Jurnal Teknovasi
Volume 03, Nomor 2, 2016, 51 – 57
ISSN : 2355-701X

ANALISIS SPAM DENGAN MENGGUNAKAN NAÏVE BAYES

Darma Juang[1]*
[1]Magister Teknik Informatika, Univertas Sumatera Utara
Jl. Almamater Kampus USU, Medan 20155,Telp 061- 8219005, Fax 061-8213250
*Email : juanglp3i@gmail.com

ABSTRAK

Perkembangan teknologi internet sekarang berkembang secara pesat.banyak bentuk – bentuk penyerangan terhadap sebuah situs web site, baik itu dilakukan oleh hacker, cracker, ataupun virus. Dengan makin banyaknya virus, kenyamanan saat berinternet-an pun ikut berkurang. Salah satu fasilitas internet yang sering kita gunakan adalah e-mail. Akhir-akhir ini e-mail merupakan hal yang sangat penting bagi kita,manfaatnya sering kita rasakan. *SPAM* (Stupid pointless Annoying Message) yang berarti e-mail sampah atau email yang tidak kita butuhkan.,merupakan kata yang sering didengar.Sebenarnya email yang dianggap *SPAM* itu tergantung dari sudut pandang masing-masing.Tujuan penelitian ini dilakukan untuk memahami cara kerja metode naive bayes di dalam menganalisis spam berdasarkan frekuensi kata.

Kata Kunci : Spam, Naïve Bayes, Frekuensi Kata.

PENDAHULUAN

Perkembangan teknologi internet sekarang berkembang secara pesat.Banyak manfaat yang dirasakan masyarakat dengan adanya teknologi ini. banyak bentuk – bentuk penyerangan terhadap sebuah situs web site, baik itu dilakukan oleh hacker, crakcker, ataupun virus. Dengan makin banyaknya virus, kenyamanan saat berinternet-an pun ikut berkurang. Salah satu fasilitas internet yang sering kita gunakan adalah e-mail. Akhir-akhir ini e-mail merupakan hal yang sangat penting bagi kita,manfaatnya sering kita rasakan. Kita dapat menerima informasi atau bertukar pesan dengancepat dan mudah. Fasilitas ini dapat digunakan jika terhubung ke internet (sebagian besar penggunaan e-mail terhubung ke internet kecuali dalam jaringan LAN/intranet).

Email salah satu media dalam penyebaran virus di internet. Banyak virus, trojan, spyware terus menyebar melalui email, dengan subject yang menarik. Biasanya, komputer yang sudah terkena satu virus, ikut menyebarluaskannya dengan mengirimkan email lain berisi virus ke listing email yang lainnya. *SPAM* (Stupid pointless Annoying Message) yang berarti e-mail sampah atau email yang tidak kita butuhkan.,merupakan kata yang sering didengar.Sebenarnya email yang dianggap *SPAM* itu tergantung dari sudut pandang masing-masing. Contoh doni yang seorang pebisnis hotel mendapatkan email yang berisi penawaran produk wallpaper untuk hotelnya walaupun email tersebut bukan yang diinginkan tetapi email tersebut bukanlah sebagai SPAM akan tetapi lain halnya andi seorang tenaga pengajar email tersebut sangat tidak ia ingkin karena tidak ada kebutuhan informasi.

Dampak negatif *SPAM* ini adalah waktu yang terbuang sia-sia. Saat mengecek email yang ada pada inbox umumya haruslah memilih *email* yang sesuai ,tentu saja pengguna akan langsung menghapus *SPAM* saat menemukannya hal ini akan menghabiskan waktu terkadang tak jarang *email* yang penting pun turut ikut terhapus,hal ini tentu saja sangat merugikan. Bagi orang-orang yang tidak dapat leluasa koneksi ke internet,bandwidth yang dibatasi oleh quota,quota pada inbox email akan sangat merugikan jika terkena dari SPAM ini. Fasilitas *e-mail* yang murah dan kemudahan untuk mengirimkan ke berapa pun jumlah penerima, maka *spam mail* menjadi semakin merajalela. Pada survey yang dilakukan oleh Cranor & La Macchia (1998), ditemukan bahwa 10% dari mail yang diterima oleh suatu perusahaan adalah *spam-mail* atau spammer.

