METODE STEMMING SEBAGAI PREPROCESSING PADA FILTER KATA PORNO MELALUI ASPEK PENDIDIKAN

Moh. Sulhan[1], Rachman Kurniawan[2]
[1]*Sistem Informasi, Fakultas Teknologi Informasi, Universitas Kanjuruhan Malang*
Jl. Soedanco Supriadi No. 48 Malang
Telp : (0341) 801488, Ext: 215
[2]*Manajemen Informatika, Fakultas Teknologi Informasi, Universitas Kanjuruhan Malang*
Jl. Soedanco Supriadi No. 48 Malang
E-mail: sulhan.unikama@yahoo.com, rachman.kurni@gmail.com

ABSTRAK

Informasi merupakan pengetahuan yang telah diterima oleh seseorang dari berbagai macam cara diantaranya didapatkan dari pembelajaran, pengalaman, atau instruksi. Ada banyak macam informasi yang dapat diterima oleh seseorang, tentunya informasi yang diinginkan adalah informasi yang berkualitas bagi sipenerimanya, Namun kenyataannya, tidak selamanya seseorang mendapatkan informasi yang berkualitas tersebut, misalkan saja informasi yang mengarah pada informasi yang mengandung pornografi. Informasi yang mengarah pada isi yang mengandung porno saat ini sangat mudah diterima oleh anak yang masih belum siap menerimanya dari berbagai media diantaranya adalah CD, majalah, televisi, internet, dan lain sebagainya, bahkan di dunia pendidikan baik sengaja atau tidak sengaja mendapat informasi yang mengarah pada pornografi, hal ini ditunjukkan seperti contoh terjadi di sidoarjo yang sangat disayangkan adanya tata bahasa yang dianggap terlalu vurgar dalam soal ujian salah satu mata pelajaran yang disajikan untuk ujian tengah semester. Berdasarkan permasalahan di atas, maka dalam penelitian ini dikembangkan sebuah aplikasi yang mampu mencegah konten porno pada soal ujian berbahasa indonesia. Aplikasi yang dikembangkan terdiri dari 5 (lima) modul yaitu Tokenizing, Affix Porn Filtering, Stemming, Root Porn Filtering, dan Phrase Porn Filtering. Dan Aplikasi telah berhasil dikembangkan dan diimplementasikan dalam lingkungan sistem operasi windows, dan menunjukkan bahwa metode kombinasi filtering lebih baik dibandingkan dengan metode filtering yang lain yakni pada kata affix, root, dan phrase.

Kata Kunci : konten porno, tokenizing, stemming, filtering

ABSTRACT

Information is a knowledge that has been received by the person from whom obtained a variety of ways of learning, experience, or instruction. There are many kinds of information that can be accepted by someone, of course, the information that is needed is a quality information for the recipient, but fact people are not always get the quality information, such as the information that leads to pornographic content. Information leads to pornographic content nowadays is easily accepted by children who are not ready to accept from a variety of media including the compact disc (CD), magazines, television, internet, and so on, even in their education either intentionally or unintentionally get information leads to pornography, As happened in Sidoarjo unfortunately the grammar is considered too vulgar in one of the exam subjects which are presented for the midterm. Based on the problems above, so in this study, an application is developed that is able to preventing pornographic contents from examinations in "Indonesia Language".The application which are developed consisting of 5 (five) module that is tokenizing, Stemming, Affix Porn Filtering, Root Porn Filtering, and Phrase Porn Filtering. Applications have been successfully developed and implemented in the Windows operating system environment, and demonstrate that the combination of filtering method is better than the other filtering methods that the word affix, root, and phrases.

Keywords: pornographic content, tokenizing, Stemming, Filtering.

1. PENDAHULUAN

Informasi merupakan pengetahuan yang telah diterima oleh seseorang dari berbagai macam cara diantaranya didapatkan dari pembelajaran, pengalaman, atau instruksi (Wiki, 2011). Ada banyak macam informasi yang dapat diterima oleh seseorang dalam kegiatan sehari – hari, tentunya informasi yang diinginkan adalah informasi yang berkualitas bagi sipenerimanya, informasi berkualitas itu sendiri merupakan informasi yang mengandung unsur – unsur seperti memadai, relevan, terpercaya, dan sikap mental yang menjunjung tinggi prinsip moral dan kebenaran. Namun kenyataannya, tidak selamanya seseorang mendapatkan informasi yang berkualitas tersebut, misalkan saja informasi yang mengarah pada informasi yang mengandung porno, hal ini sangat mengganggu bagi seseorang khususnya orang yang

belum siap menerimanya secara mental, contohnya seseorang tersebut adalah orang atau anak yang masih mengenyam pendidikan dasar, menengah, atau sekolah tinggi.

Informasi yang mengarah pada isi yg mengandung porno saat ini sangat mudah diterima oleh anak yang tentunya anak – anak yang masih belum siap menerimanya dari berbagai media diantaranya adalah CD, majalah, televisi, internet, dan lain sebagainya, bahkan di dunia pendidikanpun mereka baik sengaja atau tidak sengaja mendapat informasi yang mengarah pada pornografi, hal ini ditunjukkan seperti contoh kejadian di sidoarjo yang sangat disayangkan adanya tata bahasa yang dianggap terlalu vurgar dalam soal ujian salah satu mata pelajaran yang disajikan untuk ujian tengah semester yang dikeluarkan oleh Diknas kabupaten tersebut (Ismail, 2009).

