Penerapan Algoritma Peningkatan Porter Stemmer Dan Likelihood Untuk Identifikasi Topik Pada Artikel Berita Berbahasa Indonesia

Devi Andriyani, Imam Mukhlash, Alvida Mustika Rukmi
Matematika, Fakultas MIPA, Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
e-mail: devi12@mhs.matematika.its.ac.id, imamm@matematika.its.ac.id, alvida@yahoo.com

***Abstrak*— Setiap informasi yang disajikan dalam suatu berita memiliki tema pembicaran yang beragam sehingga tidak mungkin semua informasi tersebut bisa dicerna secara bersamaan, melainkan harus dikelompokkan berdasarkan relevansi topik dari berita tersebut. Pengelompokan tersebut dapat mempermudah pembaca untuk memilih informasi yang penting sesuai dengan topik yang ingin dibaca. Berkaitan dengan pengelompokan berita, berita memiliki karakteristik yang berbeda dengan informasi yang lain sehingga diperlukan suatu algoritma khusus yang mampu menangani penemuan topik dan klasifikasi menggunakan data training pada suatu berita. Pada penelitian ini akan diterapkan algoritma peningkatan Porter Stemmer pada proses stemming (pembentukan kata dasar) dan metode Likelihood untuk klasifikasi berita berdasarkan kategori serta identifikasi topik.**

Berdasarkan hasil pengujian menggunakan 900 data training dan 90 data uji didapatkan akurasi yang cukup tinggi, yaitu 95,56 % untuk klasifikasi kategori dan 97,78 % untuk identifikasi topik.

Kata kunci : berita, algoritma Peningkatan Porter Stemmer, likelihood, klasifikasi kategori, identifikasi topik.

I. PENDAHULUAN

Seiring dengan perkembangan teknologi yang pesat semakin meningkat pula penyebaran informasi secara online seperti halnya berita atau artikel yang mudah sekali kita jumpai pada berbagai situs. Sekumpulan informasi tersebut tentunya memiliki tema pembicaran yang beragam sehingga tidak mungkin semua informasi yang disajikan bisa dicerna secara bersamaan, melainkan harus dikelompokkan berdasarkan relevansi topik dari berita tersebut. Pengelompokan tersebut dapat mempermudah pembaca untuk memilih informasi yang paling penting sesuai dengan topik yang ingin dibaca.

Informasi dalam berita mempunyai karakteristik yang berbeda dengan koleksi dokumen lainnya yaitu aliran dinamis berupa dokumen – dokumen baru yang mungkin saja memiliki informasi yang tidak pernah ada pada dokumen sebelumnya. Maka untuk melakukan klasifikasi topik dibutuhkan sebuah algoritma khusus yang mampu menangani penemuan topik, dan klasifikasi menggunakan data training[1]. Proses identifikasi topik berita nantinya akan dilakukan pra-proses yaitu training dokumen yang terdiri dari proses filtering, stopword removal, stemming dan weighting.

Pada [1] telah dilakukan penelitian untuk identifikasi topik dan kategori berita berbahasa Inggris menggunakan perhitungan likelihood. Sedangkan pada [2] dilakukan penelitian sejenis yaitu identifikasi topik dan kategori terhadap berita Bahasa Indonesia menggunakan perhitungan likelihood serta penggunaan algoritma Confix Stripping Stemmer untuk pembentukan kata dasar (stemming). Meskipun running time yang diperlukan untuk identifikasi topik cukup lama tapi nilai precision yang dihasilkan cukup tinggi yaitu 97,26 %.

Berdasarkan hal di atas, pada Tugas Akhir ini diajukan pembuatan aplikasi identifikasi topik berita Bahasa Indonesia yang pada prosesnya akan digunakan algoritma lain dimana algoritma tersebut adalah algoritma peningkatan Porter Stemmer yang dimodifikasi oleh Putu Bagus Susastra Wiguna dan Bimo Sunarfri Hantono [5] sehingga algoritma ini memiliki performa yang lebih baik dalam hal akurasi untuk proses stemming. Dengan kemampuan performa yang baik maka metode ini akan diterapkan pada proses stemming identifikasi topik berita sebagai uji kinerja aplikasi yang nantinya akan dibandingkan dengan penelitian terdahulu. Dengan demikian, aplikasi ini diharapkan dapat menunjukkan kinerja yang lebih baik terutama dalam keakuratan pengidentifikasian topik berita sehingga keberhasilan kinerjanya dapat menjadi media penunjang dalam mempermudah pemilihan informasi berita berdasarkan topik bagi pengguna.

II. DASAR TEORI

A. Corpus

Corpus merupakan sekumpulan teks terstruktur. Secara lebih spesifik merupakan teks berita hasil pengunduhan dari situs berita online yang disimpan dalam format file teks tertentu dan memiliki keterkaitan kategori antar corpusnya. Corpus yang akan digunakan pada tugas akhir ini adalah corpus dengan eksistensi .news supaya memudahkan program mengenali corpus saat akan diproses[2]. Bila corpus tetap disimpan dalam ekstensi .txt, maka terdapat kemungkinan besar dalam sebuah folder terdapat file – file lain yang tidak berhubungan yang menggunakan ekstensi yang sama, sehingga akan ikut terproses dan mengganggu jalannya proses aplikasi.