Jurnal Teknovasi
Volume 03, Nomor 2, 2016, 51 – 57
ISSN : 2355-701X

Kebutuhan dunia bisnis yang ingin mendapatkan nilai tambah dari data yang telah terkumpul, mendorong penerapan teknik pengolahan data dari berbagai bidang pengetahuan seperti statistika dan kecerdasan buatan. Ternyata penerapan teknik tersebut memberikan tantangan baru yang akhirnya memunculkan metode baru yang disebut data mining. Ada beberapa definisi data mining yang dikenal dari berbagai sumber, diantaranya adalah :

1. Data mining adalah pencarian dan teknik analisa data yang besar untuk menemukan pola dan aturan yang berarti (Berry & Linoff, 2004).
2. Data mining adalah teknik untuk menganalisa sekumpulan data yang besar guna menemukan hubungan yang tidak diduga dan berguna bagi pemilik data (Hand, 2001).
3. Data mining adalah proses untuk menemukan pola dan hubungan dalam suatu data (Hornick, 2007).
4. Data mining adalah perangkat lunak untuk menemukan pola-pola tersembunyi dalam database yang besar dan menghasilkan aturan-aturan yang digunakan untuk memperkirakan perilaku di masa depan (Kadir, 2003)

Tahap-tahap data mining ada 6 yaitu :

1. Pembersihan data (data cleaning). Pembersihan data merupakan proses menghilangkan noise dan data yang tidak relevan. Pada umumnya data yang didapat, baik dari database memiliki isian-isian yang tidak sempurna seperti data yang hilang, data yang tidak valid atau juga hanya sekedar salah ketik.Data yang tidak relevan itu juga lebih baik dibuang. Pembersihan data juga akan mempengaruhi performasi dari teknik data mining karena data yang ditangani akan berkurang jumlah dan kompleksitasnya.
2. Integrasi data (data integration). Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru.Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya. Sebagai contoh bila integrasi data berdasarkan jenis produk ternyata menggabungkan produk dari kategori yang berbeda maka akan didapatkan korelasi antar produk yang sebenarnya tidak ada.
3. Seleksi Data (Data Selection).Data yang ada pada database sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari database. Sebagai contoh, sebuah kasus yang meneliti factor kecenderungan orang membeli dalam kasus market basket analysis, tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan saja.
4. Transformasi data (Data Transformation). Data diubah atau digabung ke dalam format yang sesuai untuk diproses. Sebagai contoh beberapa metode standar seperti analisis asosiasi dan clustering hanya bisa menerima input data kategorikal.Karenanya data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa interval. Proses ini sering disebut transformasi data.
5. Proses mining. Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.
6. Evaluasi pola (pattern evaluation). Untuk mengidentifikasi pola-pola menarik kedalam knowledge based yang ditemukan. Dalam tahap ini hasilnya berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesayang ada memang tercapai.
7. Presentasi pengetahuan (knowledge presentation). Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.Tahap terakhir adalah bagaimana memformulasikan keputusan atau aksi dari hasil analisis yang didapat? presentasi dalam bentuk pengetahuan yang bisa dipahami semua orang adalah satu tahapan yang diperlukan.Dalam presentasi ini, visualisasi juga bisa membantu mengkomunikasikan hasil data mining (Han, 2006).

Preprocessing

Preprocessing merupakan tahapan awal dalam mengolah data input sebelum memasuki proses tahapan utama dari metode Latent Semantic Analysis (LSA). Preprocessing text dilakukan untuk tujuan penyeragaman dan kemudahan pembacaan serta proses LSA. Preprocessing terdiri

Jurnal Teknovasi
Volume 03, Nomor 2, 2016, 51 – 57
ISSN : 2355-701X

dari beberapa tahapan. Adapun tahapan preprocessing berdasarkan (Triawati, 2009) , yaitu: case folding,tokenizing / parsing, filtering, stemming. Berikut penjelasan empat tahapan dalam proses preprocessing adalah sebagai berikut.