Dalam pembuatan soal ujian dapat dilakukan secara manual artinya dibuat oleh seorang guru/pengajar atau dibuat dengan mesin secara otomatis dalam hal ini adalah komputer dari bahan ajar yang telah dipelajari sebelumnya oleh siswanya, seperti halnya yang telah dilakukan oleh Wei CHEN dan rekan – rekannya yang di tulis dalam papernya yaitu pembangkitan pertanyaan secara otomatis dari teks informasi (Wei CHEN, 2011).

Pada paper lain dikatakan bahwasanya dalam pembuatan pertanyaan otomatis yang berasal dari suatu narasi dapat menggunakan pendekatan metode NLP (Natural Language Processing), yang dikembangkan sebagai upaya untuk memudahkan pengguna komputer dalam berinteraksi dengan komputer. Melalui teknologi NLP ini, pengguna komputer berkomunikasi dengan komputer dengan menggunakan bahasa sehari – hari manusia, bukan hanya menggunakan bahasa formal komputer.

Penelitian pembangkitan atau pembuatan pertanyaan secara otomatis telah banyak dilakukan salah satunya menggunakan pendekatan metode NLP, seperti halnya penelitian yang telah dilakukan oleh Idawati Widjaja dalam Tugas Akhirnya yaitu Perancangan dan Pembuatan Perangkat Lunak Pembuat Soal dengan Menggunakan NLP. Dalam Penelitian yang dilakukan oleh Idawati Widjaja (2006) akan sangat membantu untuk menghindari hal-hal yang tidak diinginkan seperti halnya terjadinya kebocoran soal ujian yang sering terjadi akhir-akhir ini yang akan mengakibatkan kerugian bagi pihak sekolah maupun negara sebagai pembuat soal tersebut. Kerugian tersebut dapat berupa kerugian waktu, pikiran, tenaga, dan biaya, karena soal yang sudah bocor harus dibuat kembali

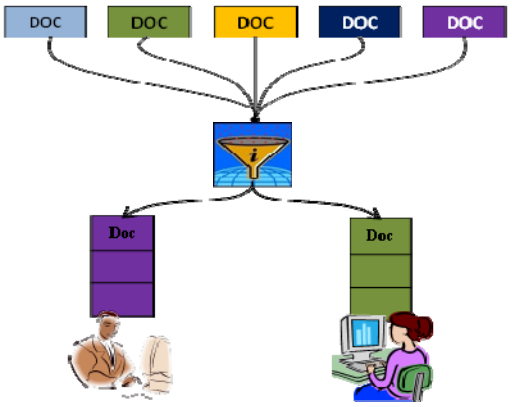
Melihat dari penelitian-penelitian yang dilakukan sebelumnya dalam perancangan dan pembuatan soal secara otomatis belum dilakukan penelitian yang melibatkan text filtering yang dimungkinkan terjadi menurunnya kualitas informasi dalam soal-soal yang telah dibuat, maka dengan ini penelitian kali ini akan melihat seberapa jauh kemampuan komputer dan algoritma yang digunakan untuk melakukan filter teks khususnya teks yang berisi porno pada pembuatan soal ujian berbahasa indonesia.

Tujuan dari penelitian ini adalah merumuskan suatu metode *text filtering* berbasis *keyword* untuk menyelesaikan permasalahan filter teks mengandung porno pada pertanyaan–pertanyaan yang dibuat untuk soal ujian.

Selain itu penelitian ini dibuat untuk dapat memberikan kontribusi sebagai *text filtering* mengandung porno pada penelitian – penelitian yang telah dilakukan sebelumnya yaitu pembuatan atau pembangkitan pertanyaan secara otomatis yang berasal dari suatu narasi. Sehingga penelitian ini berguna dan dapat memberi manfaat untuk diaplikasikan pada interaksi manusia dengan komputer khususnya dalam hal *text filtering* yang mengandung porno pada pertanyaan-pertanyaan yang akan dijadikan sebagai soal ujian

2. INFORMATION FILTERING

Information Filtering (IF) adalah salah satu metode yang secara cepat berkembang untuk mengelola aliran informasi yang datang kepada pengguna. Tujuan dari Information Filtering adalah membawa pengguna kepada hanya informasi yang relevan terhadap kebutuhan mereka. Sistem IF telah dikembangkan beberapa tahun terakhir ini untuk berbagai domain aplikasi. Beberapa contoh dari aplikasi pemfilteran adalah pemfilteran e-mail personal berdasarkan profil personal, pemfilter browser yang memblok informasi yang tidak sesuai, filter yang dirancang agar anak-anak hanya dapat mengakses informasi yang sesuai bagi mereka, dan lain-lain. Secara umum tujuan utama dari IF adalah mengarahkan informasi yang paling berharga (relevan) kepada pengguna secara otomatis dan membantu pengguna memanfaatkan waktu membaca dokumen yang terbatas secara lebih optimal (Rila Mandala,2006).



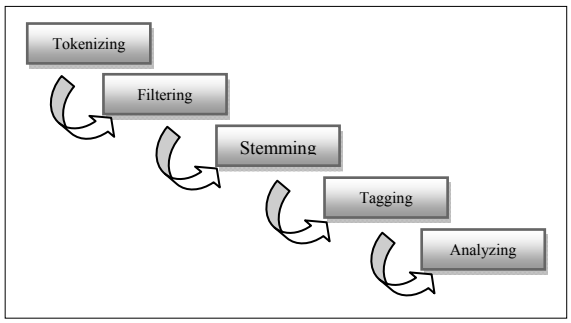
Gambar 1. Information Filtering

3. TEXT MINING

Text mining adalah proses semi otomatis penggalian pola (pengetahuan dan informasi yang berguna) dari sejumlah sember data berukuran besar yang tidak restruktur (Arfiani, Retno Laila, 2010). Text mining memiliki tujuan yang sama dan proses yang sama dengan data mining. Yang membedakan hanyalah sumber data yang digunakan. Pada data mining data yang digunakan adalah data terstruktur sedangkan dalam text mining data yang digunakan adalah data yang tidak terstruktur berupa teks. Manfaat dari text mining jelas pada area dimana sejumlah besar data tekstual dihasilkan seperti di bidang hukum, pengobatan, keuangan, dll.

