Jurnal Sains, Teknologi dan Industri, Vol. 13, No. 2, Juni 2016, pp.151 - 159
ISSN 1693-2390 print/ISSN 2407-0939 *online*

Implementasi Metode Terms Frequency-Inverse Document Frequency (TF-IDF) dan Maximum Marginal Relevance untuk Monitoring Diskusi Online

Okfalisa[1], Ahli Hidayat Harahap[2]
[1,2]Teknik Informatika UIN SUSKA Riau
Email: [1]okfalisa@gmail.com, [2]hidayatahli@gmail.com

(Received: 2 Februari 2016; Revised: 20 Juni 2016; Accepted: 20 Juni 2016)

ABSTRAK

Pemanfaatan media sosial dalam proses pembelajaran terutama di forum diskusi online semakin meningkat. Namun, melebarnya pembahasan diluar jangkauan kajian yang seharusnya bahkan mengarah kepada perdebatan negatif serta melanggar kode etik berkomunikasi seringkali ditemukan pada diskusi online. Hal ini mendorong peningkatan peran admin ataupun tenaga pengajar dalam memonitor dan mengontrol jalannya diskusi pada forum tersebut. Dengan menerapkan metode TF-IDF dan *Maximum Marginal Relevancy* sebuah aplikasi dibangun untuk memonitor proses pelaksanaan diskusi online. Serangkaian tahapan text preprocessing meliputi pemecahan kalimat, case folding, tokenizing, filtering dan stemming dilakukan guna mengekstrak postingan baik dari topik diskusi tenaga pengajar maupun komentar peserta. Selanjutnya bobot TF-IDF, Query Relevance dan Similarity dikalkulasikan. Dengan menggunakan metode Maximum marginal relevance ekstraksi optimal ringkasan dapat diberikan dengan mengurangi redudansi dan perangkingan kalimat. Komentar yang bernilai nol (0) dari nilai keseluruhan perbandingan postingan yang telah diringkas dengan komentar diklasifikasikan sebagai komentar yang "tidak layak" dan direkomendasikan untuk dieliminasi.Dari hasil pengujian akurasi, black box test dan UAT pada salah satu topik perkuliahan sistem ini telah berhasil memonitor proses pelaksanaan diskusi online. Sehingga forum diskusi yang diperuntukkan untuk peningkatan mutu pendidikan dan pengajaran dapat dimaksimalkan serta terarah sesuai dengan harapan.

Kata Kunci: *cosine similarity*, diskusi *online*, kelayakan komentar, *Maximum Marginal Relevance*, *text processing*, *TF-IDF*.

ABSTRACT

The use of social media for learning process especially in online discussion forum is gradually increased. Nevertheless, the spreading of out of scope discussion that trigger the emergence of negative debates breaks the communication ethic code in online discussion. This push forward the increasing of administrators or instructors roles in monitoring and controlling the discussion activity during the forum session. By applying TF-IDF and Maximum Marginal Relevancy methods a software apllication is developed to monitor the discussion online activity. The list of Text Processing Phase including The sentences breakdown, case folding, tokenizing, filtering and stemming are conducted to extract the document posting from the instructors as well as member comments.Then, TF-IDF, Query Relevance and Similarity values are calculated. By applying Maximum Marginal Relavancy, the optimal extraction of document summary is provided to reduce the sentences redudancy and rangking output.The comment which value is zero (0) that based on the comparison of summary between document posting and members comments will be classfied as "Unfeasible"and recommended to be eliminated. As the result of accuracy, blackbox and UAT testing in one of lecture class this application is success in monitoring the activity of online discussion. Hence the discussion forum as one of tool in increasing the learning process quality can be optimaized accordingly.

Keywords*: comments feasibility, cosine similarity, online discussion, Maximum Marginal Relevance, text processing, TF-IDF.*

Jurnal Sains, Teknologi dan Industri, Vol. 13, No. 2, Juni 2016, pp.151 - 159
ISSN 1693-2390 print/ISSN 2407-0939 *online*

Corresponding Author
Okfalisa,
Program Studi Teknik Informatika, Fakultas Sains dan Teknologi,
Universitas Islam Negeri Sultan Syarfi Kasim Riau,
Email: okfalisa@gmail.com

Pendahuluan

Kehidupan sosial manusia terus berkembang kearah teknologi informasi pada segala bidang, baik kegiatan interaksi sosial, belajar-mengajar, berkomunikasi maupun saling berbagi informasi. Para pengguna internet dapat berbagi informasi dan pengalaman tanpa harus saling mengenal melalui media sosial berupa *twitter, facebook, blog* dan forum diskusi *online*.

Pemanfaatan forum diskusi online terutama pada proses belajar mengajar menjadi salah satu trend pendidikan masa kini yang dapat meningkatkan optimalitas proses belajar mengajar, *sharing knowledge*, akuisisi pengetahuan, pembentukan pengetahuan dan karakter serta *learning process* baik bagi peserta maupun tenaga pengajar. Sehingga pendidikan yang bisa dilakukan dimana saja dan kapan saja serta oleh siapa saja yang terlibat didalamnya dapat dicapai.

Guna memaksimalkan pemanfaatan forum diskusi online teutama kaitannya dengan proses pembelajaran, monitoring dan *controlling* sistem perlu dilakukan. Sehingga informasi yang disampaikan oleh peserta maupun tenaga pengajar dapat mencapai hasil sesuai yang diharapkan. "Lain yang disampaikan, lain pula yang didiskusikan dalam forum tersebut". Bahkan, tidak sedikit materi yang dibicarakan menyimpang dari materi yang diberikan.Hal ini tentunya memicu terjadinyaperdebatan yang simpangsiur di dalam forum tersebut. Perdebatan yang positifdansalingmendukungtentunyaakanmembentukkarakterpesertadalammenghasilkansuatupengetahuanbaruyang meningkatkan capaian hasil proses pembelajaran.