1. *Case Folding.*
   Case folding merupakan tahapan 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)(Triawati, 2009).Contoh penggunaan case folding adalah sebagai berikut.

data input hasil *case folding*

Gambar 1. Proses Case Folding

Penjelasan:

Tabel 1. Case Folding

Input	Output
Kalimat/kata input dari pengguna	Kalimat/kata input menjadi huruf kecil serta tanpa karakter lain selain karakter huruf 'a-z'

2. *Tokenizing*
   Tahap *tokenizing / parsing* adalah tahap pemotongan kata berdasarkan tiap kata yang menyusunnya(Triawati, 2009). Selain itu, *spasi* digunakan untuk memisahkan antar kata.

data input hasil *tokenizing / parsing*

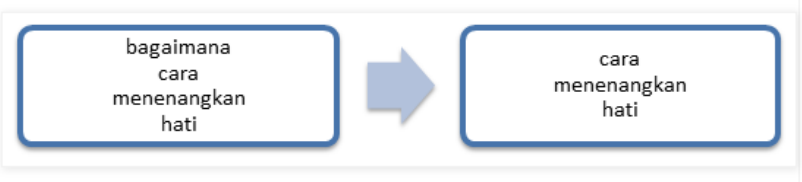
Gambar 2. Proses Tokenizing

Penjelasan:

Tabel 2. Tokenizing

Input	Output
Kalimat/kata input hasil dari proses *case folding*	Kumpulan kata

3. *Filtering*
   Tahap *filtering* adalah tahap mengambil kata - kata penting dari hasil *tokenizing*. Proses filtering dapat menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata penting). *Stoplist / stopword* adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. Contoh *stopword* adalah "yang", "dan", "di", "dari" dan lain – lain.(Triawati, 2009).



data input hasil *filtering*

Gambar 3. Proses filtering

Penjelasan:

Tabel 3. Filtering

Input	Output
Kumpulan kata hasil dari proses *tokenizing/parsing*	Kumpulan term yang siap untuk diolah dengan proses svd

Naive Bayes

Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma mengunakan teorema Bayes dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas.Definisi lain mengatakan Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Bustami, 2013).

Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Keuntungan penggunaan Naive Bayes adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (Training Data) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. Naive Bayes sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan (Pattekari,2012)

Persamaan Metode Naive Bayes

Persamaan dari teorema naïve bayes adalah

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

Dimana :

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 Hipoteses H (prior probabilitas)
P(X|H) = Probabilitas X berdasarkan kondisi pada hipotesis H
P(X) = Probabilitas X

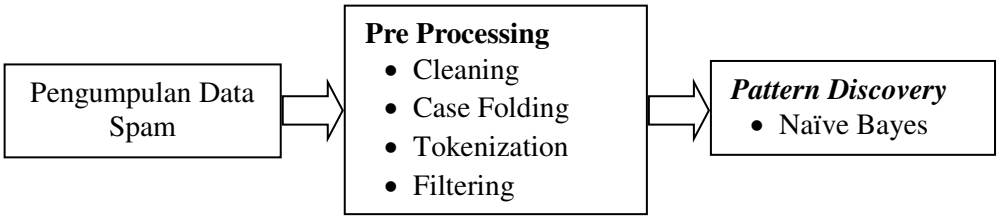
METODE PENELITIAN

Dalam pembahasan di jurnal ini, penulis ingin menganalisis spam dengan menggunakan metode naive bayes. Hasil yang diharapkan yaitu metode naïve bayes dalam memberikan nilai pengukuran dalam menganalisis spam berdasarkan frekuensi kata spam dan non spam. Penelitian dilakukan menggunakan *email* berbahasa Inggris tersimpan dalam bentuk document txt. Data yang digunakan diambil diambil dari lingspam dataset. Teknik pengumpulan data yang digunakan peneliti dalam pengumpulan data adalah Mengumpulkan literatur, jurnal, paper, dan bacaan-bacaan lainnya yang berhubungan dengan algoritma klasifikasi data mining. Mengumpulkan data penelitian yang diperoleh secara online dari Ling spam dataset.

