

PENERAPAN *TEXT MINING* PADA SISTEM KLASIFIKASI EMAIL SPAM MENGGUNAKAN NAIVE BAYES

¹Ervita Kusuma Putri (09018228), ²Tedy Setiadi (0407016801)

^{1,2}Program Studi Teknik Informatika
Universitas Ahmad Dahlan

Prof. Dr. Soepomo, S.H., Janturan, Umbulharjo, Yogyakarta 55164

²Email: tedy.setiadi@tif.uad.ac.id

ABSTRAK

Email atau Elektronik mail merupakan salah satu fasilitas internet yang murah dan mudah digunakan untuk melakukan transfer informasi atau penyebaran informasi berupa file (mail attachment) antar pengguna internet. Tetapi tidak semua pengguna memanfaatkan email dengan baik dan benar. pengguna yang kurang baik memanfaatkan email untuk menyebarkan informasi yang tidak baik seperti virus dan iklan suatu perusahaan atau mempromosikan produk bisnis tertentu. Email yang seperti itulah yang lebih dikenal dengan email spam. Email spam dikirim ke banyak orang tanpa melakukan izin terlebih dahulu ke pemilik email yang dituju. Berdasarkan permasalahan tersebut, maka dibuat suatu penelitian untuk mengembangkan suatu aplikasi text mining yang mampu mengklasifikasi email.

Text mining merupakan proses menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen. Proses dalam text mining meliputi proses tokenisasi, stemming dan filtering. Metode pengumpulan data dengan metode kepustakaan. Tahapan pengembangan aplikasi meliputi perancangan proses, perancangan tabel, implementasi dan pengujian sistem. pengujian sistem dengan black box test dan alpha test.

Dari penelitian yang dilakukan menghasilkan sebuah perangkat lunak penerapan text mining pada sistem klasifikasi email spam menggunakan metode naive bayes. Pada klasifikasi email dihitung nilai probabilitas berdasarkan kemunculan kata yang terdapat dalam data email. pengujian keakurasian sistem ditampilkan berupa grafik nilai keakurasian, false positif dan false negatif. Hasil uji coba menunjukkan bahwa aplikasi ini layak dan dapat digunakan dan memiliki nilai keakurasian sistem sebesar 89,6 %.

Kata Kunci : *Text Mining, Klasifikasi, Email spam, Naive Bayes.*

1. PENDAHULUAN

Email atau Elektronik mail merupakan salah satu fasilitas internet yang digunakan untuk melakukan komunikasi atau berdiskusi (maillist), transfer informasi atau penyebaran informasi berupa file (mail attachment) antar pengguna

internet dengan cara mengirim dan menerima pesan antar pengguna. *Email* yang sifatnya yang mudah dan murah juga membuat semakin banyak pengguna nya. Tetapi tidak semua pengguna memanfaatkan *email* dengan baik dan benar. Pengguna yang baik, hanya memanfaatkan *email* untuk melakukan komunikasi dan penyebaran informasi-informasi yang baik. Sedangkan pengguna yang kurang baik memanfaatkan *email* untuk menyebarkan informasi – informasi yang tidak baik seperti virus. Tidak sedikit pula pengguna yang menggunakan *email* untuk media iklan suatu perusahaan atau mempromosikan produk bisnis tertentu. *Email* yang seperti itulah yang lebih dikenal dengan *email spam*.

Email spam merupakan *email* yang bertujuan untuk mempromosikan iklan produk layanan-layanan suatu produk. *Email spam* dikirim ke banyak orang tanpa melakukan izin terlebih dahulu ke pemilik *email* yang dituju. Beberapa *emailspam* mengandung virus, pengirim *spam* akan mengirimkan *email* berisi virus yang dapat merusak komputer pemilik *email*. *Spam* tidak hanya mengganggu, tetapi dapat berbahaya dan mengakibatkan pencurian identitas dan kehancuran finansial[1].

Text mining merupakan proses menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen [2]. Tahapan pertama yaitu teks kemudian tahap kedua pengolahan teks (tokenisasi), tahap ketiga perubahan teks (*stemming*), tahap ke empat pemilahan teks (*filtering*), tahap kelima Data Mining (*Pattern Discovery*) dan tahap terakhir, Evaluasi adalah penafsiran pola yang ditemukan.