Text mining menggunakan natural language processing untuk memasukkan stuktur kedalam kumpulan teks. Natural language processing(NLP) adalah komponen penting dari text mining dan merupakan sub-bidang dari artificial intelligence dan komputasional linguistik. NLP berupaya memecahkan masalah untuk memahami bahasa alami manusia, dengan segala aturan gramatika dan semantiknya, dan mengubah bahasa tersebut menjadi representasi formal yang dapat diproses oleh komputer

Tahapan secara umum yang dilakukan dalam Text Mining sebagai berikut (Harlian, Milkha Ch, 2011):

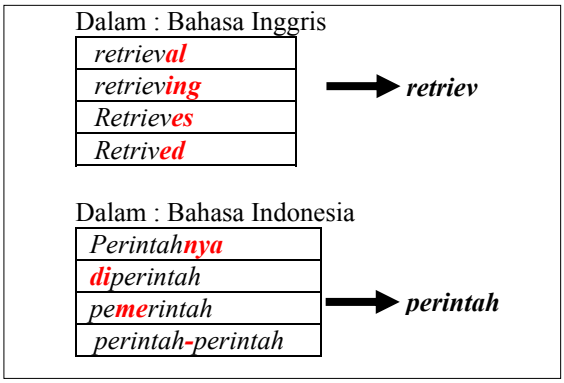


Gambar 2. Gambaran Tahapan Text Mining

- ***Tokenizing***, merupakan tahapan pemotongan string input berdasarkan tiap kata yang menyusunnya
- ***Filtering***, merupakan tahapan mengambil kata-kata penting dari hasil token. Dapat menggunakan algoritma stop list (membuang kata yang kurang penting) atau word list (menyimpan kata penting)
- ***Stemming,*** merupakan tahapan mencari root kata dari tiap kata hasil filtering
- ***Tagging***, tahapan mencari bentuk awal/root dari tiap kata lampau atau kata hasil stemming
- ***Analyzing***, merupakan tahapan penentuan seberapa jauh keterhubungan antar kata-kata antar dokumen yang ada

4. STEMMING

Salah satu teknik yang digunakan dalam Pemrosesan Bahasa Alami (NLP) untuk mengembalikan bentuk suatu kata menjadi bentuk root-nya (root word) yaitu penggunaan teknik *stemming*, dimana *stemming* ini adalah menghilangkan prefiks dan sufiks dari data yang diambil dari keri (Grossman, 2011). Sebagai contoh kata bersama, kebersamaan, menyamai, akan di-stem ke root wordnya yaitu "sama". Proses stemming pada teks berBahasa Indonesia berbeda dengan stemming pada teks berbahasa Inggris. Pada teks berbahasa inggris, proses yang diperlukan hanya proses menghilangkan sufiks. Sedangkan pada teks berbahasa Indonesia, selain sufiks, prefiks, dan konfiks juga dihilangkan (Ledy Agusta, 2009). Berikut seni *stemming* dalam bahasa inggris dan bahasa Indonesia



Gambar 3. Stemming dalam Bahasa Inggris dan Indonesia

Stemming untuk Bahasa Indonesia telah dikembangkan antara lain yang menggunakan aturan berdasarkan algoritme Porter oleh Akhmadi (2002) yang hanya melakukan pemotongan prefiks dan oleh Ridha (2002) yang melakukan pemotongan prefiks dan sufiks. *Stemming* berdasarkan kamus untuk

Bahasa Indonesia juga telah dikembangkan oleh (Nazief, 2009)

5. PENCOCOKAN STRING (*String Matching*)

Pengertian *string* menurut *Dictionary of Algorithms and Data Structures, National Institute of Standards and Technology (NIST)* adalah susunan dari karakter-karakter (angka, alfabet atau karakter yang lain) dan biasanya direpresentasikan sebagai struktur data *array*. *String* dapat berupa kata, frase,atau kalimat (NIST, 2012).

Pencocokan *string* (*string matching* atau *pattern matching*) merupakan bagian penting dari sebuah proses pencarian *string* (*string searching*) dalam sebuah dokumen. Hasil dari pencarian sebuah *string* dalam dokumen tergantung dari teknik atau cara pencocokan *string* yang digunakan. Pencocokan *string (string matching)* menurut *Dictionary of Algorithms and Data Structures, National Institute of Standards and Technology (NIST),* diartikan sebagai sebuah permasalahan untuk menemukan pola susunan karakter *string* di dalam *string* lain atau bagian dari isi teks.

Persoalan pencarian *string* dapat dirumuskan dirumuskan sebagai berikut (Syaroni & Munir, 2005) :

1. Teks (*text*), yaitu (*long*) *string* yang panjangnya *n* karakter
2. *Pattern*, yaitu *string* dengan panjang *m* karakter (*m*<*n*) yang akan dicari di dalam teks.