TF-IDF sebagai teknik temu kembali informasi berdasarkan frekuensi kata atau istilah tertentu telah banyak diterapkan pada peringkasan dokumen baik yang berbahasa Indonesia[3,4,5,6] ataupun selainnya. Proses inputan ringkasan dapat diperoleh dari dokumen tunggal ataupun berganda yang akan diringkas dalam bentuk summary. Dengan menerapkan berbagai teknik pengujian apakah melalui pemberian nilai bobot pada TF-IDF ataupun perhitungan nilai F-measure terlihat bahwa TF-IDF ini memberikan nilai akurasi yang lebih baik daripada metode lainnya salah satunya adalah algoritma genetika . Faktor panjangnya ringkasan dan kemampuan optimalitasi compression juga mempengaruhi keefektivitasan perhitungan metode ini dibandingkan yang lainnya. Optimalitas compression yang diperoleh bisa mencapai 50% dengan tingkat akurasi yang linier.

Berlatarbelakang permasalahan diatas, maka penelitian ini mendiskusikan implementasi metode TF-IDF dengan nilai maximum marginal relevance dalam memonitor proses belajar mengajar pada forum diskusi online. Dengan melakukan summarization setiap dokumen postingan materi ajar yang kemudian dibandingkan dengan ringkasan dokumen komentar peserta, pembobotan TF-IDF untuk masing-masing dokumen dilakukan. Guna menentukan sejauh mana tingkat relevancy dari hasil ringkasan berganda tersebut teknik maximum marginal relevancy diukur dan ditentukan. Sehingga sistem akan menentukan apakah komentar peserta berikut layak tetap berada pada forum diskusi atau terdeteksi sebagai *spam*. Dengan menerapkan metode ini, sistem dapat memfilter setiap komunikasi yang dilakukan oleh peserta dalam forum diskusi. Jika berkaitan bisa diteruskan sebaliknya jika tidak akan di *spam*.

1. Information Retrival System

Sistem temu kembali informasi (*information retrieval system*)merupakan sistem yang dapat digunakan untuk menemukan informasi yang *relevan* dengan kebutuhan dari penggunanya secara otomatis dari suatu koleksi informasi.

Proses yang berjalan dalam sistem temu kembali informasi adalah proses *indexing subsystem*, yangmerupakan proses persiapan ulang dilakukan terhadap dokumen sehingga dokumen siap diproses,dan *searching subsystem* (*macthingsystem*) yang merupakan proses menemukan kembali informasi (dokumen) yang *relevan* terhadap *query* yang diberikan .

Adapun tahap-tahap yang terjadi pada proses *indexing*, yaitu:

1. *Tokenizing* dokumen, yaitu proses mengubah dokumen menjadi kumpulan term dengan cara menghapus semua karakter tanda baca yang terdapat pada token. Hingga pada akhirnya yang diperoleh hanya kumpulan kata-kata dari suatu teks/dokumen.
2. *Stopword removal* dokumen, yaitu kata-kata yang sering muncul dalam dokumen namun artinya tidak deskriptif dan tidak memiliki keterkaitan dengan tema tertentu. Pada bahasa Indonesia, *stopword* disebut juga sebagai kata

yang tidak penting, misalnya "di", "oleh", "pada", "sebuah", "karena" dan lain sebagainya.

3. *Stemming* dokumen, yaitu tahap penghilangan imbuhan sehingga didapatkan kata dasar dari *term-term* dokumen inputan.
4. *Term Weighting*, yaitu proses pembobotan pada setiap term (kata) yang ada didalam dokumen.

2. *Text Processing*

Text Preprocessing adalah mempersiapkan teks menjadi data yang akan mengalami proses pengolahan pada tahapan berikutnya. Tujuan dilakukan *pre-processing* adalah memilih setiap kata dari dokumen dan merubahnya menjadi kata dasar yang memiliki arti sempit dan proses teks mining akan memberikan hasil yang lebih memuaskan.

1. Pemecahan Kalimat
   Pemecahan kalimat teks menjadi kalimat-kalimat. Adapun yang menjadi pemisah kumpulan kalimat adalah tanda tanya "?", tanda titik ".", dan tanda seru "!" [8].
2. Case Folding
   Case Folding adalah proses mengubah semua huruf yang ada pada dokumen teks menjadi huruf kecil semua.
3. Filtering Kalimat
   Filtering merupakan proses penghilangan *stopword*.
4. Tokenisasi Kata
   Pemecahan kalimat menjadi kata-kata tunggal dilakukan dengan men-*scan* kalimat dengan pemisah *white space* (spasi, tab, dan *newline*)[8].
5. *Stemming*
   Stemming merupakan suatu proses mentransformasikan kata-kata yang terdapat dalam suatu dokumen ke kata-kata akarnya (*root word*) dengan menggunakan aturan-aturan tertentu. Pada penelitian ini stemming *Porter Stemmer* digunakan.

3. *TF - IDF*

Metode ini dilakukan setelah tahapan *stemming* dan *stopword removal* dilakukan. Perhitungan terhadap nilai atau bobot suatu kata *(term)* pada dokumen dilakukan ditahap ini.