Secara sederhana, corpus secara keseluruhan adalah hasil pengunduhan berita pada situs dengan menghilangkan atribut – atribut html maupun php pada halaman tersebut yang sudah terstruktur mengikuti format yang telah dijelaskan pada Gambar 3.1.

Pada corpus data training tidak dituliskan kategori yang telah ditetapkan oleh situs kompas karena terdapat beberapa kategori yang namanya diubah menjadi nama yang lebih umum sehingga berbeda dengan nama kategori yang dituliskan pada www.kompas.com. Seluruh berita disimpan dalam folder menurut nama kategorinya. Perbedaan signifikan yang terdapat antara data training dengan data uji adalah pada atribut topik. Data uji yang mengalami pembentukan corpus juga, tidak memiliki atribut topik karena diasumsikan atribut tersebut akan menjadi hasil identifikasi pada proses identifikasi topik yang dilakukan program.

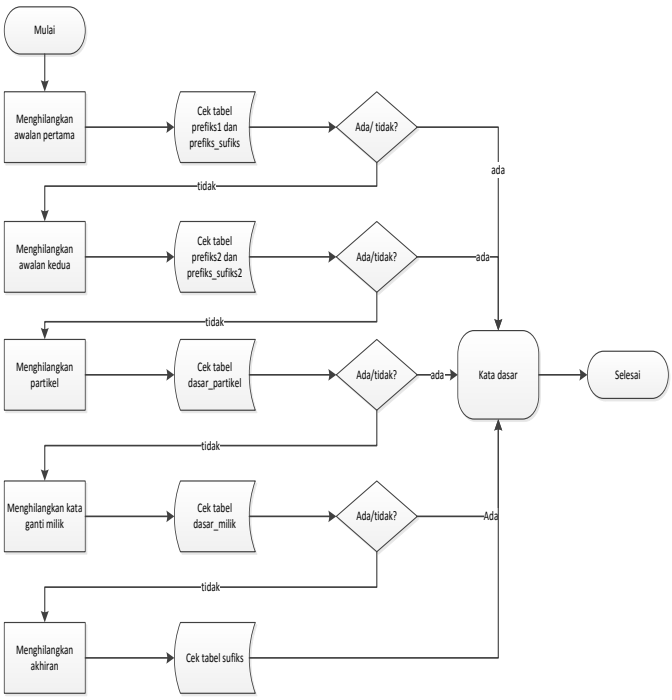
B. Metode Peningkatan Porter Stemmer

Tidak semua proses pembentukan kata dari kata dasar bisa diselesaikan dengan satu tingkat morfologi. Contoh pembentukan kata dengan penambahan imbuhan pada kata dasar dengan satu tingkat morfologi adalah “mem”+”baca” menjadi “membaca”, “men”+”cari” menjadi “mencari”.

Penambahan imbuhan pada kata dasar untuk membentuk kata baru dengan mengubah fonem dari kata dasar tidak bisa diselesaikan dengan satu tingkat morfologi. Contoh kata yang tidak bisa diselesaikan dengan satu tingkat morfologi adalah kata “memutar” berasal dari kata dasar “putar” yang mendapat imbuhan “men-”. Untuk menyelesaikan masalah ini maka diperlukan 2 tingkat morfologi untuk menyelesaikan masalah ini.

Secara umum proses stemming dibagi menjadi 5 bagian yaitu: menghilangkan awalan pertama (“meng-”, “peng-”, “mem-”, “pem-”, “meny-”, “peny-”, “men-”, “pen-” dan lainlain), menghilangkan awalan kedua (“ber-”, “per-”, “ter-”, “se-”, “pel-”, dan lain-lain), menghilangkan partikel (“-kah”, “-lah”, “-tah”, “-tah”), menghilangkan kata ganti milik (“- ku”, “-mu”, “-nya”), menghilangkan akhiran (“-kan”, “-an”, “-i”, “-isme”, “-isasi”, “-onal”). Proses stemming secara umum yang dilakukan pada penelitian ini dapat dilihat pada gambar 2.1 [5]

Untuk menunjang proses stemming dapat dilakukan dengan baik maka diperlukan database kata yang terdiri dari 7 tabel pada database yang menjadi kamus kata-kata pengecualian untuk tiap prosesnya. Table yang digunakan pada database adalah tabel dsr_milik, tabel dsr_partikel, tabel dsr_prefiks1, tabel dsr_prefiks1_sufiks1, tabel dsr_prefiks2, tabel dsr_prefiks21, tabel dsr_sufiks. Contoh Tabel dsr_prefiks1 digunakan untuk menyimpan kata dasar yang memiliki fonem awalan pertama. Tabel ini berguna agar fonem awalan pertama pada kata tersebut tidak dihilangkan karena merupakan bagian dari kata dasar.Tabel kedua adalah tabel dsr_prefiks1_sufiks1. Tabel ini digunakan untuk menyimpan kata-kata yang harus diproses 2 tingkat morfologi untuk kata dasar yg berawalan huruf “k” dan “p”. Begitu seterusnya untuk tabel yang lain.