Carilah (*find* atau *locate*) lokasi pertama di dalam teks yang bersesuaian dengan *pattern*. Aplikasi dari masalah pencocokan string antara lain pencarian suatukata di dalam dokumen (misalnya menu *Find* di dalam *Microsoft Word*)

6. REGULAR EXPRESSION

Regular Expression atau lebih dikenal regex adalah suatu cara pencarian suatu pola string tertentu. Salah satu contoh sederhana adalah mencari pola string suatu email. Pada email terdapat aturan bahwa harus terdapat suatu symbol '@' yang tertera pada email tersebut. Pencarian pola string menggunakan regular expression akan lebih menghemat waktu. Cukup dengan memasukkan pola string tertentu, maka semua string yang memiliki pola yang sama akan diolah sesuai yang diinginkan. Untuk dapat mengakses suatu regular expression, diperlukan suatu engine yang merupakan suatu potongan program tertentu. Sudah banyak bahasa pemrograman yang menyertakan regular expression engine, untuk mengakses pola string tertentu. Bahasa pemrograman tersebut diantaranya adalah Java, PHP, Visual Basic, Perl (Amak, 2008).

7. PEMBAHASAN

Untuk mendukung proses penyelesaian penelitian ini maka diperlukan tahapan-tahapan kerja yang harus dilakukan. Tahapan kerja yang dilakukan meliputi desain aplikasi dan implementasi aplikasi.

A. Desain Aplikasi

Untuk membuat aplikasi *text filtering* mengandung porno berbasis keyword, perlu dilakukan desain aplikasi yang meliputi tiga bagian yaitu desain data, desain proses, dan desain antar muka.

a. Desain Data

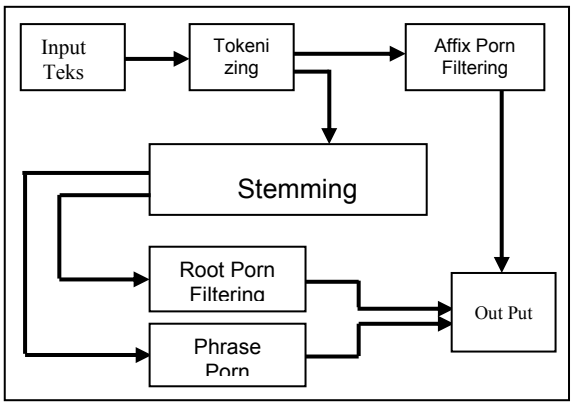
1. Data Input
   Pemilihan data pada desain ini, merupakan data-data soal yang akan dijadikan sebagai soal ujian.
2. Keyword
   Data *Keyword* atau kata kunci merupakan suatu kumpulan kata kunci yang diasumsikan sebagai kata yang mengandung porno yang tersimpan di dalam suatu file.

Ada 3 macam data yang akan dijadikan sebagai data kata kunci (Keyword-based), yakni :

1. *Root Keyword-Based*
2. *Affix Keyword-Based*
3. *Phrase Keyword-Based*

b. Desain Proses

Pada bagian ini akan diterangkan mengenai proses-proses yang terdapat dalam aplikasi text filtering porno berbasis keyword. Adapun Desain Proses yang ada pada aplikasi adalah sebagai berikut:

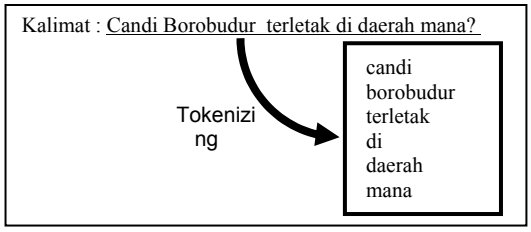


Gambar 4. Gambaran Umum Proses Filtering

Penjelasan dari masing modul di atas adalah sebagai berikut :

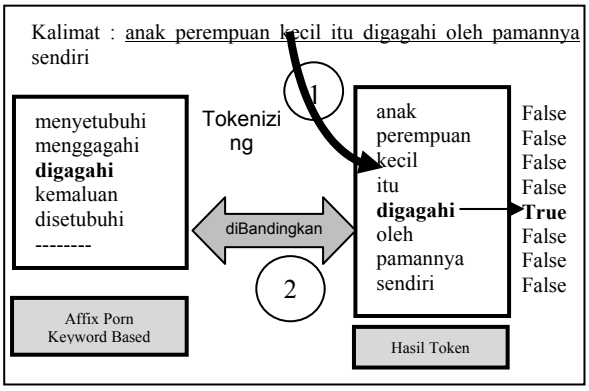
1. **Input Teks**
   Merupakan suatu proses memasukkan data soal yang sudah dibuat sebelumnya.
   Contoh : *Candi borobudur terletak di daerah mana?*

2. Pemotongan Data Input *(Tokenizing)*



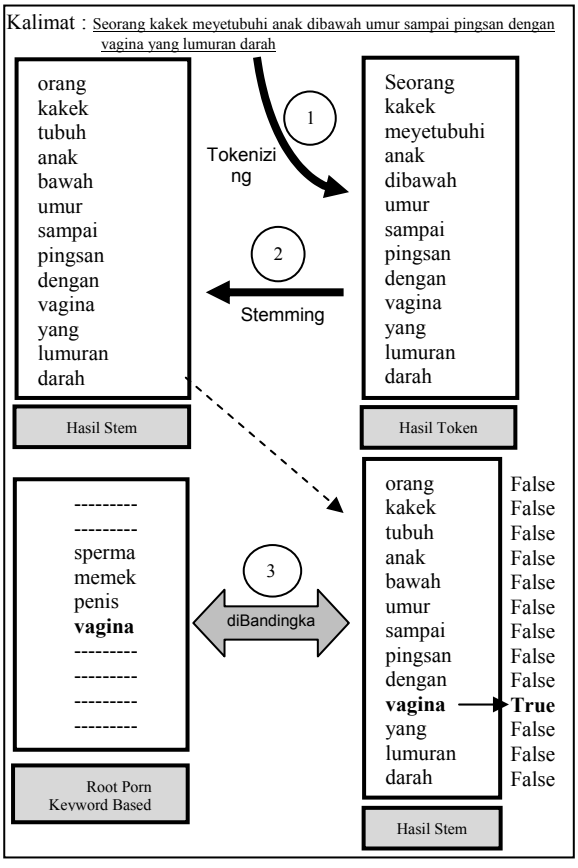
Gambar 5. Proses Pemotongan (Tokenizing)

