DETEKSI WEB BERKONTEN PORNO DENGAN METODE *BAYESIAN FILTERING* DAN *PRINCIPAL COMPONENT ANALYSIS*

Afif Rizka Wandala
Jurusan Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami No. 36 A Surakarta
wafriz01@gmail.com

Sarngadi Palgunadi Yohanes
Jurusan Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami No. 36 A Surakarta
palgunadi@uns.ac.id

Abdul Aziz
Jurusan Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami No. 36 A Surakarta
aaziz@staff.uns.ac.id

ABSTRAK

Saat ini konten porno banyak bertebaran di sebuah website baik dalam bentuk konten utama maupun hanya sebuah iklan. Salah satu cara untuk mencegah konten yang tidak diinginkan tersebut adalah dengan pendeteksian konten. Pendeteksian konten dilakukan dengan proses text mining. Setiap website diprediksi dengan mempertimbangkan karakteristik text yang ada didalamnya. Metode yang digunakan untuk memprediksi web porno dalam penelitian ini adalah bayesian filtering,

Bayesian filtering berrfungsi memperhitungkan probabilitas kemiripan suatu website dengan membandingkan munculnya tiap keyword pada data latih. Namun, banyaknya keyword atau variable mempengaruhi efisiensi dan keakuratan deteksi. Untuk mengatasi hal tersebut, peneilitan ini menggunakan Principal component analysis untuk mengurangi dan mencari variable yang memiliki pengaruh penting terhadap deteksi.

Dari penelitian ini diperoleh tingkat akurasi tertinggi deteksi web berkonten porno dengan metode bayesian filtering sebesar 89.22%, hasil deteksi tersebut menggunakan 51 variable hasil ekstraksi Principal component analysis dengan data training terdiri dari 250 web porno dan 250 bukan porno.

Kata kunci : *Text mining*, *Bayesian Filtering*, PCA, deteksi, porno

1. PENDAHULUAN

Saat ini konten porno banyak bertebaran di sebuah website baik dalam bentuk konten utama maupun hanya sebuah iklan. Menurut survey yang dilakuan TopTenReviews, dari tahun 2003-2007 terdapat 4.2 juta situs porno atau 12% dari total websites di dunia. Dari jumlah sebanyak itu tentu akan sangat mudah kita temui ketika kita browsing di internet. Sumber yang sama menunjukan ada 34% kasus , bahwa web menampilkan informasi yang tidak sesuai dari konten yang dinginakan user dalam bentuk konten porno. Dalam menangani hal ini baik pemerintah ataupun masyarakat pengguna internet sudah menerapkan beberapa langkah pencegahan. Dari Kemenkominfo sendiri sudah memblokir sekitar 835.000 situs porno sejak 2010 hingga Juni 2012 berdomain id. Kemenkominfo menyeleksi web porno melalui pengaduan masyarakat melalui kanal trustpositif.kominfo.go.id. Hal ini sangat tidak sebanding dengan pertumbuhan situs porno yang meningkat 5-10 kali lipat dalam 3 tahunnya. Artinya jumlah pemblokiran situs dan pertumbuhan situs tiap tahunnya 1:3[1].

Metode pendeteksian yang ada menggunakan badword list atau pendeteksian *keyword* yang mengandung arti buruk, misalnya porn, xxx, fuck dsb. mencari *keyword* pada sebuah webpages kemudian memastikan apakah terdapat salah satu *keyword* , jika ada maka web yang di uji tersebut akan dianggap sebagai website porno. Metode ini cenderung memiliki banyak kelemahan, yang mana web yang sebenarnya bukan termasuk berkonten porno semisal berita juga akan terdeteksi sebagai situs porno. Salah satu cara untuk memperbaiki metode deteksi konten tersebut adalah menggunakan web mining. *Web mining* merupakan proses ekstraksi pola-pola atau karakteristik penting dan bermanfaat namun tersimpan secara implisit pada kumpulan data yang relatif besar pada layanan *world wide web*. *Web mining* sendiri teridiri atas tiga bagian yaitu: *web content mining*, *web structure mining*, dan *web usage mining* [3]. Namun pada kasus deteksi konten ini kita akan fokus pada *web content mining*. *Web content mining* merupakan suatu proses otomatis untuk menemukan informasi yang berguna dari dokumen atau data. Pada prinsipnya teknik ini mengekstraksi kata kunci yang terkandung pada dokumen. Isi data web antara lain dapat berupa teks, citra, audio, video, metadata, dan hyperlink. Ada dua strategi yang umum digunakan: pertama langsung melakukan *mining* terhadap data, dan kedua melakukan pencarian serta meng*improve* hasil pencarian seperti layaknya *search engine* [4].