Gambar 2.1 Langkah-langkah Algoritma Peningkatan Porter [5]

Untuk menunjang proses stemming dapat dilakukan dengan baik maka diperlukan database kata yang terdiri dari 7 tabel yang menjadi kamus kata-kata pengecualian untuk tiap prosesnya. 7 tabel yang digunakan adalah tabel dsr_milik, tabel dsr_partikel, tabel dsr_prefiks1, tabel dsr_prefiks1_sufiks1, tabel dsr_prefiks2, tabel dsr_prefiks21, tabel dsr_sufiks. Contoh Tabel dsr_prefiks1 digunakan untuk menyimpan kata dasar yang memiliki fonem awalan pertama. Tabel ini berguna agar fonem awalan pertama pada kata tersebut tidak dihilangkan karena merupakan bagian dari kata dasar.Tabel kedua adalah tabel dsr_prefiks1_sufiks1. Tabel ini digunakan untuk menyimpan kata-kata yang harus diproses 2 tingkat morfologi untuk kata dasar yg berawalan huruf “k” dan “p”. Begitu seterusnya untuk tabel yang lain.

C. Vector Space Model

Salah satu model matematika yang digunakan pada sistem temu-kembali informasi untuk menentukan bahwa sebuah dokumen itu relevan terhadap sebuah informasi adalah Vector Space Model (VSM). Model ini akan menghitung derajat kesamaan antara setiap dokumen yang disimpan di dalam sistem dengan query yang diberikan oleh pengguna. Model ini pertama kali diperkenalkan oleh Salton (1989)[7].

Vector space model merupakan salah satu pendekatan yang paling umum digunakan untuk merepresentasikan model teks digital. Setiap dokumen dj akan direpresentasikan menjadi vector [4].

$$dj = (w1j , w2j , . . . , wij) \quad (1)$$

dimana wij adalah bobot term ke- i pada dokumen j yang bersangkutan.

D. Metode TF-IDF

Baeza-Yates dan Ribeiro-Neto (1999), menyebutkan bahwa pembobotan (tf·idf) terdiri dari dua faktor, yaitu:

1. *tf (term frequency)*

tf adalah frekuensi kemunculan suatu istilah ki di dalam sebuah dokumen dj dibandingkan dengan frekuensi istilah kl yang sering muncul pada dokumen itu. Jika dimasukan dalam rumus matematika didapatkan:

$$tf_{ij} = \frac{freq_{ij}}{max_l freq_{lj}} \qquad (2)$$

2. *idf (inverse document frequency)*

idf adalah frekuensi kemunculan suatu istilah ki di dalam seluruh dokumen. Penggunaan faktor idf didasarkan pada istilah yang muncul pada setiap dokumen tidak memberikan suatu ciri khusus untuk menentukan dokumen yang relevan dari yang tidak relevan. Jika jumlah seluruh dokumen didalam sistem dinyatakan dengan nilai N dan jumlah dokumen yang memiliki istilah ki tersebut dinyatakan dengan dfi, maka nilai idfi-nya dapat dinyatakan dengan:

$$idf_i = \log_2\left(\frac{N}{df_i}\right) \qquad (3)$$

Bobot setiap term dapat direpresentasikan dengan frekuensi invers dokumennya (TF-IDF) yang dinyatakan sebagai berikut :

$$w_{ij} = tf_{ij}.\log_2\left(\frac{N}{df_i}\right) \qquad (4)$$

dimana wij adalah bobot term ke i pada dokumen ke j yang bersangkutan, tfij adalah frekuensi term ke i pada dokumen ke j. N adalah jumlah dokumen yang diproses dan dfj adalah jumlah dokumen yang memiliki term ke i di dalamnya. [5]

E. Likelihood

Perhitungan likelihood untuk sebuah kategori dijelaskan pada rumus (5).

$$\text{Likelihood } (c_j \mid A=\{k_1,k_2,\ldots,kn\}) = -\sum_{i=1}^{n} P(k_i|c_j)\log\left(P(k_i|c_j)\right) \qquad (5)$$

dimana cj adalah kategori ke j, A adalah artikel dokumen uji, P(k_i |c_j) dihitung menggunakan "In-Document" dan perhitungan "jumlah total dokumen".