Tahapan penemuan pola adalah tahap terpenting dari keseluruhan proses *text mining*. Merupakan penemuan pola atau pengetahuan dari keseluruhan teks. Proses penemuan pola pada data *mining* dapat dilakukan dengan metode klasifikasi. Model yang digunakan untuk klasifikasi yaitu dengan formula matematis *naive bayesian*.

2. KAJIAN PUSTAKA

Penelitian ini mengacu pada penelitian terdahulu tahun 2007 yang berjudul “*Email Filtering Menggunakan Naive Bayes*”. Penelitian ini menghasilkan sebuah database filter yang digunakan untuk mengidentifikasi *email* sebagai *spam* atau *legitimate mail*[3].

Penelitian ini juga mengacu pada penelitian pada tahun 2010 yang berjudul

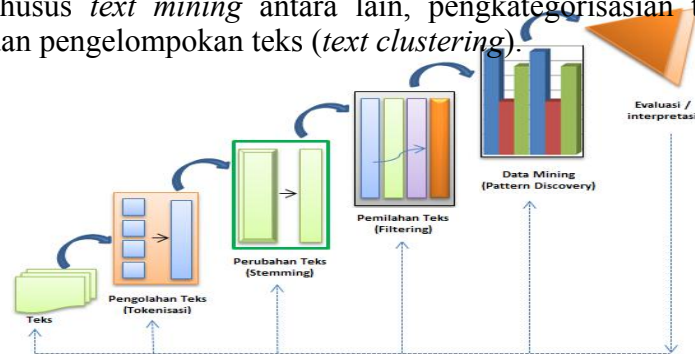
“*Klasifikasi Email Spam dengan Metode Bayes Classifier Menggunakan Java Programming*”. Penelitian ini menghasilkan sebuah sistem klasifikasi *email spam* yang dapat berhasil membuktikan metode *naive bayes classifier* dapat mengidentifikasi *spam* yang dilakukan dengan dua cara yaitu sistem klasifikasi dapat beroperasi pada *mail client(offline)* dan pada *mail server(online)*[4].

3. TEORI PENDUKUNG

3.1 Text Mining

Text mining merupakan proses menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen. Tujuan dari *text mining* adalah mengekstrak informasi yang berguna dari sumber data. Jadi, sumber data yang digunakan pada *text mining* adalah sekumpulan dokumen yang memiliki format

yang tidak terstruktur melalui identifikasi dan eksplorasi pola yang menarik. Adapun tugas khusus *text mining* antara lain, pengkategorisasian teks (*text categorization*) dan pengelompokan teks (*text clustering*).



Gambar 1. Tahapan Proses *Text Mining*

Tahapan *text mining* terdiri dari teks, Pengolahan teks (*tokenisasi*) adalah memecah kalimat menjadi kata per kata, perubahan huruf besar ke huruf kecil (kapitalisasi) dan menghilangkan tanda baca, Perubahan teks (*stemming*) adalah perubahan kata berimbuhan menjadi kata dasar, pemilahan teks (*filtering*) adalah melakukan perhitungan dan pengelompokan kata per kata, Data Mining (*Pattern Discovery*) adalah proses pencarian pengetahuan atau pola yang menarik/bernilai, Evaluasi adalah penafsiran pola yang ditemukan[5].

3.2 Klasifikasi

Klasifikasi adalah proses pencarian sekumpulan model atau fungsi yang menggambarkan dan membedakan kelas data dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas dari suatu objek yang belum diketahui kelasnya. Model itu sendiri bisa berupa aturan “jika-maka”, berbentuk pohon keputusan (*decision tree*), formula matematis seperti *naive bayesian* dan *support vector machine*. Proses klasifikasi biasanya dibagi menjadi dua fase: *learning* dan *test*. Pada fase *learning*, sebagian data yang telah diketahui kelas datanya diumpamakan untuk membentuk model prediksi. Karena menggunakan data yang telah diberikan label terlebih dahulu sebagai contoh data yang benar maka klasifikasi sering disebut juga sebagai metoda diawasi (*supervised method*). Kemudian pada fase *testing*, model prediksi yang sudah terbentuk diuji dengan sebagian data lainnya untuk mengetahui akurasi dari model tersebut. Bila akurasinya mencukupi model ini dapat dipakai untuk prediksi kelas data yang belum diketahui[5].