Nilai IDF sebuah *term* (kata) dapat dihitung menggunakan persamaan berikut:

$$\mathrm{idf}_t = \log\left(\frac{N}{df_t}\right) \quad (1)$$

Adapun algoritma yang digunakan untuk menghitung bobot (W) masing-masing dokumen terhadap kata kunci (*query*) yaitu:

$$\mathrm{W}_{dt} = tf_{d,t} \bullet idf_t \quad (2)$$

4. Cosine Similarity

Metode *Cosine Similarity* merupakan metode yang digunakan untuk menghitung *similarity* (tingkat kesamaan) antar dua objek.

Ukuran ini memungkinkan perangkingan dokumen sesuai dengan kemiripannya (relevansi) terhadap *query*. Setelah semua dokumen dirangking, sejumlah tetap dokumen *top-scoring* dikembalikan kepada pengguna. C*osine similarity* ini dihitung berdasarkan nilai *cosinus* sudut antara dua vektor. Jika terdapat dua vektor dokumen dj dan *query* q, serta *term* diekstrak dari lokasi dokumen maka nilai *cosinus* antara dj dan q didefenisikan sebagai berikut :

$$sim(D_1, D_2) = \frac{\sum_i t_{1i} t_{2i}}{\sqrt{\sum_i t_{1i}^2} \times \sqrt{\sum_i t_{2i}^2}} \quad (3)$$

5. Maximum Marginal Relevance (MMR)

Algoritma *MMR* merupakan salah satu metode *extractive summary* yang digunakan dalam peringkasan *single* dokumen maupun multidokumen, mempunyai karakter yang efektif dan sederhana.

Peringkasan dokumen dengan tipe ekstraktif, nilai akhir diberikan pada kalimat Di dalam *MMR* dihitung dengan persamaan :

$$MMR = argmax\,[\lambda * Sim_1(D, Q) - (1 - \lambda) * \max Sim_2(D_i, D')] \quad (4)$$

Nilai parameter λ adalah 0 sampai 1 atau range 0,1. Pada saat parameter λ = 1 maka nilai MMR yang diperoleh akan cenderung relevan terhadap dokumen asal. Ketika nilai parameter λ = 0 maka nilai diperoleh cenderung relevan terhadap kalimat yang telah diekstrak sebelumnya yang akan dibandingkan. Untuk peringkasan yang baik pada dokumen pendek digunakan nilai parameter λ = 0.7 atau λ = 0.8.

6. Spam

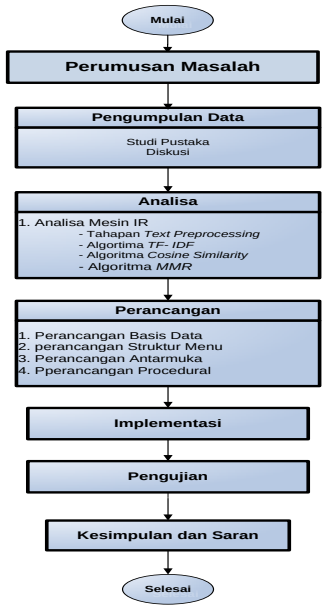
Spam adalah penyalahgunaan perangkat elektronik untuk mengirimkan pesan secara terus menerus tanpa dikehendaki oleh si penerima. Pada penelitian ini digunakan untuk mengklasifikasi hasil filter diskusi.

Metode Penelitian

Penelitian ini dilakukan dalam serangkaian proses seperti yang dijelaskan pada Gambar 3.1 meliputi: Perumusan Masalah, Pengumpulan Data, Analisa, Perancangan, Implementasi dan Pengujian. Terakhir adalah Kesimpulan dan Saran.

Pada tahap pengumpulan data, semua informasi terkait baik terhadap kasus maupun metode yang digunakan dilakukan melalui studi pustaka dari berbagai referensi baik berbentuk buku cetak, jurnal, proceeding, e-book, websites ataupun artikel lainnya. Diskusi dengan beberapa orang tenaga pengajar yang memanfaatkan forum diskusi online dalam proses pembelajaran juga dilakukan guna memahami permasalahan dengan lebih baik.

Tahap Analisa dilakukan dengan menganalisis metode mesin *Information Retrieval* yang digunakan meliputi analisis Text Preprocessing untuk koleksi dokumen (*corpus*) yaitu: Pemecahan kalimat, *Case folding*, *Filtering* kalimat, Tokenisasi kata dan *Stemming*. Selanjutnya analisis Algoritma *TF-IDF* dilakukan melalui serangkaian aktivitas untuk: Menghitung banyak kata dalam kalimat, Menghitung banyak kata dalam dokumen, Menghitung nilai *inverse document frequency*, Menghitungnilai bobot katadan Menghitung nilai akumulatif W untuk setiap kalimat.



Gambar 1. Skema alur penelitian

Analisis Algoritma *Cosine Similarity* digunakan untuk menghitung kesamaan antar 2 buah objek yang dinyatakan dalam 2 buah vektor dengan menggunakan *keyword* dari sebuah dokumen sebagai ukuran. Terakhir Algoritma *Maximum Marginal Relevance*(*MMR*) dianalisis untuk meringkas dokumen tunggal atau multi dokumen dengan menghitung kesamaan (*similarity*) antara bagian teks.

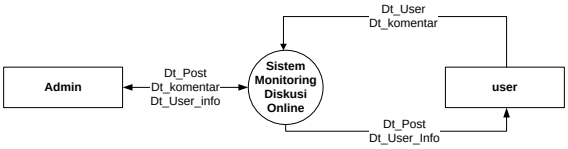
Proses pembangunan perangkat lunak pada aplikasi sistem ini menerapkan model Waterfall dengan menggunakan Context Diagram dan Data Flow Diagram sebagai tools yang menggambarkan hasil analisis model. Tahap Perancangan sistem dilakukan dengan melengkapi rancangan basis data, struktur menu, antarmuka dan procedural sistem. Implementasi dengan menerapkan bahasa pemograman PHP dan database MySQL dilakukan. Guna membuktikan keberhasilan hipotesa dan optimalitas capaian sistem, maka pengujian blackbox, akurasi dan User Accaptance Test dilakukan. Terakhir kesimpulan dan saran pengembangan sistem diperoleh sebagai hasil akhir penelitian.