Setelah semua kategori dihitung nilai likelihood-nya maka nilai ambang batas adapat diperoleh. Nilai ambang (threshold) digunakan untuk menentukan apakah sebuah kategori dapat ditetapkan untuk artikel uji atau tidak. Nilai ini didapatkan dari standar deviasi dan rata – rata. L adalah jumlah banyaknya likelihood, sementara li adalah likelihood untuk kategori ke – i. Asumsinya adalah kategori – kategori yang tepat akan memiliki nilai yang besarnya jauh berbeda dibandingkan kategori – kategori lainnya. [1]

$$Threshold = \frac{\sum_i^{|L|} l_i}{|L|} + \sqrt{\frac{\sum\left(l_i - \frac{\sum_i^{|L|} l_i}{|L|}\right)^2}{|L|}} \qquad (6)$$

F. Algoritma Identifikasi Topik

Algoritma identifikasi topik dapat dibagi menjadi dua proses, yaitu klasifikasi dan *dynamic thresholding.* Algoritma ini menghitung kemiripan antara kata kunci topik yang telah diketahui sebelumnya dan kata kunci artikel uji. Setelah itu, nilai yang memiliki *similarity* paling tinggi ditetapkan untuk artikel sebagai *conditionally assigned topic.* [1]

Untuk membandingkan antara vector kata kunci dengan vector topik, keduanya ditransformasikan ke dalam *vector-space* yang sama. Contoh :

Topic:	war	iraq	US	UK		war	iraq	US	UK	violence
	2	5	4	1	⇒	2	5	4	1	0
Article:	war	iraq	violence			war	iraq	US	UK	violence
	1	3	1		⇒	1	3	0	0	1

Gambar 2.2. Contoh Transformasi Vektor [1]

Setelah itu dihitung nilai similarity menggunakan rumus berikut :

$$CosSim\,(t_i,A) = \frac{t_i A}{|t_i||A|} \qquad (7)$$

dengan t_i adalah vector topik ke-i, A adalah artikel uji A, $|t_i|$ dan |A| masing-masing adalah panjang vector topik ke –i dan panjang vector artikel A.

Topik yang memiliki similarity terbesar nantinya akan diuji menggunakan nilai threshold dinamis (*dynamic threshold).* Nilai ambang ini akan membandingkan antara nilai topik awal yang ditentukan dengan nilai topik baru yang mungkin terbentuk *NewTSim.* [1]

$$\text{NewTSim } (t_c,A) = \frac{(0.05\,x\,|t_c|\,x\,(mean(A)-StdDev(A))\,x\,mean(t_c)}{\left(|A|\,x\,(mean(A))^2\right)\,x\,(|t_c|\,x\,(mean(t_c))^2)} \qquad (8)$$

dengan t_c merupakan topik awal yang telah ditentukan, yaitu hasil perhitungan CosSim terbesar, Mean (A) adalah rata-rata dokumen A, StdDev(A) adalah standar deviasi vector dokumen A, dan mean(t_c) adalah rata-rata topik awal yang telah ditentukan.

Langkah selanjutnya adalah *dynamic thresholding*, yaitu membandingkan nilai NewSTim dengan nilai topik awal yang telah ditentukan sebagai berikut (9) :

(i) $CosSim\,(t_c,A) > 0.1 \wedge CosSim\,(t_c,A) > NewTSim\,(t_c,A)$

(ii) $NumTopics > 10 \wedge CosSim\,(t_c,A) > (2\,x\,StdDev(AllTopicSims) + Mean(AllTopicSims))$

Dengan $CosSim\,(t_c,A)$ adalah hasil perhitungan Cosine Similarity terbesar yang diperoleh dari rumus (7) dan diasumsikan sebagai topik awal yang ditentukan. NumTopic adalah jumlah keseluruhan topik yang telah diketahui sebelumnya, StdDev(AllTopicSims) dan Mean(AllTopicSims) masing-masing adalah standar deviasi dan rata-rata seluruh similarity topik. [1]

G. Metode Evaluasi Uji Coba

Pelaksanaan evaluasi uji coba seringkali menggunakan rumus *precision*, *recall*, *F-Measure* dan *accuration*. Adapun pengertian dari beberapa metode di atas adalah :

- *Precision* adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh system yang dirumuskan sebagai berikut :

  Precision (P) = TP / (TP + FP)

- *Recall* adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi yang dirumuskan sebagai berikut :

  Recall (R) = TP / (TP + FN)

- *F-Measure* adalah harmonic mean dari precision dan recall yang dirumuskan sebagai berikut :

  F-Measure (F) = 2 * P * R / (P + R)

- *Accuracy* didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual yang dirumuskan sebagai berikut :

  Accuration (A) = (TP + TN) / (TP + FP + FN + TN)

Tabel 2.1 *Item* Penyusun *Precision, Recall, F-measure, Accuration* [8]

		Nilai Sebenarnya	
		TRUE	FALSE
Nilai prediksi	TRUE	TP (True Positive) *Corect result*	FP (False Positive) *Unexpected result*
	FALSE	FN (False Negative) *Missing result*	TN (True Negative) *Corect absence of result*

III. METODE PENELITIAN DAN PERANCANGAN SISTEM

A. Studi Literatur

Pada tahap pertama ini dilakukan identifikasi masalah dan akan dilakukan pengkajian tentang *preproses* identifikasi topik yang meliputi pencarian dan pemahaman informasi soal representasi text dokumen, *stopword elimination*, *stemming*, ekstraksi kata kunci, metode evaluasi, serta penerapan metode klasifikasi dan identifikasi topik pada dokumen berita.