Data yang digunakan dalam proses web mining didapat dari proses *crawling*, *crawling* merupakan proses penjelajahan halaman sebuah website yang secara berurutan sesuai daftar antrian URL yang sudah ada, kemudian menyimpan data dari website yang telah dijelajahi dan kemudian disimpan dalam bentuk teks atau metadata[5]. Kemudian dari data ini dilakukan proses training untuk mengenali karakteristik dari sebuah web. Proses training memerlukan dua jenis web yang mengandung kata porno dari web porno dan bukan porno.

Untuk mendeteksi sebuah web benar porno atau bukan diperlukan sebuah klasifikasi menggunakan probabilitas. Salah satu metode klasifikasi adalah Algoritma Bayesian Filtering. Bayesian filtering merupakan metode statistik yang digunakan untuk mencari sebuah kemungkinan munculnya kejadian dengan memperhitungkan kejadian lain. Bayesian filtering mampu memprediksi dan mengklasifikasikan

sebuah kejadian dengan menggunakan data latih. Selain itu Bayesian filter mampu memperhitungkan kemunculan sebuah kejadian yang belum pernah ditemui berdasarkan dengan data latih atau kumpulan kejadian yang dimilikinya tersebut. Dalam kasus deteksi porno ini sangat dimungkinkan untuk untuk memprediksi sebuah web dengan memperhitungkan munculnya sebuah *keyword* dari badword list dan mengklasifikasikannya dengan menghitung probabilitas dengan berdasar pada data latih[6].

Principal Component Analysis (PCA) mampu mereduksi jumlah *keyword*. Prinsip dasar dari algoritma PCA adalah mengurangi dimensi suatu set data namun tetap mempertahankan sebanyak mungkin informasi dalam set data tersebut. Secara matematis PCA mentrans-formasikan sejumlah variabel yang berkorelasi ke dalam bentuk yang bebas tidak berkorelasi. *Principal Component* satu dengan yang lain tidak saling berkorelasi dan diurutkan sedemikian rupa sehingga Principal *Component* yang pertama memuat paling banyak variasi dari data set[7]. Dalam deteksi web ini kita akan mencari variable paling berpengaruh kemudian mereduksi variable yang kurang pengaruh terhadap deteksi. Dengan dikombinasikannya *Bayesian Filtering* dan *Principal component analysis* ini diharapkan mendapatkan hasil deteksi lebih akurat dan lebih efektif.

2. DASAR TEORI

2.1 *Web crawling*

Web crawler merupakan suatu program yang dapat mendownload halaman dari suatu web dan biasanya digunakan dalam search engine. Crawler umumnya dimulai dengan inisial set dari URL , kemudian crawler mendownload isi dari halaman web yang telah didefinisikan tersebut. Web crawler juga dapat mengekstraks URL dalam sebuah website. URL yang ditemukan dalam halaman website tersebut dapat dicrawling lagi. Dari proses ekstraksi URL tersebut dapat diperoleh juga keterkaian antar web satu dengan lainnya[5].

2.2 Pengertian *Text Mining*

Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Sumber data yang digunakan pada *text mining* adalah kumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur[4].

2.2.1 *Text Preprocessing*

Berikut adalah tahapan dalam *text preprocessing* [4]:

1. *Case folding*

Mengubah semua huruf pada dokumen menjadi huruf kecil. Karakter selain huruf dihilangkan dan dianggap delimiter.

2. *Parsing / tokenizing*

Proses pemotongan string input berdasarkan tiap kata yang menyusunnya.

3. *Filtering*

mengambil kata - kata penting dari hasil token. Bisa menggunakan algoritma *stoplist / stopword* yang kurang penting) atau *wordlist* (menyimpan kata penting).