Hasil dan Pembahasan

1. Analisa Sistem Baru

Sistem yang akan dibangun ini merupakan sistem peringkasan dokumen. Dokumen yang sudah diringkas diukur nilai bobot relevansinya dengan komentar yang diberikan oleh peserta, sehingga dapat ditentukan kelayakan komentar tersebut untuk tetap dipertahankan atau tidak pada forum diskusi. Komentar yang tidak layak, adalah komentar yang bernilai nol (0) dari nilai keseluruhan perbandingan postingan yang telah diringkas dengan komentar. Komentar yang tidak layak tersebut akan dieliminasi dari *front page* diskusi.

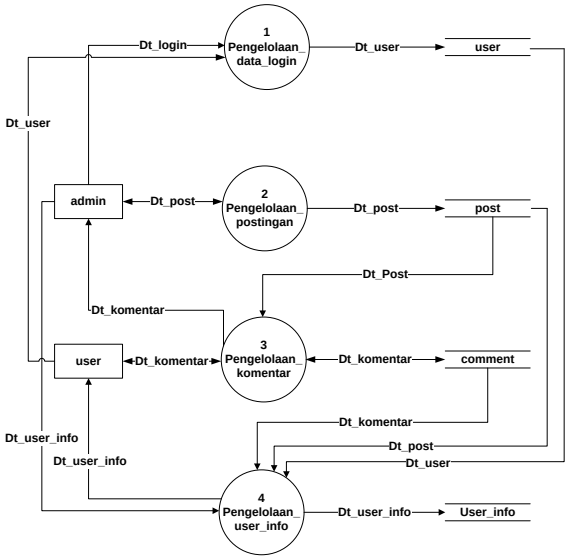
2. *Context Diagram (CD)*



Gambar 2. *Context diagram*

Gambar 2 menjelaskan gambaran sistem secara keseluruhan. Sistem memiliki 2 aktor utama yaitu admin yang bisa juga berfungsi sebagai tenaga pengajar dan user sebagai peserta yang ikut berpartisipasi pada forum diskusi. Admin mampu memposting materi ajar pada blog atau websites, melihat dan mengupdate informasi dan komentar pada forum tersebut.

3. Data *Flow* Diagram



Gambar 3. Data *flow* diagram

Informasi lengkap dari CD diuraikan pada Gambar 4.2 yang terdiri dari 4 proses yaitu Pengelolaan data_login, Pengelolaan_Postingan, Pengelolaan_komentar dan Pengelolaan_user_info. Database yang terlibat terdiri dari database user, post, comment dan user_info.

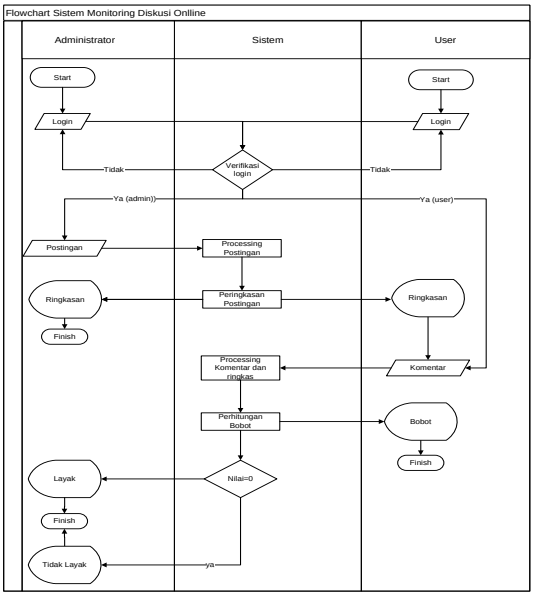
4. *Flowchart* Sistem *Monitoring* Diskusi *Online*

Flow proses sistem dapat dilihat pada Gambar 4.

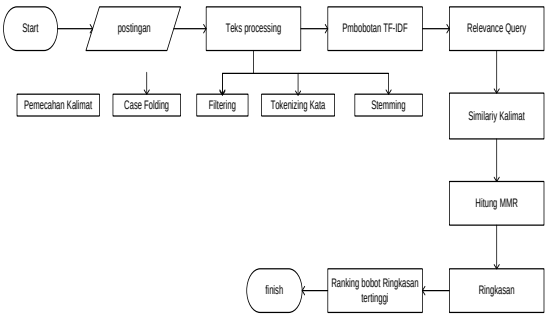
5. Proses Peringkasan Dokumen *Postingan*

Proses pada peringkasan dokumen *postingan* dapat dilihat pada Gambar 5 yang dilakukan sesuai dengan tahapan di metodologi. Pada proses teks processing flowchart yang dilakukan dapat dilihat pada Gambar 6.

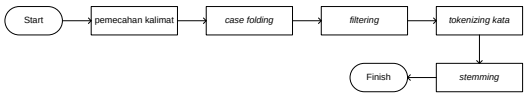
Analisa Proses sesuai dengan Studi Kasus dijelaskan dibawah ini. Gambar 4.6 adalah contoh dokumen postingan materi ajar pada forum diskusi online dengan judul "Stopword Bahasa Indonesia".