B. Perancangan perangkat lunak (PL)

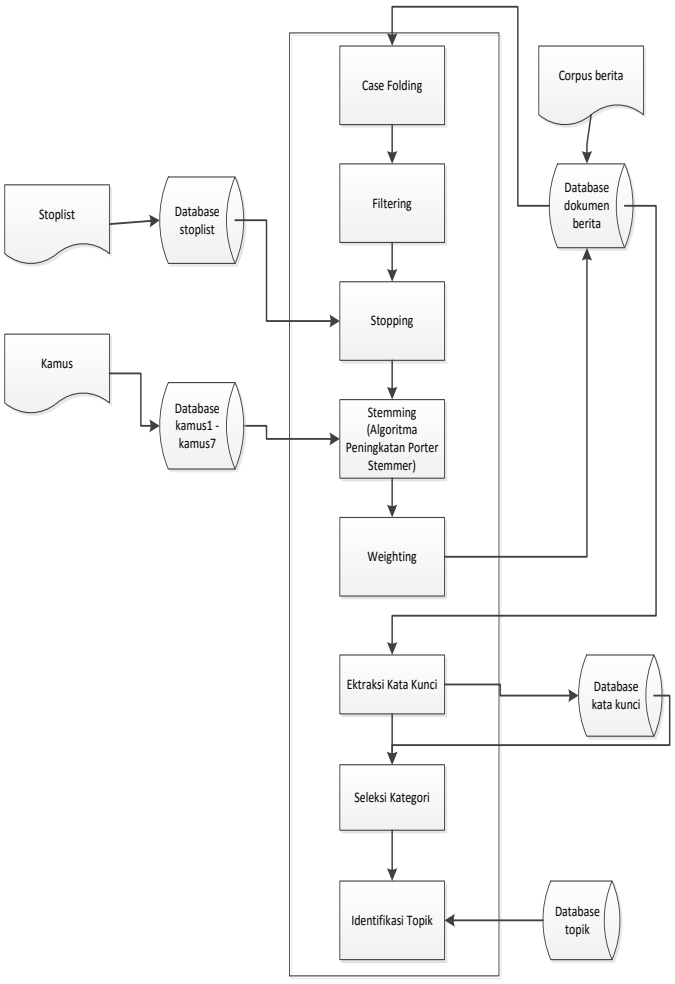
Perancangan PL ini terdiri dari perancangan *software enginering*, *use case* diagram, rancangan database, rancangan *interface* yang dapat dilihat pada gambar 3.2-3.5.

C. Pengumpulan dokumen berita dan Corpus

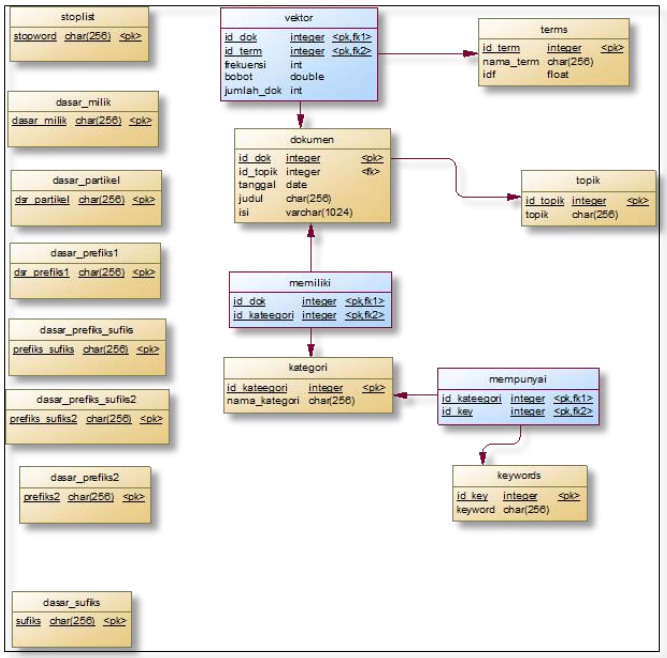
Data input aplikasi ini berupa corpus dokumen berita berbahasa Inggris dengan ekstensi .news. Ekstensi .news digunakan untuk mempermudah pengambilan file baik saat preproses ataupun proses pembelajaran aplikasi dilakukan. Corpus mempunyai format tanggal, kode sumber, judul dan isi dokumen berita. Corpus akan diambil melalui situs www.kompas.com. Corpus yang digunakan dalam tugas akhir ini menggunakan data yang sudah digunakan pada [2] yaitu 932 data training dan 10 data testing untuk masing-masing kategori.