3.3 Teorema Bayesian

Teorema Bayes adalah teorema yang digunakan dalam statistika untuk menghitung peluang untuk suatu hipotesis, *Bayes Optimal Classifier* menghitung peluang dari suatu kelas dari masing-masing kelompok atribut yang ada, dan menentukan kelas mana yang paling optimal. Persamaan dalam *teorema bayes* adalah sebagai berikut[6] :

$$(I) \quad \frac{P(A) \cdot P(B|A)}{P(B)} \text{ atau } (II) \quad \frac{P(A) \cdot P(B|A)}{P(A) \cdot P(B|A) + P(A) \cdot P(B|A)} \quad [1]$$

Teorema Bayes memanipulasi persamaan diatas ke dalam sebuah pernyataan probabilitas dalam hal kemungkinan (*likelihood*).

$$(1) \frac{P(C)}{P(F)} \quad (1) \dots \dots \dots [2]$$

3.4 Naive Bayesian Filtering

Naive Bayesian Filtering memanfaatkan metode klasifikasi bayesian dengan dua asumsi dasar yaitu nilai atribut dari kelas yang didefinisikan independen (bebas) dari nilai atribut yang lain dan prior probabilitas suatu *email* sebagai spam tidak diketahui. Asumsi pertama dikenal dengan sebutan naive Bayesian[4]. Algoritma Naïve Bayes adalah algoritma yang digunakan untuk mengklasifikasikan suatu *email* sebagai *email spam* atau *non spam*.

Model probabilitas untuk klasifikasi adalah model p bersyarat ($C | F_1, F_2, \dots, F_n$) atas kelas variabel dependen C dengan sejumlah kecil hasil atau kelas, tergantung pada beberapa variabel fitur F_1 melalui F_n . Masalahnya adalah bahwa jika jumlah fitur n besar atau ketika fitur dapat mengambil sejumlah besar nilai-nilai, maka mendasarkan model seperti pada tabel probabilitas tidak layak. Teorema bayes berkaitan probabilitas kondisional dan marginal peristiwa stokastik C, dan F [6]:

$$P(C) = \frac{P(C|F) \cdot P(F)}{P(F|C)} \quad [3]$$

dimana : P (C) adalah probabilitas sebelumnya hipotesis C , P (F) adalah probabilitas sebelumnya dari data training F , P (C|F) adalah probabilitas yang diberikan F dan , P (F|C) adalah probabilitas F diberikan C. Menggunakan teorema Bayes untuk beberapa variabel fitur F_n , kita dapat menulis ulang ini sebagai :

$$P(C) = \frac{P(C|F) \cdot P(F)}{P(F|C)} \quad [4]$$

Karena penyebut tidak tergantung pada C diberikan dan nilai-nilai dari fitur, sehingga penyebut secara efektif konstan . Pembilang setara dengan model probabilitas gabungan (3) yang dapat ditulis ulang menggunakan aplikasi berulang dari definisi probabilitas bersyarat sebagai :

$$P(C) = \frac{P(C|F) \cdot P(F)}{P(F|C)} \quad [5]$$

Klasifikasi yang sesuai untuk model ini didefinisikan sebagai berikut :

$$P(C) = \frac{P(C|F) \cdot P(F)}{P(F|C)} \quad [6]$$

Dari persamaan diatas akan di substitusikan dengan kasus untuk klasifikasi *email spam*. Klasifikasi *naive bayes* untuk dokumen *email*. Mengklasifikasi dokumen dengan konten, misalnya menjadi *email spam* dan *email nonspam*. Dokumen yang diambil dari sejumlah kelas dari dokumen yang dapat dimodelkan sebagai set kata-kata (independen) probabilitas bahwa ke-i kata dari dokumen yang diberikan terjadi dalam sebuah dokumen. Maka probabilitas diberikan *Email E* berisi semua kata-