3. *Affix Porn Filtering*



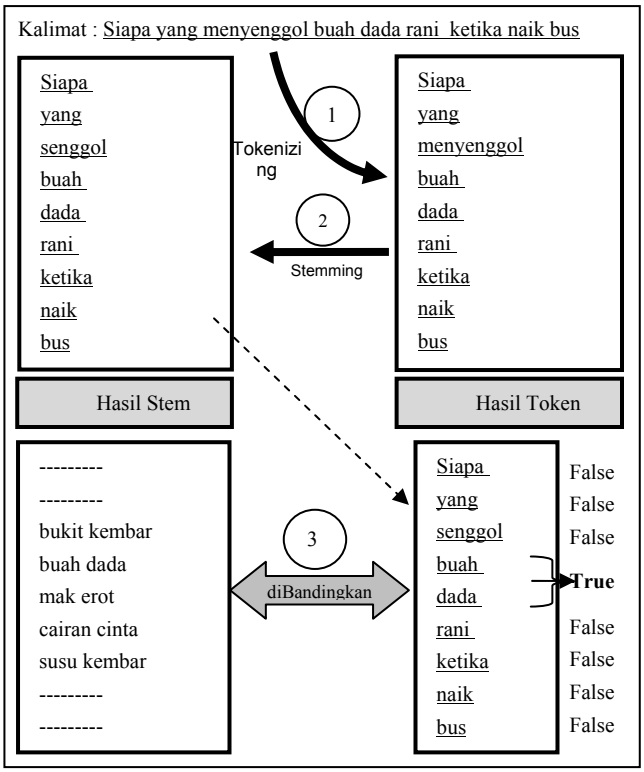
Gambar 6. Proses Affix Porn Filtering

4. *Root Porn Filtering*



Gambar 7. Proses Root Porn Filtering

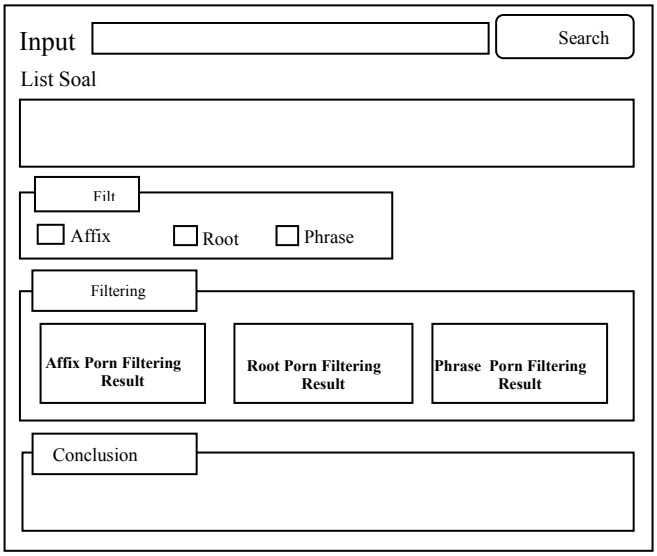
5. *Phrase Porn Filtering*



Gambar 8. Proses Phrase Porn Filterin

c. Desain Antar Muka

Desain tampilan program aplikasi yang akan dikembangkansepertipadagambarberikut.



Gambar 9. Desain Antar Muka (Interface)

8. UJI COBA PERANGKAT LUNAK

Pada sub bab ini akan dibahas tentang skenario uji coba tiap modul yang dibuat. Dalam pengujian akan dilakukan beberapa skenario ujicoba pada tiap modul yang dibuat, dimana dari hasil pengujian ini akan dianalisa kinerjanya yaitu dengan menghitung nilai akurasi ***true positive, false positive, true***

negative dan ***false negative***. Dan proses pengujian yang akan dilakukan adalah sebagai berikut:

a) Pengujian dengan menyisipkan *Stemming* sebagai *Preprocessing*.
Adalah Pengujian yang dilakukan dengan menyisipkan metode stemming suatu *filtering* yang dilakukan pada semua jenis kata yaitu : kata berimbuhan, kata dasar, dan kata frase.

b) Pengujian tanpa menyisipkan *Stemming* sebagai *Preprocessing*
Adalah Pengujian yang dilakukan tanpa menyisipkan metode stemming pada suatu *filtering* yang dilakukan pada semua jenis kata yaitu : kata berimbuhan, kata dasar, dan kata frase.

c) Pengujian dengan cara kombinasi
Adalah Pengujian yang dilakukan pada suatu bagian yang perlu disisipkan *stemming* dan ada juga bagian yang tidak perlu disisipkan *stemming*. Pada penelitian ini untuk pengujian dengan cara kombinasi tersebut pada *filtering* kata berimbuhan tidak disisipkan *stemming*, sedangkan untuk kata dasar dan kata frase akan disisipkan stemming.

Data yang digunakan adalah data soal yang disimpan pada file notepad yang memiliki ekstensi .txt. Data soal didapat dari bahan pelajaran anak setingkat SD, SMP, dan SMA dengan mata pelajaran sejarah, bahasa Indonesia, dan sosiolagi. Sumber data soal tersebut berasal dari buku pelajaran dan kumpulan soal – soal ujian negara (UN) yang diambil dari internet yang berbahasa indonesia dan dikelompokkan berdasarkan nama mata pelajaran yakni bahasa *indonesia.txt*, *sejarah.txt*, dan *sosiologi.txt*.