Tanggal_berita	<Day, DD Month YYYY>
Topik_berita	<Topik Berita>
ID_Sumber	<ID Sumber Berita>
Judul_berita	<Judul Berita>
Isi_berita	<Isi berita>

Gambar 3.1 Format Corpus [2]

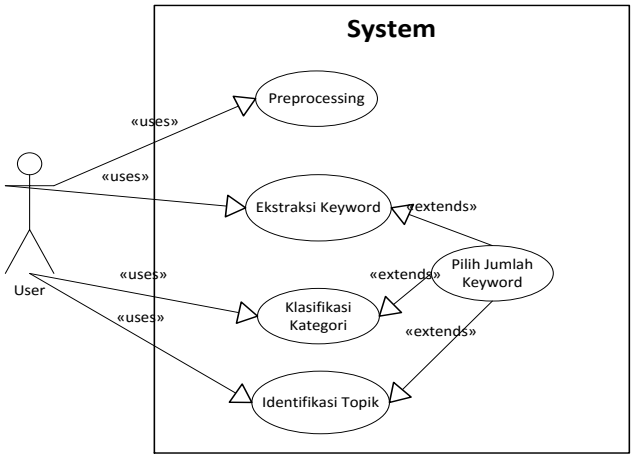


Gambar 3.2 Diagram Alir Sistem

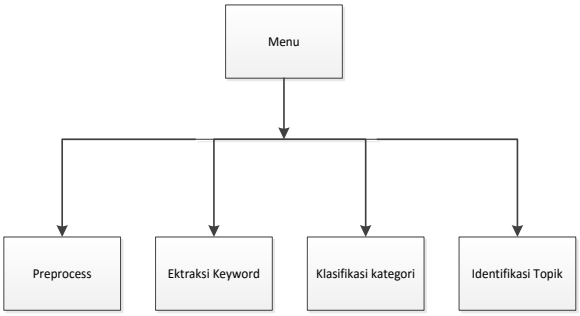


Gambar 3.3 *Desain Physical Data Model*

Desain Physical Data Model di atas selanjutnya akan digunakan sebagai dasar untuk merancang database menggunakan MySQL.



Gambar 3.4 *Use Case Diagram*



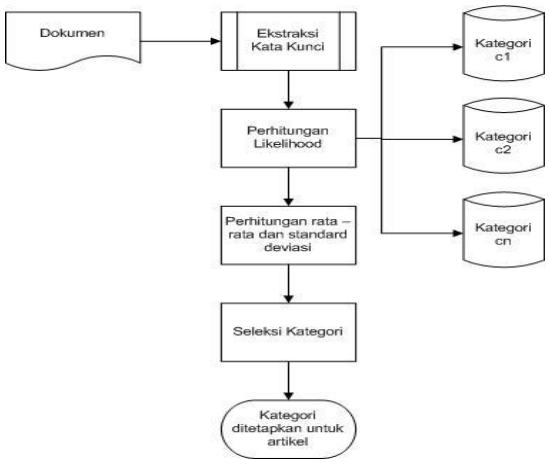
Gambar 3.5 Diagram Rancangan *Interface*

D. Training Teks Dokumen

Dalam tahap ini akan dilakukan *Preprocess*, yaitu tahap perancangan fungsi-fungsi yang dapat diterapkan dalam aplikasi. Diantaranya adalah dokumen *training* harus direpresentasikan dalam bentuk vector yang meliputi *Case folding*, *Filtering*, *Stoplist Removal*, *Stemming*, dan *Weighting*.

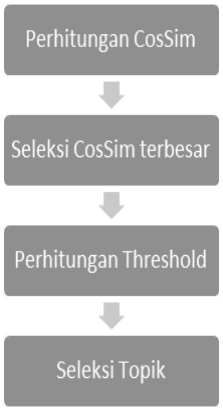
1) *Case Folding :* Seluruh huruf pada setiap kata dalam dokumen diubah menjadi huruf kecil
2) *Filtering :* eliminasi tanda baca
3) *Stoplist Removal :* penghilangan karakter yang memiliki frekuensi tinggi, karena dianggap bukan merupakan kata penting. Kata – kata tersebut antara lain: preposisi, konjungsi, dan lain – lain. Kata –kata yang termasuk dalam stoplist disebut stopword dan telah disimpan dalam database.
4) Penerapan Algoritma Peningkatan Porter pada Proses Stemming [5]
5) *Weighting :* pembobotan setiap terms yang telah di-stem melalui metode TF-IDF.

E. Klasifikasi Kategori



Gambar 3.6 Proses Klasifikasi Kategori

F. Identifikasi Topik



Gambar 3.7 Proses Identifikasi Topik

G. Uji coba dan Evaluasi Hasil Klasifikasi dan Identifikasi

Pada tahap ini akan dilakukan uji coba pada program. Yaitu menguji data *testing* Corpus yang sudah disimpan sebelumnya dengan jumlah 10 data untuk masing-masing kategori. Hasil dari klasifikasi setiap kategori akan dihitung nilai *precision, recall, f-measure, accuracy*. Sedangkan untuk identifikasi topik menggunakan 90 data testing dan akan dihitung nilai *accuracy* berdasarkan jumlah teridentifikasi benar dibagi dengan total data uji.

IV. UJI COBA DAN PEMBAHASAN

A. Data Uji Coba

Karakteristik : Data berupa corpus berita online berbahasa Indonesia yang didapatkan dari www.kompas.com. Berita diunduh berdasar kategori yang telah ditetapkan. Kategori primitif dalam uji coba berguna untuk mengevaluasi hasil klasifikasi.