kata () , mengingat kelas C, adalah

$$P(C) = \frac{P(C|F) \cdot P(F)}{P(F|C)} \quad [7]$$

Menggunakan hasil persamaan (7) dan dengan asumsi bahwa hanya ada dua kelas, Spam dan NSpam (misalnya *spam* dan *non spam*) dapat ditulis dengan :

$$\frac{P(w_i | Spam)}{P(w_i | NonSpam)} \prod_{i=1}^n \frac{P(w_i | Spam)}{P(w_i | NonSpam)} \dots [8]$$

Dengan demikian, rasio probabilitas $P(Spam | Email) / P(NonSpam | Email)$ dapat dinyatakan dalam serangkaian rasio kemungkinan. Sebenarnya probabilitas $P(Spam | Email)$ dapat dengan mudah dihitung dari $\log(P(Spam | Email) / P(NonSpam | Email))$ berdasarkan pengamatan bahwa $P(Spam | Email) + P(NonSpam | Email) = 1$. Dengan mengambil logaritma dari semua rasio ini, dapat didefinisikan :

$$\log \left(\frac{P(w_i | Spam)}{P(w_i | NonSpam)} \right) = \sum_{i=1}^n \log \left(\frac{P(w_i | Spam)}{P(w_i | NonSpam)} \right) \dots [9]$$

Sifat Suatu *email* dapat diklasifikasikan sebagai berikut :

Dimana , Jika hasil klasifikasi menghasilkan nilai kurang dari dan sama dengan 0 maka sifat dari *email* tersebut *non spam* [6].

Keterangan :

: kata

email spam

: *email non spam*

() : Peluang kemunculan kata dalam *email spam*.

() : Peluang kemunculan kata dalam *email non spam*.

() : Peluang kata *spam* dengan *email keseluruhan*.

() : Peluang kata *non spam* dengan *email keseluruhan*.

C* : Klasifikasi *naive bayes*

4. METODE PENELITIAN

4.1 Metode Pengumpulan Data

Metode pengumpulan data yang dilakukan dalam penelitian ini adalah Metode Kepustakaan. Metode Kepustakaan ini merupakan metode yang dilakukan dengan cara mengumpulkan, mempelajari dan memahami buku-buku referensi serta laporan tugas akhir termasuk pula pustaka-pustaka digital dari hasil browsing di internet yang relevan dengan topik penelitian ini seperti metode *naive bayes* untuk pengklasifikasian *email spam*, *text mining* dan cara kerja *spam*.

4.2 Metode Pengumpulan Sistem

Tahap pengumpulan sistem ini dilakukan dengan menganalisis terhadap metode yang akan digunakan dalam sistem klasifikasi *email spam* yaitu metode *naive bayes*, bagaimana cara kerja *text mining* dengan metode *naive bayes* dan apakah metode ini memiliki tingkat keakuratan yang tinggi dalam melakukan klasifikasi *email spam* dan *email* yang *non spam*.

4.3 Pengembangan Sistem

4.3.1 Pengolahan Teks

Pengolahan teks merupakan tahapan proses awal terhadap teks, untuk mempersiapkan teks menjadi data yang akan diolah lebih lanjut. Pengolahan teks adalah proses memecah teks menjadi kalimat dan kata atau token. *Tokenisasi* juga digunakan untuk mengekstraksi fitur-fitur token. Beberapa fungsi sederhana, seperti penyamaan tipe kapitalisasi, deteksi keberadaan digit, eliminasi tanda baca, karakter spesial, dan sebagainya akan membantu dalam menggambarkan suatu properti yang harus dipenuhi token dalam barisan karakter sebagai calon *token*.

4.3.2 Perubahan Teks

Perubahan teks merupakan tahapan yang dipergunakan untuk mengubah kata-kata ke dalam bentuk dasar, sekaligus untuk mengurangi jumlah kata-kata tersebut. Pendekatan yang dapat dilakukan dengan *stemming* dan *stopword removal*.