Dari data soal yang sudah disiapkan yang terbagi menjadi tiga macam mata pelajaran kemudian masing – masing data soal tersebut disisipkan kalimat yang dianggap mengandung kata porno, baik kata porno yang berimbuhan (*affix porn*), kata dasar (*root porn*), dan kata frase (*phrase porn*).

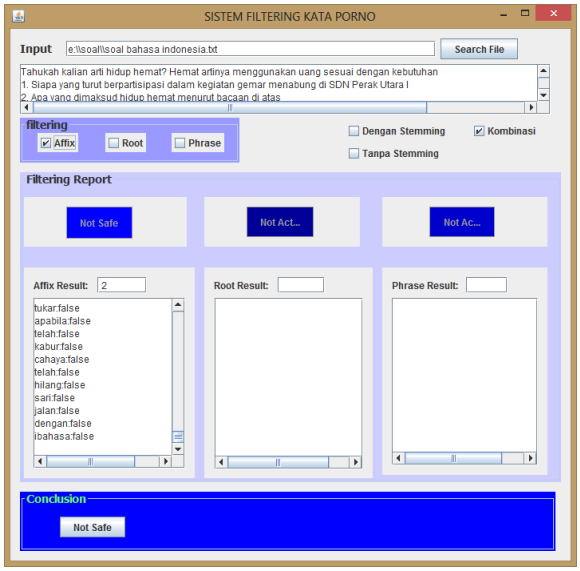
Tabel 1
Data Soal yang akan uji cobakan

No	Jenis soal yang diuji	Jumlah Kata	Kata Porno	Kata Tidak Porno
1	Bhs Indonesia	421	14	407
2	Sejarah	238	5	233
3	Sosiologi	491	17	474
Jumlah		1150	36	1114

Berikut akan ditunjukkan proses pengujian yang dilakukan sesuai dengan skenario yang telah dijelaskan di sebelumnya :

a. **Pengujian pada kata Berimbuhan *(Affix Filtering)***

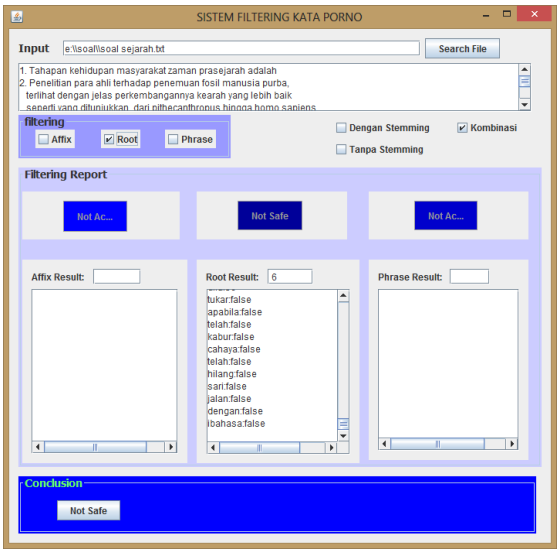
Berikut ditunjukkan user interface untuk pengujian pada kata berimbuhan, yang ditunjukkan sebagai berikut



Gambar 10. Pengujian pada kata berimbuhan (*Affix Filtering*)

b. **Pengujian pada kata dasar (*Root Filtering)***

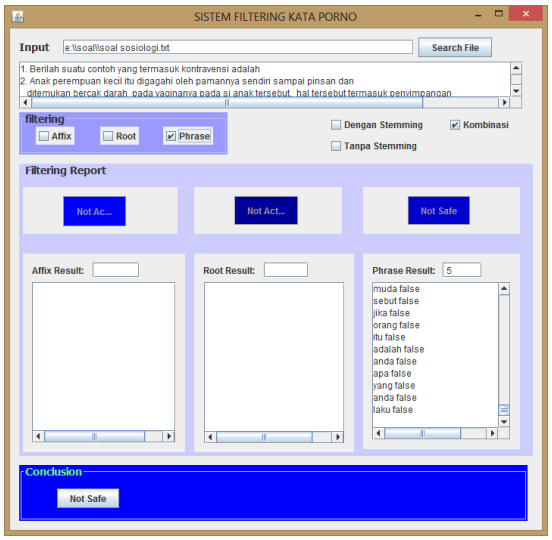
Berikut ditunjukkan user interface untuk pengujian pada kata dasar, yang ditunjukkan sebagai berikut



Gambar 11. Pengujian pada kata dasar (*Root Filtering*)

c. Pengujiam pada kata Frase (*Phrase Filtering)*

Berikut ditunjukkan user interface untuk pengujian pada kata frase, yang ditunjukkan sebagai berikut



Gambar 12. Pengujian pada kata frase (*Phrase Filtering*)

Setelah dilakukan pengujian pada kata Berimbuhan, Kata Dasar. dan Kata Frase, dimana masing-masing telah dilakukan pengujian seperti yang dijelaskan sebelumnya yakni dengan menyisipkan *Stemming*, tanpa menyisipkan *Stemming*, dan dengan cara Kombinasi maka akan didapatkan data hasil pengujian yang akan ditunjukkan pada tabel hasil uji coba. Dari Hasil beberapa pengujian yang telah dilakukan dengan menggunakan beberapa skenario pengujian dapat di jabarkan pada tabel untuk dapat dianalisa dan diambil kesimpulan dari proses penelitian tersebut.