Gambar 4. *Flowchart* sistem *monitoring* diskusi *online*



Gambar 5. *Flowchart* peringkasan dokumen *postingan*



Gambar 6. *Flowchart teks processing*

a. Analisa Pemecahan Kalimat

Kalimat pada dokumen dipecahkan menjadi 9 kalimat sesuai dengan aturan pemecahan. Hasilnya dapat dilihat pada Gambar 8.

b. Case Folding

Proses yang dilakukanya itu mengubah teks menjadi huruf kecil, menghilangkan angka dan tanda baca maupun simbol-simbol karena dianggap sebagai delimeter, sistem hanya menerima karakte rhuruf saja seperti pada Gambar 9.

c. *Filtering* Kata

Membuang kata yang kurang penting dengan Algoritma*stopword*. Hasil filtering diperoleh pada Gambar 10.

Stopword adalah kata umum (common words) yang biasanya muncul dalam jumlah besar dan tidak memiliki makna. Stopword umumnya dimanfaatkan dalam task information retrieval, termasuk oleh google. Contoh stopword untuk bahasa inggris diantaranya of dan the. Sedangkan untuk bahasa indonesia diantaranya yang, di dan ke.

Saya pernah membuat stopword bahasa Indonesia untuk tugas salah satu matakuliah. Tujuannya waktu bukan untuk information retrieval, tapi untuk klasifikasi. Saya gunakan stopword untuk mengurangi jumlah kata yang harus diproses.

Saya membuat daftar stopword dengan cara mengumpulkan kata paling banyak muncul pada korpus (saya menggunakan berita kompas), setelah diurutkan kemudian diperiksa secara manual satuper satu. Karena daftar itu dibuat secara manual dan untuk task klasifikasi, ada beberapa kata yang mungkin diperdebatkan, jadi silahkan edit sesuai kebutuhan.

Gambar 7. Contoh berita

No	Kalimat
Q	Stopword untuk bahasa Indonesia
1	Stopwords adalah kata umum (common words) yang biasanya muncul dalam jumlah besar dan tidak memiliki makna
2	Stopword umumnya dimanfaatkan dalam task information retrieval, termasuk oleh google
3	Contoh stopword untuk bahasa inggris diantaranya of dan the
4	Sedangkan untuk bahasa indonesia diantaranya yang di dan ke
5	Saya pernah membuat stopword bahasa Indonesia untuk tugas salah satu matakuliah
6	Tujuannya waktu bukan untuk information retrieval, tapi untuk klasifikasi.
7	Saya gunakan stopword untuk mengurangi jumlah kata yang harus diproses
8	Saya membuat daftar stopword dengan cara mengumpulkan kata paling banyak muncul pada korpus (saya menggunakan berita kompas), setelah diurutkan kemudian diperiksa secara manual satu per satu
9	Karena daftar itu dibuat secara manual dan untuk task klasifikasi, ada beberapa kata yang mungkin diperdebatkan, jadi silahkan edit sesuai kebutuhan

Gambar 8.Analisa pemecahan kalimat

No	Kalimat
Q	stopword untuk bahasa Indonesia
1	stopword adalah kata umum common words yang biasanya muncul dalam jumlah besar dan tidak memiliki makna
2	stopword umumnya dimanfaatkan dalam task information retrieval termasuk oleh google
3	contoh stopword untuk bahasa inggris diantaranya of the
4	suntuk bahasa indonesia diantaranya yang di ke
5	saya pernah membuat stop word bahasa indonesia untuk tugas salah satu matakuliah
6	tujuannya waktu bukan untuk information retrieval tapi untuk klasifikasi
7	saya gunakan stopword untuk mengurangi jumlah kata yang harus diproses
8	saya membuat daftar stopword dengan cara mengumpulkan kata paling banyak muncul pada korpus saya menggunakan berita kompas setelah diurutkan kemudian diperiksa secara manual satu per satu
9	karena daftar itu dibuat secara manual dan untuk task klasifikasi ada beberapa kata yang mungkin diperdebatkan jadi silahkan edit sesuai kebutuhan

Gambar 9. Hasil *case folding*

No	Kalimat
Q	stopword bahasa Indonesia
1	stopword commonwords makna
2	stopword umumnya dimanfaatkan task informationretrieval google
3	contoh stopword bahasa inggris diantaranya
4	bahasa indonesia diantaranya
5	stopword bahasa indonesia tugas matakuliah
6	tujuannya information retrieval klasifikasi
7	gunakan stopword mengurangi diproses
8	daftar stopword mengumpulkan korpus berita kompas diurutkan diperiksa manual per
9	daftar dibuat manual task klasifikasi diperdebatkan silahkan edit

Gambar 10. Hasil filtering kata

d. *Tokenizing* Kata

Pada tahap ini proses yang dilakukanya itu pemotongan string kalimat-kalimat hasil *filtering*. Hasilnya dapat dilihat pada Gambar 4.10.

Kata	Kata	Kata	Kata
stopword	google	mengumpulkan	edit
diperdebatkan	contoh	berita	umumnya
bahasa	inggris	kompas	diantaranya
indonesia	tugas	diurutkan	per
common	matakuliah	diperiksa	diproses
makna	tujuannya	manual	retrieval
dimanfaatkan	klasifikasi	korpus	daftar
task	gunakan	dibuat	silahkan
information	mengurangi	words	-

Gambar 11. Tokenizing kata

e. *Stemming* dengan Algoritma Nazief dan Andriani

Stemming untuk mendapatkan kata dasar dilakukan dengan menggunakan algoritma Nazief dan

Andriani. Hasil yang diperoleh terlihat pada Gambar 12.