4.3.2.1 *Stemming*

Teknik untuk meningkatkan performa calon *token*, yaitu dengan cara menemukan variasi *token* dari *token* pencarian yang dimasukkan. Misalnya, pengguna memasukkan *token* '*stemming*' sebagian dari *query*, seharusnya akan mendapatkan variasi token lainnya. Seperti '*stemmed*' dan '*stem*'. Kerugian dari proses *stemming* adalah informasi mengenai *token* awal akan hilang atau harus dipergunakan penyimpanan tambahan untuk menyimpan bentuk awalnya (*unstemmed*).

4.3.2.2 *Stopword Removal*

Stopword removal adalah proses untuk menghilangkan kata-kata yang kurang relevan atau dianggap tidak akan memberikan kontribusi yang besar jika muncul dalam suatu dokumen, dan akan memperlama proses. Kata-kata tersebut biasanya berupa kata sandang dan sambung (misalnya a, an, the, on pada Bahasa Inggris) dan dalam proses ini sekumpulan kata tersebut disebut sebagai sekumpulan *stopword* (*stoplist*).

4.3.3 Pemilahan Teks

Pemilahan Teks merupakan tahapan seleksi fitur ini digunakan setelah melakukan proses pemilahan teks, tetapi tidak semua kata yang telah dilakukan proses pemilahan teks menggambarkan isi dari dokumen. Tahap seleksi fitur ini bertujuan mengurangi dimensi dari suatu kumpulan teks. Dengan kata lain, menghapus kata-kata yang dianggap tidak penting atau tidak menggambarkan isi dokumen berdasarkan frekuensi kemunculan kata di dalam teks yang bersangkutan.

4.3.4 Data Mining

Pada tahapan data *mining* akan dilakukan tahapan penemuan pola atau pengetahuan dari keseluruhan teks. Tahapan ini adalah tahap terpenting dari keseluruhan proses *text mining*. Proses penemuan pola pada data

mining dapat dilakukan dengan metode klasifikasi. Klasifikasi adalah proses pencarian sekumpulan model atau fungsi yang menggambarkan dan membedakan kelas data dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas dari suatu objek yang belum diketahui kelasnya. Model itu sendiri bisa berupa aturan “jika-maka”, berbentuk pohon keputusan (*decision tree*), formula matematis seperti *naive bayesian* dan *support vector machine*.

4.3.5 Evaluasi Pola

Pada tahapan terakhir ini, penelitian ditujukan kepada *end user* atau pengguna *email*. Dalam penggunaannya diharapkan pengguna *email* dapat mengetahui ciri-ciri *email spam* dan *email* yang *non spam* dari kata-kata yang ada pada *email* dan dapat memfilter *email* agar tidak membuang waktu dalam penghapusan *email* yang tidak diinginkan atau *email spam*.

5. HASIL DAN PEMBAHASAN

5.1 Pengolahan Teks

Proses pengolahan teks dalam penelitian ini adalah dengan melakukan proses *Tokenisasi*, yaitu proses memecah teks menjadi kata atau *token* atau proses yang digunakan untuk mengekstraksi fitur-fitur *token*. Beberapa fungsi sederhana, seperti penyamaan tipe kapitalisasi, deteksi keberadaan digit, eliminasi tanda baca, karakter spesial, dan sebagainya akan membantu dalam menggambarkan suatu properti yang harus dipenuhi *token* dalam barisan karakter sebagai calon *token*.

5.2 Perubahan Teks

Perubahan kata dilakukan dengan cara mengambil *token* hasil proses pengolahan teks kemudian dilakukan proses perubahan kata dengan merubah kata menjadi kata dasar.

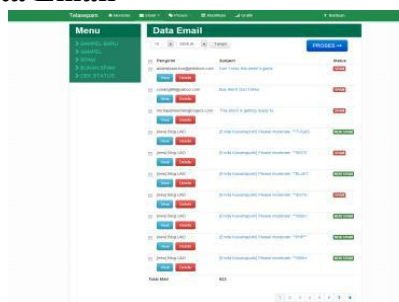
5.3 Pemilahan Teks

Pemilahan kata dilakukan dengan mengambil hasil dari proses perubahan kata yang sudah dalam bentuk kata dasar kemudian dilakukan proses pemilahan kata dengan menggabungkan kata yang sama kemudian dilakukan perhitungan frekuensi tiap katanya.