4. *Stemming*

mencari *root* kata dari tiap kata hasil *filtering*. Pada tahap ini dilakukan proses pengembalian berbagai bentukan kata ke dalam suatu representasi yang sama.

2.3 *Bayesian Filtering*

Bayesian Filtering merupakan sebuah metode yang awalnya digunakan sebagai pemfilteran sebuah spam dengan mengaplikasikan *Teorema Bayes*. *Bayesian filtering* memanfaatkan rule yang biasanya digunakan untuk mengkalsifikasikan sebuah objek atau biasa disebut dengan metode *Naïve Bayes Classifier*. Keuntungan penggunaan *Naive Bayes Filtering* ini adalah metode ini hanya membutuhkan jumlah data pelatihan (*training data*) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian [6]. Rumus *Bayesian Filtering* ditunjukan pada formula 1.

$$P(X_K|Y) = \frac{P(Y|X_K)}{\sum_i P(Y|X_I)} \quad \dots(1)$$

$P(X_k|y)$: Probabilitas Xk pada dokumen Y
$P(Y|X_K)$: Probabilitas dokumen Y mengandung Xk

2.4 *Multikolinearitas*

Multikolinieritas adalah suatu kondisi dimana terjadi korelasi antara variabel bebas atau antar variabel bebas tidak bersifat saling bebas[10].

2.4.1 Analisis Komponen Utama

Analisis komponen utama merupakan teknik statistik yang dapat digunakan untuk menjelaskanstruktur variansi-kovariansi dari sekumpulan variabel melalui beberapa variabel baru dimana variabel baru ini saling bebas, dan merupakan kombinasi linier dari variabel asal.Selanjutnya variabel baru ini dinamakan komponen utama (*principal component*).Secara umum tujuan dari analisis komponen utama adalah mereduksi dimensi data dan mencari variable yang paling berpengaruh[10].

2.4.2 Komponen utama yang Dibentuk Berdasarkan Matriks Kovariansi

Misal Σ merupakan matriks kovariansi dari vektor acak $X' = [X_1 X_2, \dots X_p]$dengan pasangan nilai eigen dan vektor eigen yang saling ortonormal adalah $(\lambda_1, e_1), (\lambda_2, e_2), \dots (\lambda_p, e_p)$. Dimana $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, maka komponen utama ke-i didefinisikan sebagai berikut[10]:

$$W_i = e'_i = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ij}X_j, \quad \dots(1)$$

W_1adalah komponen pertama yang memenuhi maksimum nilai $e_1' \sum e_1 = \lambda_1$. W_2adalah komponen kedua yang memenuhi sisa keragaman selain komponen pertama dengan memaksimumkan nilai $e_2' \sum e_2 = \lambda_2$ .W_p adalah komponen ke- p yang memenuhi sisa keragaman selain $W_1, W_2, \dots, W_{p-1}$dengan memaksimumkan nilai = $e_p' \sum e_p = \lambda_p$. Urutan $W_1, W_2, \dots, W_p$harus memenuhi persyaratan $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Sementara itu, proporsi total variansi yang dijelaskan komponen utama ke- *k* adalah :

$$\begin{pmatrix} Proporsi\ total\ variansi\ populasi \\ yang\ dijelaskan\ oleh\ komponen\ utama\ ke-k \end{pmatrix}$$

$$= \frac{\lambda_k}{tr(\sum)} = \frac{\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p} \dots(2)$$

Dengan $k = 1,2, \dots, p$

2.4.3 Kriteria Pemilihan Komponen Utama

Salah satu tujuan dari analisis komponen utama adalah mereduksi dimensi data asal yang semula terdapat *p* variabel bebas menjadi *k* komponen utama (dimana *k*<*p*). Kriteria pemilihan *k* yaitu :

1. Proporsi kumulatif keragaman data asal yang dijelaskan oleh *k* komponen utama minimal 80% , dan proporsi total variansi populasi bernilai cukup besar [11].
2. Dengan menggunakan scree plot yaitu plot antara *i* dengan *I* , pemilihan nilai *k* berdasarkan scree plot ditentukan dengan melihat letak terjadinya belokan dengan menghapus komponen utama yang menghasilkan beberapa nilai eigen kecil membentuk pola garis lurus [12].