Jurnal Teknovasi
Volume 03, Nomor 2, 2016, 51 – 57
ISSN : 2355-701X

Proses Analisis Spam pada Dokumen

Berikut ini adalah metode yang digunakan untuk proses analisis yang digunakan dalam penelitian ini.



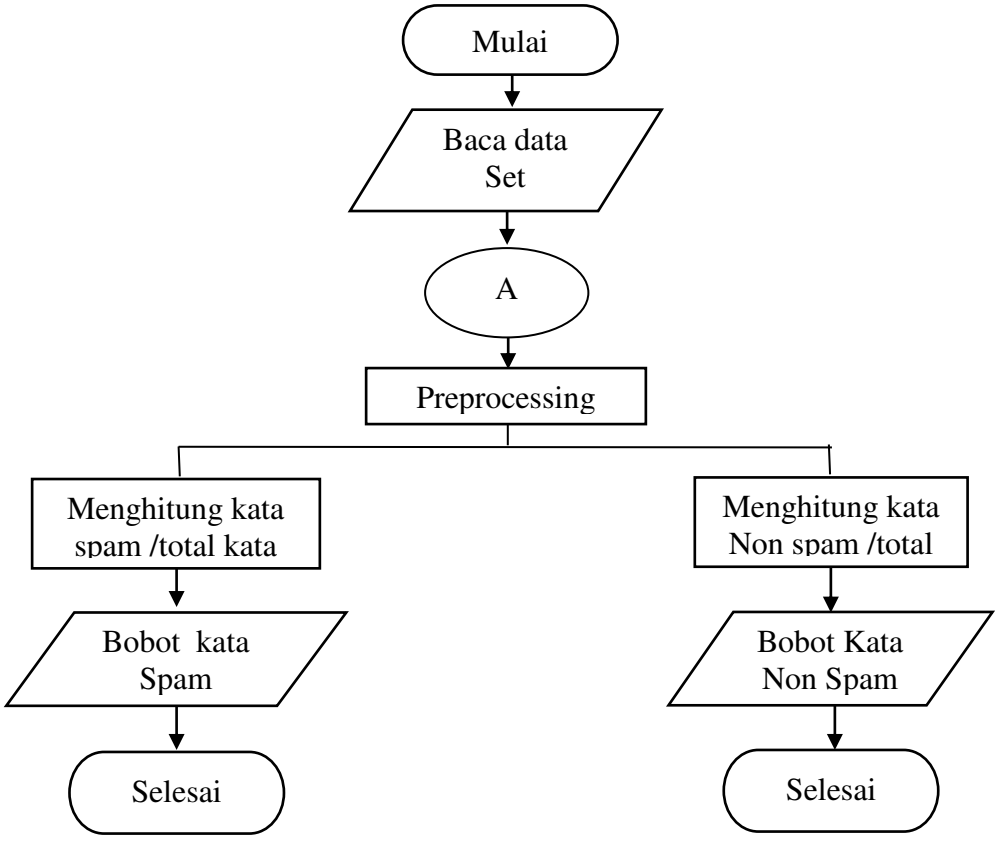
Gambar 5. Proses Analisis Spam pada Dokumen

Peprocessing

Pada tahapan ini yang dilakukan yaitu case folding, mengubah semua huruf dalam teks menjadi huruf kecil. Kemudian dilakukan proses parsing. Parsing yang digunakan adalah parsing sederhana yaitu memecah sebuah teks menjadi kumpulan kata-kata tanpa memperhatikan keterkaitan antar kata dan peran atau kedudukannya dalam kalimat. Hasil yang diperoleh yaitu karakter huruf saja. Selanjutnya proses yang dilakukan yaitu penghilangan stopword. Pada saat selesai melakukan parsing, sistem akan melakukan pengecekan ke dalam daftar stopword. Jika kata merupakan stopword maka kata itu akan dibuang. .