Jumlah : Antara sebuah kategori dengan kategori lainnya memiliki jumlah dokumen uji yang sama. Spesifikasi jumlah dokumen untuk setiap kategori dapat dilihat pada Tabel 4.1

Tabel 4.1 Spesifikasi Jumlah Dokumen Setiap Kategori

Kategori	Jumlah Dokumen
Nasional	10
Regional	10
Internasional	10
Metropolitan	10
Bisnis dan Ekonomi	10
Olahraga	10
Sains dan Teknologi	10
Edukasi	10
Pariwisata	10
Total	**90**

B. Analisis Hasil Uji Coba

Tabel 4.2 Evaluasi Klasifikasi Kategori dengan 5 Kata Kunci

No	Kategori	Precision	Recall	F-Measure	Accuracy
1	Internasional	1.0000000000	0.8000000000	0.8888888889	1.0000000000
2	Nasional	0.8888888889	0.8888888889	0.8888888889	0.9000000000
3	Regional	0.8750000000	0.7777777778	0.8235294118	0.8000000000
4	Metropolitan	0.8750000000	0.7777777778	0.8235294118	0.8000000000
5	Bisnis Ekonomi	0.8888888889	0.8888888889	0.8888888889	0.9000000000
6	Olahraga	0.8750000000	0.7777777778	0.8235294118	0.9000000000
7	Pariwisata	0.8750000000	0.7777777778	0.8235294118	0.9000000000
8	Sains Tekologi	0.8888888889	0.8888888889	0.8888888889	1.0000000000
9	Edukasi	0.8888888889	0.8888888889	0.8888888889	0.9000000000
Rata-Rata		0.8950617284	0.8296296296	0.8598402324	0.9000000000

Tabel 4.3 Evaluasi Klasifikasi Kategori dengan 10 Kata Kunci

No	Kategori	Precision	Recall	F-Measure	Accuracy
1	Internasional	1.0000000000	0.8000000000	0.8888888889	1.0000000000
2	Nasional	0.8888888889	0.8888888889	0.8888888889	1.0000000000
3	Regional	0.7777777778	0.7777777778	0.7777777778	0.7000000000
4	Metropolitan	0.8750000000	0.7777777778	0.8235294118	0.8000000000
5	Bisnis Ekonomi	0.8888888889	0.8888888889	0.8888888889	1.0000000000
6	Olahraga	0.8888888889	0.8888888889	0.8888888889	0.9000000000
7	Pariwisata	0.8888888889	0.8888888889	0.8888888889	1.0000000000
8	Sains Tekologi	0.8888888889	0.8888888889	0.8888888889	1.0000000000
9	Edukasi	0.8888888889	0.8888888889	0.8888888889	0.9000000000
Rata-Rata		0.8873456790	0.8543209877	0.8692810458	0.9222222222

Tabel 4.4 Evaluasi Klasifikasi Kategori dengan 15 Kata Kunci

No	Kategori	Precision	Recall	F-Measure	Accuracy
1	Internasional	0.8888888889	0.8000000000	0.8421052632	0.9000000000
2	Nasional	0.8888888889	0.8888888889	0.8888888889	1.0000000000
3	Regional	0.8888888889	0.8888888889	0.8888888889	0.9000000000
4	Metropolitan	0.8750000000	0.7777777778	0.8235294118	0.8000000000
5	Bisnis Ekonomi	0.8888888889	0.8888888889	0.8888888889	1.0000000000
6	Olahraga	0.8888888889	0.8888888889	0.8888888889	0.9000000000
7	Pariwisata	0.8888888889	0.8888888889	0.8888888889	1.0000000000
8	Sains Tekologi	0.8888888889	0.8888888889	0.8888888889	1.0000000000
9	Edukasi	0.8888888889	0.8888888889	0.8888888889	0.9000000000
Rata-rata		0.8873456790	0.8666666667	0.8764285441	0.9333333333

Tabel 4.5 Evaluasi Klasifikasi Kategori dengan 20 Kata Kunci

No	Kategori	Precision	Recall	F-Measure	Accuracy
1	Internasional	0.8000000000	0.8000000000	0.8000000000	0.8000000000
2	Nasional	0.8888888889	0.8888888889	0.8888888889	1.0000000000
3	Regional	0.8888888889	0.8888888889	0.8888888889	0.9000000000
4	Metropolitan	1.0000000000	0.8888888889	0.9411764706	1.0000000000
5	Bisnis Ekonomi	0.8888888889	0.8888888889	0.8888888889	1.0000000000
6	Olahraga	0.8888888889	0.8888888889	0.8888888889	0.9000000000
7	Pariwisata	0.8888888889	0.8888888889	0.8888888889	1.0000000000
8	Sains Tekologi	0.8888888889	0.8888888889	0.8888888889	1.0000000000
9	Edukasi	0.8888888889	0.8888888889	0.8888888889	0.9000000000
	Rata-Rata	0.8913580247	0.8790123457	0.8848220770	0.9444444444