2.5 *Entropy dan Information Gain*

Entropy adalah ukuran dari teori informasi yang dapat mengetahui karakteristik dari *impurity* dan *homogenity* dari kumpulan data. Setelah mendapat nilai *entropy* untuk suatu kumpulan data, maka kita dapat mengukur efektivitas suatu atribut dalam mengklasifikasikan data. Ukuran efektifitas ini disebut *information gain*. Secara matematis,untuk mendapatkan nilai *entropy* dan *infomation gain* dari suatu atribut A, dirumuskan sebagai berikut[13]:

$$Entropy(S) = \sum_{i=1}^{n} -p_i log_2 p_i \quad \dots (3)$$

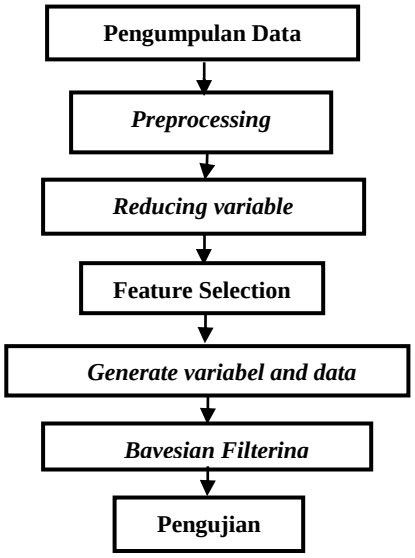
$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \dots (4)$$

Dimana:

S: himpunan kasus
n: jumlah partisi S
p_i: proporsi dari Si terhadap S
A: atribut
|Si|: jumlah kasus pada partisi ke-i
|S|: jumlah kasus dalam S
Entropy (Si): *entropy* untuk sampel-sampel yang memilki nilai i

3. METODOLOGI PENELITIAN

Penelitian akan dilaksanakan berdasarkan tahap penelitian yang ditunjukkan pada Gambar 3.1.



Gambar 3.1 Proses Tahap Penelitian

3.1 Pengumpulan Data

Dalam tahapan ini, terdapat dua step yang digunakan:

3.1.1 Studi literatur

Studi literatur dilaksanakan dengan melakukan studi kepustakaan untuk memahami teori yang berasal dari buku maupun artikel dan jurnal yangbersumber dari internet.Studi kepustakaan dilakukan untuk mempelajari hal-halyang berhubungan dengan penelitian secara umum, literatur tentang *Web Mining* dan *Multivariate.*

3.1.2 Crawling data

Data berupa link situs porno dan non porno dalam bahasa inggris. Dari situs yang sudah dikumpulkan. Kemudian dilakukan crawling terhadap situs-situs yang terdapat pada list. Hasil crawling tersebut disimpan pada database berupa text html.

3.2 *Preprocessing*

Pada preprocessing ini dilakukan dalam beberapa step. Pertama menghapus semua tag html. Kemudian mencocokkan setiap kata pada text pada list yang terdapat pada database. Jika tidak cocok maka akan dihapus. Kemudian membuang katat yag termasuk dalam *stopword*, Jika setiap kata sudah diproses kemudian masuk ke proses *stemming*. Dari proses stemming inilah setiap kata akan menjadi kata dasar. Sehingga tidak ada redundansi kata.

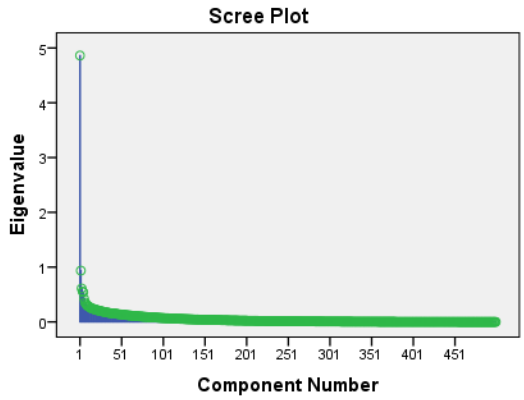
3.3 *Reducing variable*

Setelah data sudah siap. Maka setiap kata akan disimpan menjadi variable penentu klasifikasi. Kemudian memilih *keyword* yang memiliki frekuensi yang cukup besar dengan pembobotan information Gain. Apabila *keyword* tidak melewati threshold maka *keyword* tidak bisa dijadikan variable. Untuk perhitnugan nilai gain menggunakan formula 3 dan 4.

