua metode sebelumnya.

Tabel 8
Performa Regresi Logistik

Dataset	error train	akurasi	error test	akurasi	falsepos
50:50	0.190	0.810	0.098	0.902	0.151
60:40	0.120	0.880	0.212	0.788	0.198
70:30	0.180	0.820	0.129	0.871	0.102
80:20	0.108	0.892	0.106	0.894	0.197
90:10	0.099	0.901	0.110	0.890	0.102

Dari tabel di atas dapat diketahui bahwa akurasi terbaik dari regresi logistik terdapat pada partisi 70:30 yaitu saat false positive ratio menyentuh angka terkecil.

E. Perbandingan NBC, Perceptron, dan Regresi Logistik

Setelah mengetahui hasil masing-masing ketepatan klasifikasi pada ketiga metode maka langkah selanjutnya adalah membandingkannya. Berikut merupakan perbandingan antara kedua metode berdasarkan akurasi, dan *false positive ratio.* Untuk nilai yang diambil sebagi pembanding adalah nilai partisi yang memiliki *false positive ratio* terkecil.

Tabel 9
Perbandingan Hasil Ketepatan Klasifikasi Antara NBC dan Perceptron data *Testing*

	Testing			
	Tanpa Optimasi		**Optimasi NBC (Intervensi Prob Prior) Perceptron (Bobot Awal)**	
Metode	Akurasi	FP Ratio	Akurasi	FP Ratio
NBC	*0.946	*0.074	0.902	*0.009
Perceptron	0.928	0.098	*0.924	0.0862
Regresi Logistik	0.871	0.102	-	-

NBC lebih unggul atau lebih baik dibanding Perceptron dan Regresi Logistik dalam mengklasifikasikan email iklan pada data *Testing* tanpa optimasi, namun akurasi NBC berada di bawah Perceptron pada data *Testing* dengan optimasi, hal ini dikarenakan *False Positive Ratio* NBC yang mampu ditekan sampai mendekati 0, yaitu 0,9%. Berdasarkan hasil di atas, pada NBC *False Positive Ratio* lebih mudah untuk dikontrol.

V. KESIMPULAN DAN SARAN

Berdasarkan analisis diperoleh kesimpulan Metode *Naive Bayes Classifier* dapat melakukan klasifikasi email iklan dan non iklan dengan sangat baik. Hasil akurasi tertinggi yang didapatkan pada saat data *testing* tanpa intervensi *probabilitas prior* adalah 94,6% dengan *False Positive Ratio* 7,4%. Dan dengan intervensi Probabilitas Prior menghasilkan akurasi 90,2% dan *False Positive Ratio* 0,9%, Perceptron dalam melakukan klasifikasi email juga menghasilkan akurasi yang cukup baik. Menggunakan data *testing* tanpa optimasi didapatkan Akurasi 92,8% dan *False Positive Ratio* 9,8% dan dengan optimasi menghasilkan akurasi 92,4% dan *False Positive Ratio* 8,62%. Regresi Logistik memiliki tingkat *false positive ratio* tertinggi pada partisi data 70: 30, yaitu sebesar 0.102. Dan hasil penelitian menunjukkan bahwa pada NBC False Positive Ratio lebih mudah untuk dikontrol.

Saran yang dapat diberikan pada penelitian ini adalah perlunya *GUI* (Graphical User Interface) agar menjadi suatu bentuk otomatisasi yang benar-benar bisa diterapkan dalam kehidupan sehari-hari

DAFTAR PUSTAKA

[1] Suyanto. (2014). *Artificial Intelligence, Searching - Reasoning - Planning-Learning*. *Informatika*. Bandung: Informatika Bandung.

[2] Alo, Liliweri. (2011). *Komunikasi Serba Ada Serba Makna. Jakarta. Kencana Prenada Media Group.*

[3] Siang, J.J. 2005. *Jaringan Syaraf Tiruan & Pemrogramannya Menggunakan MATLAB*. Yogyakarta: ANDI.

[4] Kufadinimbwa, Owen & Gotora, Richard (2012). Spam Detection Using Artificial Neural Networks (Perceptron Leraning Rule). *Department of Computer Science, Faculty of Sciences, University of Zimbabwe.*

[5] Miller, T. (2005). *Data and Text Mining A Business Application.* New Jersey, USA: Prentice Hall.

[6] Darujati, C., & Gumelar, A.B. (2012). Pemanfaatan Teknik Supervised Untuk Klasifikasi Teks Bahasa Indonesia, *Jurnal Bandung Text Mining*, 16(1),pp.5-1 – 5-8.

[7] Anugroho, Prastyo & Winarno Idris (2012). Klasifikasi Email Spam dengan Menggunakan Metode Naïve Bayes Classifier Menggunakan Java Programming. *Politeknik Negeri Surabaya* [14] Agresti, A., (2002). *Categorical Data Analysis Second Edition*. New York: John Wiley & Sons.

[8] Kurniawan, B., Effendi, S., & Sitompul, O. S. (2012). Klasifikasi Konten Berita Dengan Metode. *JURNAL DUNIA TEKNOLOGI INFORMASI Vol. 1, No. 1*, 14-19.

[9] Lestari, N. A., Putra, I. G., & Cahyawan, A. A. (2013). Personality Types Classification for Indonesian Text in Partners Searching Website Using Naïve Bayes Methods. *IJCSI International Journal of* Asian, J. A. (2007). Stemming Indonesian: A Confix-Stripping Approach. *ACM Trnsactions on Asian Language Information Processing (TALIP), 6(4),*, 1-33.

[10] Kasali, Rhenald. (1995). *"Manajemen Periklanan"*. Pustaka Grafiti, Jakarta.

[11] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools and Techniques.* USA: Elsevier.

[12] Moch. Agus Taufiqurrahman, Suyanto, Moch Arif Bijaksana, 2007, "Analisa Dan Implementasi Personalized *Spam Filtering* Menggunakan Jaringan Syaraf Tiruan" Jurusan Teknik Informatika, STT Telkom, Bndung.

[13] Agresti, A., (2002). *Categorical Data Analysis Second Edition*. New York: John Wiley & Sons.

[14] Hosmer, David W. & Lemeshow, Stanley. (2000). *Applied Logistik Regression*. Willey