Tabel 2
Hasil Pengujian dengan menyisipkan *Stemming*

Soal Yang diuji	Jumlah Kata	Kata Porno	Kata tidak Porno	Jumlah kata yang Terdeteksi			Jmlh
				root	affix	phrase	
Bahasa Indonesia	421	14	407	9	0	3	12
Sejarah	238	5	233	6	0	3	9
Sosiologi	491	17	474	10	0	5	15
Jumlah	1150	36	1114	25	0	11	36

Data pada tabel di atas merupakan hasil ujicoba yang dilakukan pada data soal dengan menyisipkan pada stemming. Hal ini akan dilakukan juga pada data yang sama, namun dengan metode tanpa menyisipkan stemming dan metode kombinasi.

Tabel 3
Hasil Pengujian tanpa menyisipkan *Stemming*

Soal Yang diuji	Jumlah Kata	Kata Porno	Kata tidak Porno	Jumlah kata yang Terdeteksi			Jmlh
				root	affix	phrase	
Bahasa Indonesia	421	14	407	8	2	1	11
Sejarah	238	5	233	6	0	1	7
Sosiologi	491	17	474	8	2	2	12
Jumlah	1150	36	1114	22	4	4	30

Soal Yang diuji	Jumlah Kata	Kata Porno	Kata tidak Porno	Jumlah kata yang Terdeteksi			Jmlh
				root	affix	phrase	
Bahasa Indonesia	421	14	407	9	2	3	14
Sejarah	238	5	233	6	0	3	9
Sosiologi	491	17	474	10	2	5	17
Jumlah	1150	36	1114	25	4	11	40

Tabel 4
Hasil Pengujian dengan cara Kombinasi

9. ANALISIS HASIL UJI COBA

Analisis Dari Hasil Uji Coba yang telah dijabarkan dalam bentuk tabel adalah sebagai berikut :

Pada hasil uji coba tersebut jika dilihat dari data yang diujicobakan seperti yang sudah dijelaskan sebelumnya pada tabel 1 yakni kata porno yang disisipkan sejumlah 14 kata porno untuk mata pelajaran Bahasa Indonesia, sejumlah 5 kata porno untuk mata pelajaran Sejarah, dan sejumlah 17 kata porno untu mata pelajaran Sosiologi.

Hasil Pengujian dengan menyisipkan Stemming dihasilkan kata yang terdeteksi untuk bahasa Indonesia 12 kata porno, sejarah 9 kata porno, dan Sosiologi sejumlah 15 kata porno, sehingga kata porno yang tidak terdeteksi untuk mata pelajaran bahasa Indonesia 2 kata porno, Sosiologi 2 kata porno, sedangkan untuk mata pelajaran Sejarah kata yang terdeteksi melebihi kata porno yang disisipkan yakni 9 kata, artinya untuk mata pelajaran Sejarah tersebut kata yang terdeteksi lebih 5-9=-4 (-4 menunjukkan jumlah kata yang yang terdeteksi melebihi dari jumlah kata yang disisipkan yakni sejumlah 5 kata porno). Hal ini dimungkinkan kata yang dianggap bukan porno ternyata terdeteksi sebagai kata porno atau ***False Positive*** dan bisa juga kata porno yang disisipkan akan terdeteksi lebih dari satu kali. Dan untuk kata porno yang berimbuhan (*Affix porn*) tidak ada yang terdeteksi, hal ini terjadi karena semua kata berimbuhan yang dianggap porno menjadi kata dasar karena pengujian ini semuanya

melewati proses *Stemming*, seperti yang telah dijelaskan sebelumnya bahwasanya jika kata porno berimbuhan dijadikan sebagai kata dasar, maka kata tersebut tidak lagi menjadi porno sepert : disetubuhi →kata dasarnya : tubuh, yang mendapatkan imbuhan di-se-tubuh-i.

Hasil Pengujian tanpa menyisipkan Stemming dihasilkan kata yang terdeteksi untuk bahasa Indonesia 11 kata porno, sejarah 7 kata porno, dan Sosiologi sejumlah 12 kata porno, sehingga kata porno yang tidak terdeteksi untuk mata pelajaran bahasa Indonesia 3 kata porno, Sosiologi 5 kata porno, sedangkan untuk mata pelajaran Sejarah kata yang terdeteksi tetap melebihi kata porno yang disisipkan yakni 7 kata, artinya untuk mata pelajaran Sejarah tersebut kata yang terdeteksi lebih 5-7=-2 (-2 menunjukkan jumlah kata yang yang terdeteksi melebihi dari jumlah kata yang disisipkan yakni sejumlah 5 kata porno). Hal ini dimungkinkan kata yang dianggap bukan porno ternyata terdeteksi sebagai kata porno atau ***False Positive*** dan bisa juga kata porno yang disisipkan akan terdeteksi lebih dari satu kali, atau bahkan terjadi karena tanpa melewati proses *stemming* sehingga tidak terdeteksi.

Pada proses pengujian ini hasilnya lebih buruk diabndingkan dengan proses pengujian sebelumnya yang melakukan dengan menyisipkan metode *Stemming,* terbukti dengan ditunjukkannya kata porno yang terdeteksi lebih sedikit dibandingkan dengan yang sebelumnya.

Hasil Pengujian dengan cara Kombinasi dihasilkan kata yang terdeteksi untuk bahasa Indonesia 14 kata porno, sejarah 9 kata porno, dan Sosiologi sejumlah 17 kata porno. Pada pengujian ini ditunjukkan bahwanya kata porno yang disisipkan semuanya terdeteksi. Namun untuk mata pelajaran Sejarah kata yang terdeteksi tetap melebihi kata porno yang disisipkan yakni 9 kata, artinya untuk mata pelajaran Sejarah tersebut kata yang terdeteksi lebih 5-9=-4 (-4 menunjukkan jumlah kata yang yang terdeteksi melebihi dari jumlah kata yang disisipkan yakni sejumlah 5 kata porno).