5.4 Data Mining

Implementasi aplikasi klasifikasi *email spam* adalah sebagai berikut:

5.4.1 Tampilan FormData Email



Gambar 2. FormData Email

Form data *email* menampilkan data *email* yang sudah tersimpan dalam database data *email* yang kemudian akan diuji. Dalam form data *email* ini hanya 3 field yang ditampilkan yaitu data pengirim, subject *email* dan status *email* apakah *email* tersebut *spam* (s) atau *non spam* (n). Tetapi dalam database dataemail terdapat banyak field yaitu id *email*, pengirim, penerima, subject, tanggal, isi dan status email. Pada menu data *email* dapat memilih *email* mana saja yang akan di uji pada proses *tokenisasi*, setelah memilih *email* maka langkah selanjutnya mengklik tombol proses yang kemudian akan dilakukan proses *tokenisasi*.

5.4.2 Tampilan Form Sampel Baru

Gambar 3. *Form* Sampel baru

Tampilan sampel baru yang digunakan untuk menginputkan data *email* yang akan diuji tanpa harus mengimport database. Setelah selesai mengisi data yang sesuai dengan format *form*, maka klik tombol simpan. Data yang sudah disimpan akan masuk ke dalam database data *email*.

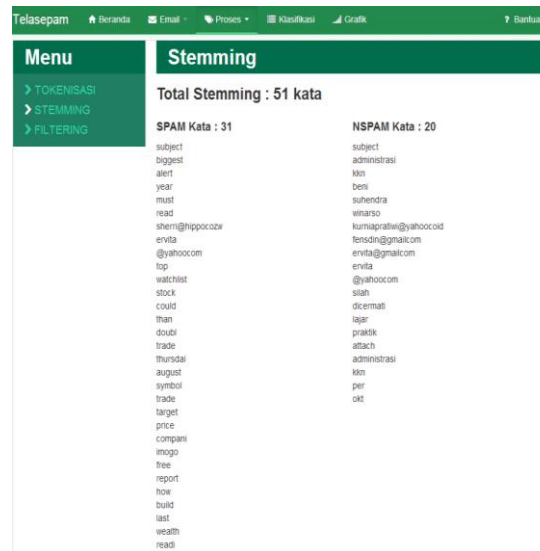
5.4.3 Tampilan Form Pengolahan Teks

Telaseparam		Home	Email	Program	Knowledgebase	Contact Us
Menu		Tokenisasi				
<ul style="list-style-type: none"> > TOKENISASI > STEMMING > FILTERING 		<p>Total Kata : 53</p> <hr/> <p>SPAM Kata : 33</p> <pre> subject logged alert year must read shern@hippopotuz ervita @yahoo.com top watchtest stock could than double trade Thursday August symbol traded m ic target price company emopo free report how found lasting wealth ready </pre>				
		<p>NPSM Kata : 20</p> <pre> subject administrasi kin berni suruhendra winanso kurniasaprathe@yahoo.co.id ferrado@gmail.com ervita@gmail.com ervita @yahoo.com shabanun de-emails dipenilaian diaperiksaan attachments admonistrasi kam per out </pre>				

Gambar 4. *Form* Pengolahan Teks

Tampilan *tokenisasi* yaitu proses pemecahan kata. Dalam menu ini menampilkan total kata yang mengikuti proses tokenisasi. Kemudian, kata-kata yang sudah dipecah dibagi kembali menjadi kata yang termasuk dalam kata *email spam* (kata *spam*) dan kata yang termasuk dalam kata *email non spam* (kata *non spam*).