3.4 Feature selection

Memilih variable yang paling berpengaruh dengan memperhitungkan tingkat korelasi antar variable dengan metode *Principal component analysis* (PCA). Penelitian ini menggunakan SPSS 2.1 untuk menghitung *Principal component analysis*. Berikut langkah-langkah pemilihan varaibel dengan *Principal component analysis:*

1. Menentukan jumlah *component* dengan melihat nilai eigen. Apabila nilai eigen tidak menalami perubahan, maka titik terakhir itu jumlah *component* yang digunakan untuk proses selanjutnya. *Component* merupakan variabel baru yang dihasilkan PCA. Variabel tersebut berbentuk formula untuk mencari komposisi terbaik dari variabel lama. Besarnya pengaruh *component* ditunjukan oleh nilai eigen semakin besar nilai eigen maka semakin baik *component* tersebut. Grafik eigen value terhadap component ditunjukan pada gambar 3.2



Gambar 3.2 Grafik PCA number

dari gambar 3.2 dapat dilihat bahwa pada component 101 tidak mengalami perubahan nilai maka component yang akan digunakan adalah component 1 sampai component 101.

2. Mencari variabel dengan nilai kovarians lebih besar dari 0.5 pada component matriks yang dihasilkan. Variabel yang memiliki nilai kovarians kurang dari 0.5 diabaikan. Karena nilai kovarians merupakan nilai yang menunjukan keterikatan variabel satu dan lainnya. Semakin tinggi nilai covarians maka semakin bebas dan tidak tergantung variabel tersebut dengan lainnya. nilai covarians pada component matriks ditunjukan pada tabel 3.1.

Tabel 3.1. Contoh nilai covarians variabel PCA

Keyword (variabel)	Component		
	1	2	3
Amateur	**-.595**	.051	.454
Articl	**.705**	.079	-.100
Tube	**-.551**	.072	.392
Babe	**-.535**	.014	.492
Fuck	-.482	.088	.464
Brunett	**-.534**	.077	**.517**

Tabel 3.1 menunjukan nilai pengaruh variabel lama terhadap variabel baru.

Contoh:
Rumus Variabel 1 adalah
-0.595.amateur + 0.75.article - 0.551.babe - 0.482.Fuck - 0.534. Brunett.

Pada kasus pemilihan variabel ini nilai bobot diabaikan dan hanya mengambil nilai eigen yang lebih dari 0.5 maka variabel yang dipakai terdapat pada component1 dan component 3. Dengan total 4 variabel yaitu *Amateur, Article, Tube, Babe, Brunett.*

3.5 Generate variabel and data training

Menentukan data yang akan dijadikan data training dan data testing. Kemudian memberi label pada setiap data masuk dalam salah satu kategori. Yaitu porno dan non porno. Kemudian menyimpan hasil pelabelan dan hasil perhitungan variable dari proses sebelumnya.

3.6 Bayesian Filtering

Mengklasifikasikan data testing menggunakan variable yang sudah direduksi dan variable yang telah dipilih pada proses PCA. Pengkalsifikasian ini menggunakan bayesian filtering. Untuk rumus perhitungan bayesian filtering ditunjukan pada formula 5 dan 6.

$$P(Porn\ |web\ i) = \frac{\sum P(keyword\ x \mid Porn)}{\sum P(Keyword\ x)} \dots (5)$$

$$P(Safe\ |web\ i) = \frac{\sum P(keyword\ x \mid Safe)}{\sum P(Keyword\ x)} \dots (6)$$

Keterangan :

$P(Porn\ \|web\ i)$	Probabilitas porno dari web *i*
$P(keyword\ x \mid Porn)$	Probabilitas munculnya *keyword* x pada dokumen porno dari data training
$P(Keyword\ x)$	Probabilitas munculnya *Keyword* x pada semua dokumen dari data training
$P(Safe\ \|web\ i)$	Probabilitas safe dari web *i*
$P(keyword\ x \mid Safe)$	Probabilitas munculnya *keyword* x pada dokumen *safe* dari data training