Kata	Kata	Kata	Kata
antara	google	kurang	silah
bahasa	guna	makna	stopword
berita	information	manfaat	task
buat	indonesia	manual	tugas
common	inggris	matakuliah	tujuan
contoh	klasifikasi	per	umum
daftar	kompas	periksa	urut
debat	korpus	proses	words
edit	kumpul	retrieval	-

Gambar 12. *Stemming Nazief dan Andriani*

f. Analisa Pembobotan *TF-IDF*

Pembobotan dilakukan dengan menggunakan proses TF dan IDF dan perangkingan. Dengan menggunakan metode *cosine similarity* nilai bobot *relevance query* dan bobot *similarity* diperoleh. Hasil yang diperoleh dapat dilihat pada Gambar 13.

	D1	D2	D3	D4	D5	D6	D7	D8	D9
D1	1	0.009	0.012	0	0.012	0	0.011	0.007	0
D2	0.009	1	0.010	0	0.010	0.287	0.009	0.006	0.092
D3	0.012	0.010	1	0.398	0.103	0	0.012	0.007	0
D4	0	0	0.398	1	0	0	0	0	0
D5	0.012	0.010	0.103	0.398	1	0	0.012	0.007	0
D6	0	0.287	0	0	0	1	0	0	0.125
D7	0.011	0.009	0.012	0	0.012	0	1	0.007	0
D8	0.007	0.006	0.007	0	0.007	0	0.007	1	0.137
D9	0	0.092	0	0	0	0.125	0	0.137	1

Gambar 13. Hasil *Similarity* antar kalimat pada *Postingan*

g. *Maximum Marginal Relevance (MMR)*

Pada penelitian ini MMR menggunakan nilai parameter λ = 0.7[3]. Hasil iterasi dan bobot MMR dapat dilihat pada Gambar 14 dan 15.

iterasi ke	D1	D2	D3	D4	D5	D6	D7	D8
1	0.016	0.013	0.138	**0.532**	0.366	0	0.016	0.009
2	0.016	0.013	0.019	-	**0.247**	0	0.016	0.009
3	0.016	0.013	**0.019**	-	-	0	0.016	0.009
4	**0.012**	0.010	-	-	-	0	0.012	0.008
5	-	0.010	-	-	-	0	**0.012**	0.008
6	-	**0.010**	-	-	-	0	-	0.008
7	-	-	-	-	-	-0.086	-	**0.008**

Gambar 14. Hasil iterasi MMR

Iterasi ke	Kalimat	Bobot ArgMax MMR
MMRMAX1	D4	**0.532**
MMRMAX2	D5	**0.247**
MMRMAX3	D3	**0.019**
MMRMAX4	D1	**0.012**
MMRMAX5	D7	**0.012**
MMRMAX6	D2	**0.010**
MMRMAX7	D8	**0.008**

Gambar 15. Hasil bobot *MMR* maksimum iterasi *MMR*

Output hasil ringkasan akhir dapat dilihat pada Gambar 16. Proses TF-IDF yang dilakukan pada dokumen postingan juga diberlakukan pada dokumen komentar peserta. Selanjutnya perbandingan antara dokumen tersebut dilakukan untuk menentukan kelayakan komentar.

Sedangkan untuk bahasa indonesia diantaranya "yang" "di" "ke". Saya pernah membuat stopword bahasa Indonesia untuk tugas salah satu matakuliah. Contoh stopword untuk bahasa inggris diantaranya "of" "the". Stopwords adalah kata umum (common words) yang biasanya muncul dalam jumlah besar dan tidak memiliki makna. Saya gunakan stopword untuk mengurangi jumlah kata yang harus diproses. Stopword umumnya dimanfaatkan dalam task information retrieval, termasuk oleh google. Saya membuat daftar stopword dengan cara mengumpulkan kata paling banyak muncul pada korpus (saya menggunakan berita kompas), setelah diurutkan kemudian diperiksa secara manual satu per satu.

Gambar 16. Hasil ringkasan

h. Proses Perbandingan Komentar

Berdasarkan hasil nilai iterasi bobot *MMR* pada D4, D5, D3, D1, D7, D2 dan D8. Maka kalimat yang akan dibandingkan dengan komentar adalah kalimat 4, 5, 3, 1, 7, 2 dan 8 seperti pada Gambar 18.

Komentar yang tidak layak adalah komentar yang bernilai 0 dan akan dikatagorikan sebagai Spam seperti pada Gambar 20.

K1	K2	K3	K4	K5	K6	K7	K8
1.368	0.157	0	0.762	0.157	1.368	0.157	1.883

Gambar 17 Hasil bobot komentar

Dari hasil pembobotan komentar diatas, maka komentar ke-3 tidak berkaitan dengan *postingan*yang di*post* oleh *admin* atau tenaga pengajar dan bisa di eliminasidari forum.

6. Perancangan Struktur Menu

Rancangan struktur menu sistem dapat dilihat pada Gambar 19. Sistem ini memiliki 4 fungsi utama yaitu Home sebagai informasi umum sistem, menu kelola Postingan, menu kelola komentar dan menu log out.