5.4.4 Tampilan Form Perubahan Teks

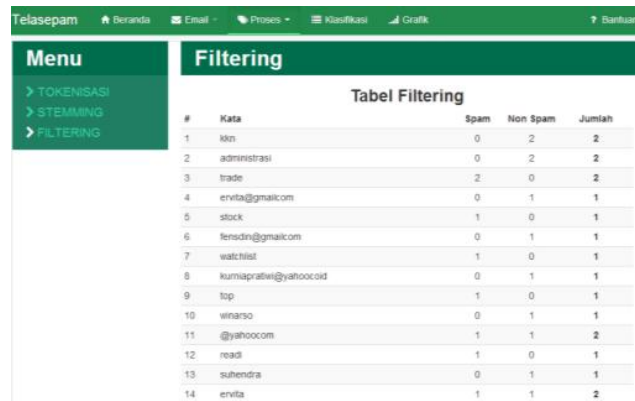


SPAM Kata : 31	NSPAM Kata : 20
subject	subject
biggest	administrasi
alert	kin
year	besi
must	suhendra
read	winarso
shern@hnpocozi	kurniapratiwi@yahoo.co.id
erita	fensdn@gmail.com
@yahooom	erita@gmail.com
top	erita
watchlist	@yahooom
stock	sih
could	dicermati
than	ajar
drive	pratik
trade	attach
thursdai	administrasi
august	kin
symbol	per
trade	okt
target	
price	
compani	
imogo	
free	
report	
how	
build	
last	
wealth	
readi	

Gambar 5. Form Perubahan Teks

Tampilan proses *stemming* yaitu proses mengubah kata berimbuhan menjadi kata dasar. Form ini akan menampilkan total kata yang sudah di proses dan hasil kata-kata yang sudah jadi kata dasar. Pada pemrosesan *stemming*, akan berjalan lama dikarenakan setiap katanya akan di cek bahasanya kemudian melakukan penghapusan imbuhan sesuai struktur yang ada pada morfologi atau dalam sistem ini pada *library porter stemmer* untuk morfologi bahasa inggris dan *porter stemming* untuk morfologi bahasa inggris.

5.4.5 Tampilan Form Pemilahan Teks

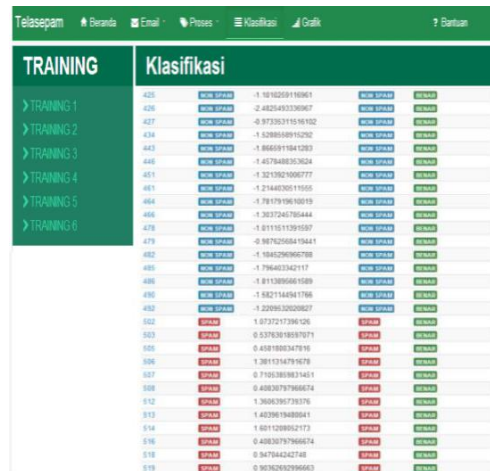


#	Kata	Spam	Non Spam	Jumlah
1	kin	0	2	2
2	administrasi	0	2	2
3	trade	2	0	2
4	erita@gmail.com	0	1	1
5	stock	1	0	1
6	fensdn@gmail.com	0	1	1
7	watchlist	1	0	1
8	kurniapratiwi@yahoo.co.id	0	1	1
9	top	1	0	1
10	winarso	0	1	1
11	@yahooom	1	1	2
12	readi	1	0	1
13	suhendra	0	1	1
14	erita	1	1	2

Gambar 6. Form Pemilahan Teks

Tampilan *filtering* yaitu proses penyaringan kata yang memiliki jumlah frekuensi yang sama kemudian digabungkan menjadi satu. Form ini juga menampilkan total kata yang sudah di proses dan kata-kata yang diurutkan berdasarkan banyak kata. Kata yang memiliki frekuensi paling banyak terdapat pada urutan pertama.