Nilai probabilitas Status 1 dan 0 dibandingkan, Jika nilai P(status=1) lebih besar maka dokumen tersebut dideteksi sebagai *porn*, apabila nilai P(Status=0) lebih besar maka dideteksi sebagai *safe.*

3.7 Pengujian

Pengujian ini bertujuan untuk mengetahui komposisi data training yang tepat, menghitung tingkat akurasi deteksi *bayesian filtering* dan mengetahui efektifitas dari principal component analysis. Testing dilakukan dengan porsi data dan variable yang berbeda. Variable yang digunakan untuk pengujian adalah 1000 variable dengan nilai gain tertinggi dan 1000 variable dengan gain tertinggi yang diseleksi lagi dengan PCA. Dengan jumlah data yang digunakan untuk pengujian sebagai data training dan data testing adalah 1000 website. 500 data sebagai training dan 500 sebagai data testing.

Dari 500 data training dilakukan dengan 3 porsi komposisi. data training pertama 100 web porno dan 400 web safe. Kedua 250 web porno dan 250 web safe. Ketiga 400 web porno dan 100 web safe. Perbedaan porsi ini bertujuan untuk membandingkan porsi yang paling akurat. Untuk menghitung akurasi menggunakan rumus(4) :

$$Akurasi = \frac{Jumlah\ data\ benar}{Jumlah\ total\ data} \times 100\% \dots (4)$$

untuk mengetahui tingkat optimalisasi hasil Principal component analysis, penelitian ini memembandingkan deteksi menggunkan variabel PCA dan deteksi menggunakan 100 variabel yang memiliki nilai gain tertinggi.

4 PEMBAHASAN DAN ANALISIS HASIL

4.1 Perolehan data

Dari 1136 domain yang dikumpulkan diperoleh 1046 website yang diperoleh pada proses crawling. Hasil crawli ng tersebut terdiri dari 454 website porno dan 592 website non porno(*safe*).

4.2 *Preprocessing dan Reducing variabel*

Dari 1136 website yang telah di crawling, diambil 500 website digunakan sebagai data training, kemudian dilakukan langkah preprocessing. Maka diperoleh *keyword* (variabel) sebagai berikut :

Tabel 4.1 Tabel total *keyword*(variabel)

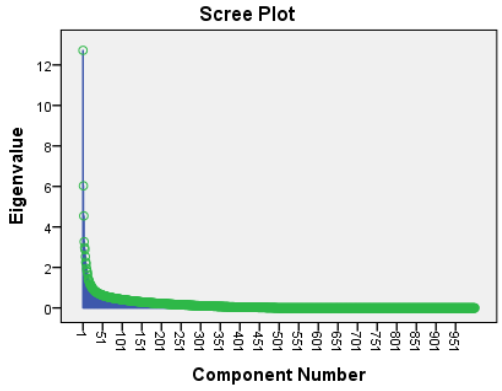
No. training	Data Training		Jumlah *Keyword*
	PORN	SAFE	
1	100	400	4007
2	250	250	15564
3	400	100	15564

Setelah mendapatkan *keyword* untuk masing-masing data training, kemudian dilakukan *reducing variabel* dengan menghitung nillai *Gain* pada tiap *keyword*. Berdasarkan nilai gain yang dimiliki, kemudian diambil 1000 *keyword* dengan nilai tertinggi pada tiap data training.

4.3 *Feature Selection*

1000 *keyword* yang telah diperoleh kemudian dihitung dengan metode *Principal component Analysis*. Hasil perhitungan PCA sebagai berikut:

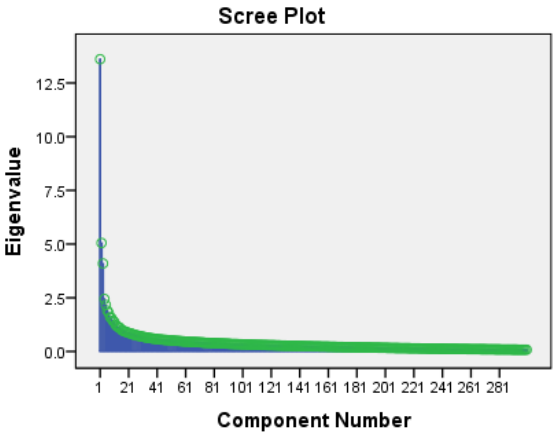
a. *Principal component analysis* pada data training 100 web porno dan 400 web safe



Gambar 4.1 Grafik PCA number 100 web porno: 400 web *safe*

Gambar 4.1 menunjukan nilai eigen dari *component* yang terbentuk oleh 1000 variabel. Dari tabel tersebut dapat dilihat bahwa tidak ada perubahan nilai eigen yang signifikan pada *component* 244. Dari 244 *component* yang terbentuk diperoleh 4 *component* yang memenuhi syarat yaitu terdapat *keyword* dengan nilai *covarians* sebesar 0.5. 4 *component* tersebut adalah *component*1 21 *keyword* ,*component2* 1 *keyword*, *compoent3* 5 *keyword* dan *component7* 1 *keyword*, total *keyword* yang memenuhi syarat adalah 28 *keyword*. Total covarians yang dihasilkan PCA pada data training 100 web porno dan 400 web safe adalah 87.96%. Artinya 28 *keyword* tersebut dapat mewakili 1000 variabel dengan nilai keaslian 87.96%.

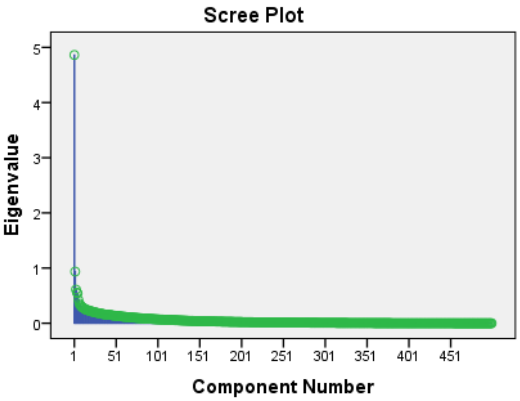
b. *Principal component analysis* pada data training 250 web porno dan 250 web safe



Gambar 4.2 Grafik PCA number 250 web porno: 250 web *safe*

Dari gambar 4.2 dapat dilihat bahwa tidak ada perubahan nilai eigen yang signifikan pada *component* 236. Dari 236 *component* yang terbentuk diperoleh 3 *component* yang memenuhi syarat. 3 *component* tersebut adalah *component1* 25 *keyword* ,*component2* 22 *keyword*, *compoent3* 4 *keyword*, total *keyword* yang memenuhi syarat adalah 51 *keyword*. Total covarians yang dihasilkan PCA pada data training 250 web porno dan 250 web safe adalah 89.53%. Artinya 53 *keyword* tersebut dapat mewakili 1000 variabel dengan nilai keaslian 89.53%.

c. *Principal component analysis* pada data training 400 web porno dan 100 web safe



Gambar 4.3 Grafik PCA number 400 web porno: 100 web *safe*

Dari gambar 4.3 nilai eigen yang signifikan pada *component* 182. Dari 182 *component* yang terbentuk diperoleh 3 *component* yang memenuhi syarat. 3 *component* tersebut adalah *component*1 40 *keyword* ,*component2* 1 *keyword*, *compoent7* 7 *keyword*, total *keyword* yang memenuhi syarat adalah 48 *keyword*. Total covarians yang dihasilkan PCA pada data training

400 web porno dan 100 web safe adalah 91.99 %. Artinya 48 *keyword* tersebut dapat mewakili 1000 variabel dengan nilai keaslian 91.99 %.

Setelah variabel tepilih dengan principal component analysis kemudian variabel dismipan sebagai keyword utama sebagai pendeteksi (*generate variabel)* dan menyimpan data sebagai data latih acuan deteksi (*data training*).