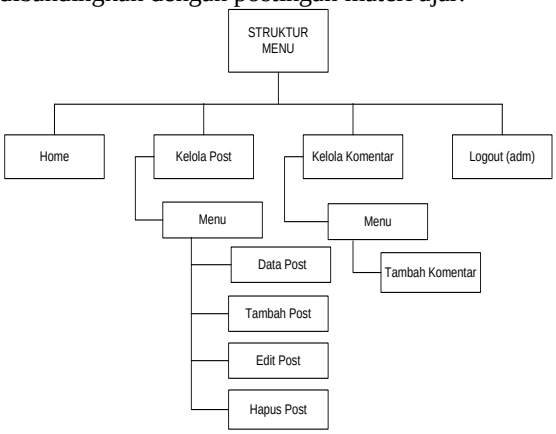
1. Bapak minta daftar stopwordnya untuk tugas semester membuat program menghitung kata. Saya sudah coba download, tapi filenya rusak. Bila bapak berkenan membantu saya mohon dikirim ke email saya. Terima kasih banyak bapak.
2. Bagaimana cara melihat satu kata itu termasuk stopword atau bukan? Kalau kata mengetahui dan mempunyai (dan kata kata berimbuhan lainnya) bukannya bisa distemming merubah kata menjadi kata dasar lagi?
3. Terima kasih bang. Memang lagi butuh, setiap hari saya cek dan akhirnya hari ini bisa saya download.
4. Maaf pak, itu daftar stopword secara keseluruhan pak? Tapi kenapa masih ada kata berimbuhan di daftar stopwordnya? Mohon bantuannya pak, buat skripsi saya tentang plagiat.
5. Pak, izin mendownload buat referensi TA. Sekalian mau tanya pak, dataset saya menggunakan artikel artikel tenang review satu hal (misal, hape, komputer, film). Kira kira dengan stopword yang bapak susun ini, bisa mengcover dataset saya pak ? Terima kasih pak.
6. Terima kasih pak, izin download buat TA pak. Saya ingin bertanya, bapak membuat daftar stopword acuannya apa ya pak? Misalnya apakah harus memiliki kemunculan dalam bilangan tertentu atau yang jadi stopword kata kata bilangan, penghubung dan laninya.
7. Pak mau tanya, sudah pernah menggunakan library lucene untuk tokenisasi, stopword dan stemming? Kalau sudah pernah bagaimana algoritmanya untuk memanggil fungsi fungsinya dijava.
8. Pak, saya ingin download stopword list bahasa indonesia untuk keperltan TA. Dataset yang saya gunakan adalah abstrak paper atau karya ilmah. Apakah sesuai jika menggunakan stopword list dari bapak?

Gambar 18. Contoh komentar

7. Implementasi dan Pengujian

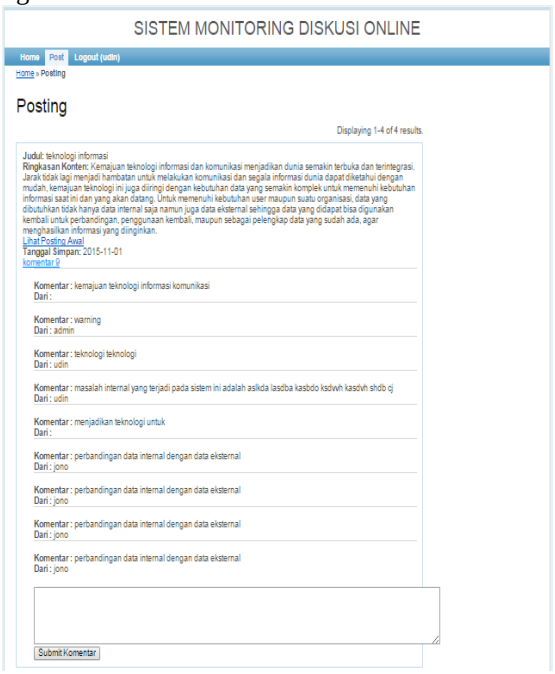
Beberapa tampilan sistem sesuai dengan proses filtering TF-IDF dapat dilihat pada Gambar 19 dan 20.

Gambar 20 menjelaskan tampilan komentar user terhadap materi pembelajaran yang diberikan oleh admin atau tenaga pengajar. Sementara Gambar 21 menampilkan hasil perhitungan bobot komentar dibandingkan dengan postingan materi ajar.



Gambar 19. Struktur menu sistem

Dari hasil pengujian yang dilakukan diperoleh nilai akurasi yang cukup baik dari proses perhitungan sistem yang dibandingkan dengan perhitungan manual (98%) untuk 35 kali percobaan. Pengujian blackbox terhadap setiap fungsi aplikasi yang digunakan 100% berjalan dengan baik sesuai dengan tahapan dan metode yang diterapkan. Hasil pengujian User Accaptance Test (UAT) dari 10 orang responden (2 orang admin atau instruktur dan 8 orang mahasiswa/peserta diskusi) yang menggunakan sistem menyatakan bahwa sistem tersebut memiliki penerimaan yang baik (99%), user friendly dan interface yang menarik dan mudah digunakan.

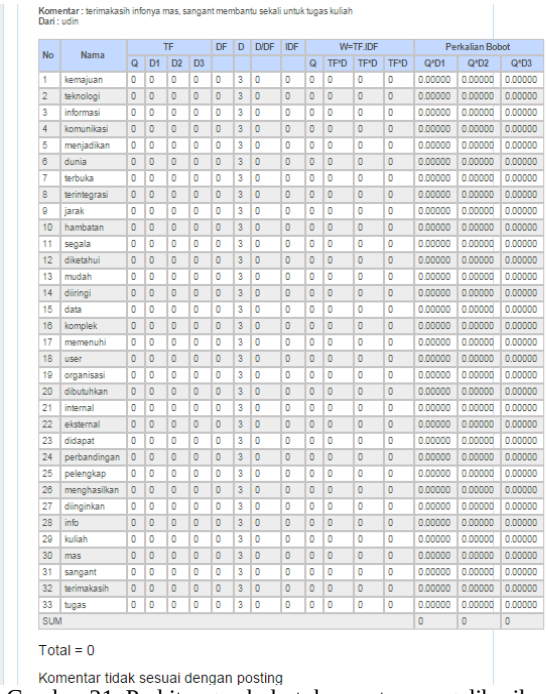


Gambar 20. Tampilan halaman komentar *user*

Kesimpulan

Dari hasil penelitian diperoleh beberapa kesimpulan diantaranya adalah: Metode *TF-IDF*, *Cosine Similarity* dan *MMR* berhasil diterapkan pada peringkasan dokumen untuk monitoring diskusi *online*. Penerapan nilai MMR dalam penentuan kelayakan komentar untuk dapat dipertahankan pada forum diskusi online telah berhasil diterapkan. Pengujian blackbox dan UAT telah memberikan hasil yang dapat menyimpulkan bahwa sistem ini layak dan dapat digunakan untuk membantu proses monitoring diskusi online.