Pada pengujian dengan menggunakan cara Kombinasi dimana cara tersebut adalah melakukan penyisipan *Stemming* pada suatu bagian tertentu yang dibutuhkan saja. Dimana dalam hal ini tanpa dilakukan penyisipan *Stemming* untuk *filtering* kata yang berimbuhan (*affix*), sedangkan untuk kata dasar (*root*) dan kata frase (*phrase*) diperlukan untuk menyisipkan *Stemming*. Hal ini menghasilkan lebih baik dibandingkan dengan pengujian – pengujian yang dilakukan sebelumnya, ditunjukkan bahwasanya semua kata porno yang disisipkan dapat terdeteksi semuanya.

10. KESIMPULAN

Dalam penelitian ini telah berhasil dibuat sebuah aplikasi *text filtering* untuk mencegah konten porno pada soal berbahasa indonesia berbasis keyword.

Setelah dilakukan Uji Coba dengan beberapa sampel data input dengan menggunakan Metode pengujian pada beberapa mata pelajaran yang sudah ditentukan sebelumnya didapatkan hasil bahwasanya penggunaan metode Kombinasi yang memperhatikan spesifikasi data input dalam hal butuh tidaknya metode *stemming* sebagai *preprocessing* disisipkan pada proses *filtering* dapat dikatakan lebih baik dibandingkan dengan menyisipkan metode *stemming* untuk semua *filtering* atau bahkan tidak menyisipkan metode *stemming* tersebut pada semua proses *filetering*. Hal ini ditunjukkan dari hasil beberapa pengujian bahwasanya penggunaan metode kombinasi pada *False Negative = 0*, artinya semua kata porno dapat terdeteksi dengan baik, berbeda dengan metode yang lain yang masih ada beberapa kata porno yang tidak terdeteksi. Hal ini terjadi karena kata yang dianggap porno bisa berupa kata berimbuhan, kata dasar, dan kata frase, dimana pengecekan kata pada kata porno yang berupa imbuhan tidak memelukan proses *stemming*, sedangkan pengecekan kata porno yang berupa kata dasar dan kata frase membutuhkan proses stemming.

11. DAFTAR PUSTAKA

Agusta, Ledy. 2009. *Perbandingan algoritma stemming porter dengan algoritma nazief & adriani untuk stemming dokumen teks bahasa Indonesia.* Konferensi Nasional Sistem dan Informatik, Bali.

Arfiani, Retno Laila. 2010. *Text dan Web Mining*, Makalah, Informatika UNS. Surakarta.

Bank Soal. Desember 2011. http://www. Banksoal .sebarin.com/.

Chaer, Abdul. 2008. *Morfologi bahasa Indonesia.* Rineka Cipta. Jakarta.

CHEN, Wei., Gregory AIST., and Jack MOSTOW, *Generating Questions Automatically from Informational Text*. Januari 2011. http://www.cs.cmu.edu/~weichen/QG.pdf

Dictionary of Algorithms and Data Structures. Maret 2012. National Institute of Standards and Technology. http://www.nist.gov/dads/.

Fernando, Hary. 2009. *Perbandingan dan Pengujian Beberapa Algoritma Pencocokan String*, Makalah IF2251 Strategi Algoritmik, Institut Teknologi Bandung.

Grossman, D. IR Book. Juli 2011. http://www.ir.iit.edu/~dagr/cs529/files /ir_book/

Harlian, Milkha Ch. *Text Mining*. Juli 2011. http://lecturer.eepis-ts.edu/~iwanarif/kuliah/dm /6Text%20Mining.pdf

Kumpulan Soal - soal Sekolah Dasar. Desember 2011. http://www.sekolahdasar.net/.

Ismail, M. Soal UTS di Sidoarjo Kebobolan Kalimat Porno. http://beritajatim.com Selasa, 27 Oktober 2009 (diakses 21 Juni 2012)

Mandala, Rila. 2006. *Evaluasi Kinerja Sistem Penyaringan Informasi Model Ruang Vektor*, Seminar Nasional Aplikaasi Teknologi Informasi (SNATI). UII. Yogyakrta.

Nazief., Bobby., dan Mirna Adriani. *Confix-Stripping: Approach to Stemming Algorithm for Bahasa Indonesia*, 2009. Fakulty of Computer Science University of Indonesia.

Regular Expressions in Java. Juni 2011. http://www.cis.upenn.edu/~matuszek/cit591-2002/Lectures/java-regex.ppt.

Syaroni Mokhamad., dan Munir Rinaldi. 2005. *Pencocokan String Berdasarkan Kemiripan Ucapan (Phoneticstring Matching) Dalam Bahasa Inggris. Seminar Nasional Aplikasi Teknologi Informasi (SNATI).Yogyakarta*

Tesaurus Bahasa Indonesia. Kamus Sinonim dan Antonim, Mei 2012. http:// sinonimkata.com

Widjaja, Idawati. 2006, *Perancangan dan Pembuatan Perangkat Lunak Pembuat Soal dengan Menggunakan NLP*, Teknik Informatika Universitas Kisten Petra Surabaya.

Yunus, Amak. 2008. *Klasifikasi Dokumen Web Berdasarkan Frase Kunci Pada Bagian Informatif.* Tesis: ITS Surabaya.

_____, Informasi. Oktober 2011**,** http://id.wikipedia.org/wiki/Informasi

.