Perancangan *Pattern Discovery*

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



Gambar 6. *Flowchart* patern discovery

Jurnal Teknovasi
Volume 03, Nomor 2, 2016, 51 – 57
ISSN : 2355-701X

Ket :
Tahapan yang harus dilakukan pada proses naïve bayes adalah

1. User memasukkan dokumen spam yang ingin dianalisis
2. Sistem akan melakukan processing pada teks dokumen
3. Sistem akan menampilkan total kata ,jumlah spam serta tidak spam
4. Sistem akan memproses untuk menentukan nilai bobot dari masing-masing

Analisis Proses Perhitungan dengan metode naïve bayes

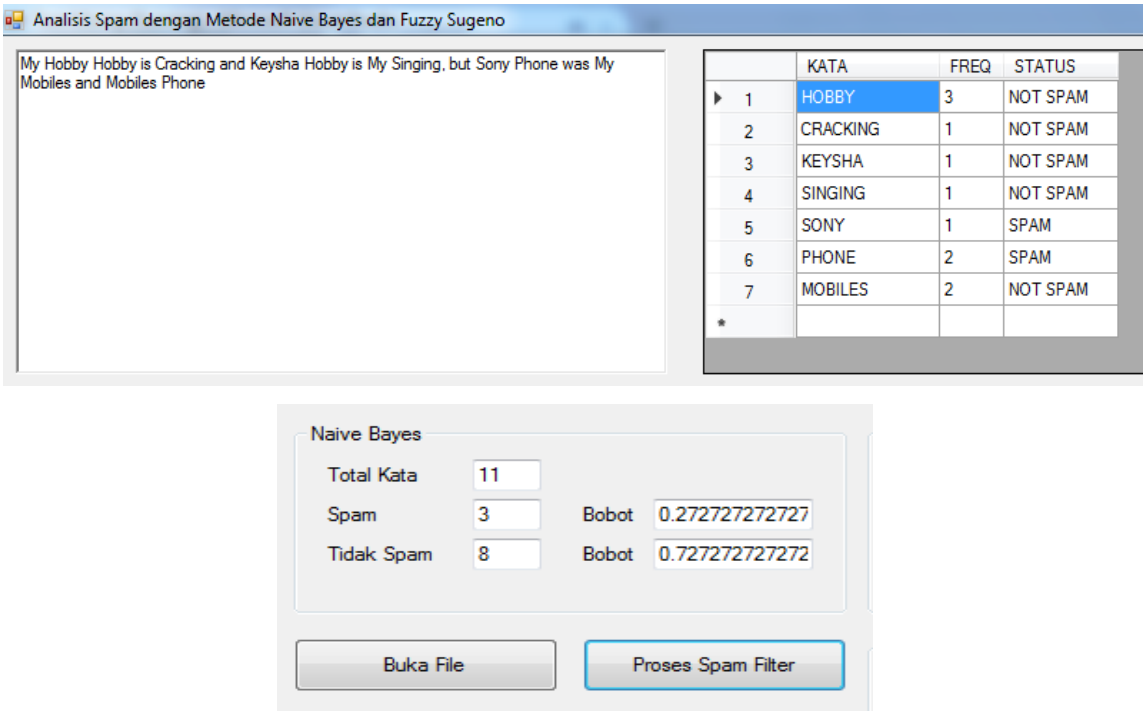
Pada tahapan ini document mengalami proses preprocessing. Setelah itu dipecah menjadi token. Token tersebut dihitung berdasarkan frekuensi katanya yang kemudian dikategorikan berdasarkan spam dan non spam**.** Adapun proses perhitungan seperti contoh berikut :

Tabel 4. Analisis Proses Perhitungan dengan metode naïve bayes

Dokumen Kata My Hobby Hobby is Cracking and Keysha Hobby is My Singing, but Sony Phone was My Mobiles and Mobiles Phone				
Kata	Frek	Status	Proses Analisis	
Hobby	3	Not Spam	Total Kata = 11	
Cracking	1	Not Spam	Spam = 3	bobot spam = 0.27
Keysha	1	Not Spam	Not Spam = 8	bobot not spam = 0.72
Singing	1	Not Spam		
Sony	1	Spam		
Phone	2	Spam		
Mobile	2	Not Spam		

Pengujian

Berdasarkan hasil pengolahan data pada masing-masing dokumen spam, dapat diperoleh nilai-nilai,yang mana nilai tersebut dapat digunakan untuk klasifikasi sebagai spam atau bukan spam. Form menu aplikasi ini, digunakan untuk melihat proses analisis spam dengan menggunakan algoritma naïve bayes.