Saran pengembanganaplikasiselanjutnya adalah Aplikasi sistem monitoring diskusi online ini dapat dijalankan atau digabungkan pada aplikasi pembelajaran elektronik lainnya baik melalui website, blog, twitter, e-learning, facebook dan sebagainya. Adanya dinamika konten stopwatch yang digunakan sehingga materi ajar yang bisa dianalisis menjadi bervariasi. Sistem dapat memberikan komentar peringatan terhadap komentar yang mendekati ketidaklayakan sebagai *alert* sebelum komentar yang tidak relevan di spam.



Komentar : terimakasih infonya mas, sangant membantu sekali untuk tugas kuliah
Dari : udin

No	Nama	TF				DF	D	D/DF	IDF	W=TF.IDF			Perkalian Bobot		
		Q	D1	D2	D3					Q	TF*D	TF*D	Q*D1	Q*D2	Q*D3
1	kemajuan	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
2	teknologi	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
3	informasi	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
4	komunikasi	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
5	menjadikan	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
6	dunia	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
7	terbuka	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
8	terintegrasi	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
9	jarak	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
10	hambatan	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
11	segala	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
12	diketahui	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
13	mudah	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
14	diiringi	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
15	data	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
16	komplek	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
17	memenuhi	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
18	user	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
19	organisasi	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
20	dibutuhkan	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
21	internal	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
22	eksternal	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
23	didapat	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
24	perbandingan	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
25	pelengkap	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
26	menghasilkan	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
27	diinginkan	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
28	info	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
29	kuliah	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
30	mas	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
31	sangant	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
32	terimakasih	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
33	tugas	0	0	0	0	0	3	0	0	0	0	0	0.00000	0.00000	0.00000
SUM													0	0	0

Total = 0

Komentar tidak sesuai dengan posting

Gambar 21. Perhitungan bobot komentar yang diberikan *user*

Ucapan TerimaKasih

Ucapan terima kasih peneliti tujukan kepada civitas akademika di Jurusan Teknik Informatika dan Fakultas Sains dan Teknologi UIN Suska Riau yang telah memberikan kontribusi dalam penelitian ini.

Daftar Pustaka

[1] Aristoteles (2013), "Penerapan Algoritma Genetika pada Peringkasan Teks Dokumen Bahasa Indonesia", Prosiding Semirata 2013, vol. 1, no. 1.

[2] Harjanto,Dhony Syafe'i, Endah,Sukmawati Nur, Bahtiar,Nurdin. (2012).*SistemTemu KembaliInformasi pada Dokumen Teks Menggunakan MetodeTF IDF*. Fakultas Sains dan Matematika, UniversitasDiponegoro

[3] Karmayasa, Oka. &Mahendra, Ida Bagus. (2012). Implementasi Vector Space Model dan Beberapa Notasi Metode Term Frequency Inverse Document Frequency (TF-IDF) Pada Sistem Temu Kembali Informasi. Bali : Program Studi Teknik Informatika Jurusan Ilmu Komputer Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Udayana.

[4] Lui, A. K.-F., Li, S. C., & Choy, S. O. (2007). An Evaluation of Automatic Text Categorization in Online Discussion Analysis. *Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007)*, (pp. 205 - 209).

[5] Mandala,Rila, Setiawan,Hendra. (2002). *Peningkatan Preformansi dengan Perluasan Query Secara Otomatis*. Departement Teknik Informatika, Institut Teknologi Bandung

[6] Mulyana I, Ramadona S, Herfina (2012). *Penerapan Terms Frequency-Inverse Document Frequency PadaSistem Ringkasan Teks Otomatis Dokumen Tunggal Berbahasa Indonesia*. Jurnal.

[7] Mustaqhfiri, Muchammad (2011). *Peringkasan Teks Otomatis BeritaBerbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance*. Fakultas Sains dan Teknologi, Jurusan Teknik Informatika, Universitas Islam Negeri Maulana Malik Ibrahim. Malang.

[8] Pradipa, Enggar (2013). *Klasifikasi Pola Konten E-mail dengan Menggunakan Jaringan Syaraf Tiruan Metode Back Propogation untuk Pengecekan Spam E-Mail dengan Acuan DMC.* Fakultas Ilmu Komputer. Universitas Dian Nuswantoro. Semarang.

[9] Pradnyana. (2012).*Perancangan dan Implementasi Automated DocumentIntegration dengan Menggunakan Algoritma Complete Linkage Aglomerative Hierarchial Clustering.*Teknik Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,Universitas Udayana.

[10] Septiawan,Danny, Suprayogi,Dwi Aries (2015). *Klasifikasi Iklan pada Online Shop dengan Metode Naive Bayes.* Teknik Informatika, Program Teknologi dan Ilmu Komputer, Universitas Brawijaya, Malang.

[11] Tala, Fadillah Z.(2003).*A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. Institute for Logic*, Language and Computation Universeit Van Amsterdam.

[12] Tata, S., Patel,J.M., Science, C., Arbor, A. (2007) Estimating the Selectivity of TF-IDF based Cosine Similarity Predicates, 36,7-12.