5.4.6 Tampilan Form Klasifikasi



TRAINING	ID	STATUS AWAL	NILAI HASIL PROBABILITAS	NILAI KEBAKARAN
TRAINING 1	425	spam	-1.18102259118061	spam
TRAINING 1	426	spam	-2.4823493230907	spam
TRAINING 1	427	spam	-0.97325191516102	spam
TRAINING 1	434	spam	-1.5286808915292	spam
TRAINING 1	443	spam	-1.866891841283	spam
TRAINING 1	446	spam	-1.4578488353624	spam
TRAINING 1	451	spam	-1.3213821008777	spam
TRAINING 1	461	spam	-1.2144030111935	spam
TRAINING 1	464	spam	-1.781919610019	spam
TRAINING 1	466	spam	-1.3837246780444	spam
TRAINING 1	478	spam	-1.6111611391087	spam
TRAINING 1	479	spam	-0.9870256819441	spam
TRAINING 1	482	spam	-1.1846296066788	spam
TRAINING 1	485	spam	-1.796423342117	spam
TRAINING 1	486	spam	-1.811389661088	spam
TRAINING 1	486	spam	-1.6821148541766	spam
TRAINING 1	482	spam	-1.2208632003827	spam
TRAINING 1	502	spam	1.0737217386126	spam
TRAINING 1	503	spam	0.5376381883771	spam
TRAINING 1	505	spam	0.4081808347816	spam
TRAINING 1	506	spam	1.3811216471678	spam
TRAINING 1	507	spam	0.71053888634431	spam
TRAINING 1	508	spam	0.40838787868674	spam
TRAINING 1	512	spam	1.3606396739376	spam
TRAINING 1	513	spam	1.4829819480041	spam
TRAINING 1	514	spam	1.4011208621773	spam
TRAINING 1	516	spam	0.40838787868674	spam
TRAINING 1	518	spam	0.347844242748	spam
TRAINING 1	519	spam	0.3678282896683	spam

Gambar 7. Form Klasifikasi

Tampilan *form* klasifikasi yang menampilkan hasil data *email* yang telah melalui proses-proses *text mining* dan kemudian dilakukan perhitungan dengan metode *naive bayes*. Form ini menampilkan data *email* yang terdiri dari id *email*, status awal, nilai hasil probabilitas dan hasil klasifikasi serta nilai kebenaran.

5.4.7 Tampilan Form Grafik Keakurasian Sistem



Gambar 8. Form Grafik Keakurasian Sistem

Tampilan hasil grafik keakurasian sistem terhadap jumlah *email training* yang telah di klasifikasi. Grafik keakurasian sistem di hitung dengan prosentase.

6. KESIMPULAN

- Hasil percobaan yang telah dilakukan dalam proses *text mining* memiliki hasil yang cukup baik dalam pemrosesan kata melalui proses *tokenisasi*, *stemming* dan *filtering* untuk memproses data menjadi kata. Hanya saja dalam proses *stemming* membutuhkan waktu yang lama dalam pemrosesan dikarenakan lamanya dalam pengecekan bahasa pada tiap katanya.
- Hasil percobaan yang telah dilakukan dalam klasifikasi *email spam* menggunakan metode *naive bayes* menghasilkan nilai keakurasian yang cukup tinggi.
- Sistem yang telah dibuat mampu menghasilkan keakurasian sebesar 89,6%.

7. DAFTAR PUSTAKA

- [1] Graham, Paul. 2002. A Plan for Spam.
(<http://www.paulgraham.com/spam.html>), Diakses: 3 november 2012).
- [2] Feldman, Ronen., Sanger, James. 2007. *The Text Mining Handbook: Advanced Unstructure Data*. New York : Cambridge University Press.
- [3] Rachli, Muhamad. 2007. *Email Filtering Menggunakan Naive Bayes*. Tugas Akhir. Program Studi Teknik Elektro, Institut Teknologi Bandung, Bandung.
- [4] Anugroho, Prasetyo. 2010. *Klasifikasi Email Spam Dengan Metode Naive Bayes Classifier Menggunakan Java Programming*. Skripsi. Politeknik Elektro Negeri Surabaya, Institut Teknologi Sepuluh Nopember, Surabaya.
- [5] Han, Jiawei and M. Kamber. 2001. *Data Mining: Concepts and Techniques*. USA: Academic Press.
- [6] Pop, Ioan. 2006. *An approach of the Naive Bayes classifier for the document classification*. Jurnal. Volume 14, No.4,
<http://www.emis.de/journals/GM/vol14nr4/pop/pop.pdf>.