Gambar 7. Pengujian Spam

Jurnal Teknovasi
Volume 03, Nomor 2, 2016, 51 – 57
ISSN : 2355-701X

Pengujian Dokumen Spam Dan Non Spam Dengan Naïve Bayes

Adapun perhitungan dengan menggunakan rasio terbobot pada setiap variabel terlihat pada tabel 5.

Tabel 5. Pengujian dokumen spam dan non spam dengan naïve bayes

No	Document	Total Kata	Spam	Bukan Spam	Bobot Kata Spam	Bobot Kata Bukan Spam	Kategori
1	spmsga1	188	124	64	0.65	0.34	spam
2	spmsga2	27	15	12	0.55	0.44	spam
3	spmsga3	42	27	15	0.64	0.35	spam
4	spmsga4	106	81	25	0.76	0.23	spam
5	spmsga5	1409	980	422	0.69	0.30	spam
6	spmsga6	380	259	121	0.68	0.31	spam
7	spmsga7	26	22	4	0.84	0.15	spam
8	spmsga8	181	96	85	0.53	0.86	spam
9	spmsga9	44	26	18	0.59	0.40	spam
10	spmsga10	464	311	153	0.67	0.32	spam

Dari table 5. di atas dapat menerangkan jika nilai bobot kata spam yang lebih besar daripada bobot kata bukan spam maka dokumen tersebut tergolong spam.Jika dokumen email memiliki nilai bobot spam lebih rendah dari bobot spam makan dokumen email tersebut tergolong bukan spam.

KESIMPULAN

Dari hasil penelitian yang sudah dilakukan bahwa algoritma naïve bayes dapat mengklasifikasikan suatu pesan ke dalam dua kelas yaitu spam dan non spam. Dari pengklasifikasian tersebut sangat dipengaruhi oleh proses training sehingga dapat disimpulkan hasil pengklasifikasian yang disajikan dalam bentuk tabel dapat terlihat dengan jelas informasi kategori spam atau bukan spam. Diharapkan untuk penelitian selanjutnya dapat dilakukan perbandingan dengan metode algoritma yang lain dan data training yang lebih banyak lagi.

DAFTAR PUSTAKA

Agusta, L. 2009. *Perbandingan Algoritma* Stemming *Porter Dengan Algoritma Nazief & Adriani Untuk* Stemming *Dokumen Teks Bahasa Indonesia*. *Proceedings* Konferensi Nasional Sistem dan Informatika (KNS&I09-036), hlm 196-201.

Berry and Linof, 2004,data mining techniques for marketing ,sales,crm

Bustami., 2013, Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi*, TECHSI : Jurnal Penelitian Teknik Informatika*, Vol. 3, No.2, Hal. 127-146.

Candra Triawati. Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia. http://digilib.ittelkom.ac.id/index.php?option=com_content&view=article&id=590:text-mining&catid=20:informatika&Itemid=1 [21 November 2012]

David Hand,2001, David Hand, Heikki Mannila and Padhraic Smyth, Principles of Data Mining

Kadir,T. Brady,M Scale,Saliency dan Image Description, International Journal of Computer Vision. 45(2),83-105 (2001)

Mark F. Hornick, Erik Marcade, Sunil Venkayala: "Java Data Mining: Strategy, Standard, And Practice: A Practical Guide for Architecture, Design, And Implementation" (Broché) Lorrie Faith Cranor and Brian A. LaMacchia. Spam! *Communications of the ACM.* Vol. 41, No. 8 (Aug. 1998), Pages 74-83.

Pattekari, S. A., Parveen, A., 2012, Prediction System for Heart Disease Using Naïve Bayes, *International Journal of Advanced Computer and Mathematical Sciences,* I. ISSN 2230-9624, Vol. 3, No 3, Hal 290-294