4.4 Analisis Hasil

Hasil deteksi web dengan *bayesian filtering* dapat dilihat pada tabel 4.2 :

Tabel 4.2 Hasil deteksi

No	Jumlah Data Training		Jumlah variabel Hasil PCA	Akurasi PCA	Akurasi 100 variabel
	Porno	Safe			
1	100	400	28	29.94	31.74
2	250	250	51	89.22	90.62
3	400	100	44	10.98	10.98

a. Data training dengan perbandingan 100:400 antara web porno dan bukan dengan 28 variabel hasil esktraksi PCA memiliki tingkat akurasi 29.94%. Sedangkan 100 variabel tanpa PCA sebesar 31.74%. Nilai akurasi selisih 1.8% .
b. Data training dengan perbandingan 250:250 antara web porno dan bukan dengan 51 variabel hasil esktraksi PCA memiliki tingkat akurasi 89.22%. sedangkan 100 variabel tanpa PCA sebesar 90.62%. Nilai akurasi selisih 1.4% .
c. Data training dengan perbandingan 400:100 antara web porno dan bukan dengan 44 variabel hasil esktraksi PCA memiliki tingkat akurasi 10.98%. sedangkan 100 variabel tanpa PCA sebesar 10.98%. Nilai akurasi sama.

Dari hasil tersebut nilai akurasi terbesar dimiliki oleh data training dengan 250 web porno dan 250 bukan porno. Nilai akurasi sebesar 89.22 dengan 51 variabel pca dan 90.62 dengan 100 variabel tanpa PCA. Selisih akurasi sebesar 1.4%.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian dapat disimpulkan bahwa :

1. Metode Bayesian filtering dapat mendeteksi web berkonten porno dengan cukup akurat dengan nilai akurasi tertinggi sebesar 90.62%.
2. *Principal component analysis* dapat memperkecil jumlah variabel dan mencari variabel paling penting tanpa harus kehilangan informasi. Hasil ini dibuktikan dengan nilai akurasi yang hanya memiliki nilai selisih terbesar 1.8%.
3. Perbandingan antara web porno dan tidak porno pada data training mempengaruhi tingkat akurasi, semakin seimbang proporsi data semakin akurat deteksi yang dihasilkan.

5.2 Saran

Saran terhadap penelitian ini yaitu:

1. Meningkatkan jumlah data training.
2. Menambahkan metode pendeteksian konten untuk file bertipe *video, image.*
3. Menambahkan metode untuk mendeteksi kemiripan karakter dari frase text dengan *text similarity.*

6. DAFTAR PUSTAKA

[1] TopTenReviews, *Internet Pornography Statistics*, (http://internet-filter-review.toptenreviews.com/ internet-pornography-statistics.html). Diakses 28 Januari 2014 .

[2] Liu Bing, 2005. *Web Content Mining*, Department of Computer Science University of Illinois at Chicago.

[3] Liu Bing, 2011. *Web Data Mining*, Department of Computer Science University of Illinois at Chicago.

[4] Feldman Ronen , 2007. *The Text Mining Handbook Advanced Approaches In Analyzing Unstructured Data*. Cambridge University Press, New York.

[5] Dhenakaran S.S , Thirugnana Sambanthan .K , 2011. *Web Crawler*.Department of Computer Science and Engineering.

[6] Rachli Muhammad, 2007. *Email Filtering Menggunakan Naïve Bayesian.* Teknik Elektro. Institut Teknologi Bandung.

[7] Jolliffe, 1972.*Discarding Variables in a Principal Component Analysis. I: Artificial Data.* Blackwell Publishing, Royal Statistical SocietyStable

[8] Olston Christopher, 2010. Web crawling, The essence of knowledge.

[9] Azhagusundari, 2013. *Feature Selection based on Information Gain.* International Journal of Innovative Technology and Exploring Engineering.

[10] Haris Bhakti Prasetyo, 2008.*Analisis Regresi Komponen Utama untuk Mengatasi Masalah Multikolinieritas dalam Analisis Regresi Linier Berganda*.FMIPA, Universitas Negeri Jakarta. Johnson, R.A. & Wichern, D.W. 2002.*Applied Multivariate Statistical Analysis, 5 Thedition* .Pearson Education International.

[11] *Rencher Alvin .C,*1998. *Christensen William .F ,Methods of Multivariate Analysis.* . Blackwell Publishing, Royal Statistical SocietyStable.