Tabel 4.6 Evaluasi Klasifikasi Kategori dengan 25 Kata Kunci

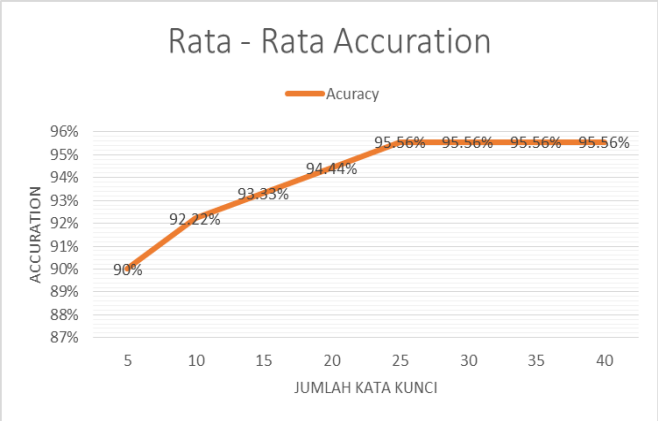
No	Kategori	Precision	Recall	F-Measure	Accuracy
1	Internasional	0.8888888889	0.8000000000	0.8421052632	0.9000000000
2	Nasional	0.8888888889	0.8888888889	0.8888888889	0.9000000000
3	Regional	0.8888888889	0.8888888889	0.8888888889	0.9000000000
4	Metropolitan	1.0000000000	0.8888888889	0.9411764706	1.0000000000
5	Bisnis Ekonomi	1.0000000000	0.8888888889	0.9411764706	1.0000000000
6	Olahraga	0.8888888889	0.8888888889	0.8888888889	0.9000000000
7	Pariwisata	1.0000000000	0.8888888889	0.9411764706	1.0000000000
8	Sains Tekologi	1.0000000000	0.8888888889	0.9411764706	1.0000000000
9	Edukasi	1.0000000000	0.8888888889	0.9411764706	1.0000000000
	Rata-Rata	0.9506172840	0.8790123457	0.9127393648	0.9555555556

Pada tabel 4.2-4.6 dapat kita lihat bahwa hasil klasifikasi kategori menunjukkan hasil yang paling baik menggunakan evaluasi *accuracy*. Sehingga *accuracy* klasifikasi kategori dari tabel-tabel di atas dapat dirangkum sebagai berikut :

Tabel 4.7 *Accuracy* Klasifikasi Kategori

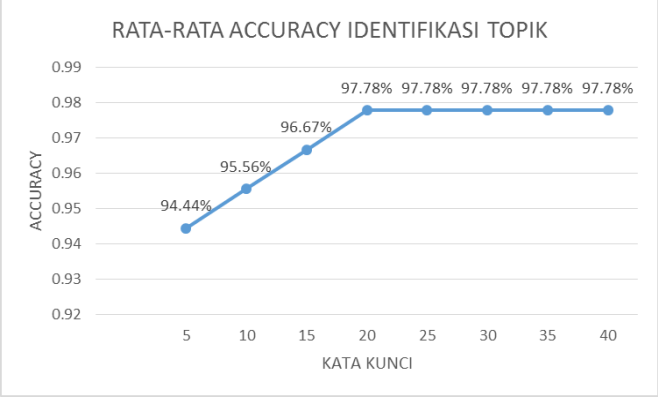
	5	10	15	20	25
	Accuration	Accuration	Accuration	Accuration	Accuration
Kategori					
Internasional	100.00%	100.00%	90.00%	80.00%	100.00%
Nasional	90.00%	100.00%	100.00%	100.00%	100.00%
Regional	80.00%	70.00%	90.00%	90.00%	90.00%
Metropolitan	80.00%	80.00%	80.00%	100.00%	100.00%
Bisnis Ekonomi	90.00%	100.00%	100.00%	100.00%	100.00%
Olahraga	90.00%	90.00%	90.00%	90.00%	90.00%
Pariwisata	90.00%	100.00%	100.00%	100.00%	100.00%
Sains & Teknologi	100.00%	100.00%	100.00%	100.00%	100.00%
Edukasi	90.00%	90.00%	90.00%	90.00%	80.00%
Rata - rata	90.00%	92.22%	93,33%	94,44%	95,56%

Data hasil perhitungan pada tabel 4.7 merupakan perhitungan *accuracy* untuk masing-masing kategori. Sebenarnya evaluasi dilakukan menggunakan pemilihan jumlah kata kunci sebesar 5, 10, 15, 20, 25, 30, 35, 40 dan didapatkan nilai paling maksimal adalah dengan menggunakan 25 kata kunci. Sebab pada jumlah kata kunci 30-40 menghasilkan rata-rata *accuracy* yang sama dengan 25. Hal tersebut dapat kita lihat pada Gambar 4.1.



Gambar 4.1 Rata-Rata Akuarasi Klasifikasi Kategori

Sedangkan untuk identifikasi topik diperoleh hasil sebagai berikut :



Gambar 4.2 Rata-Rata Akurasi Identifikasi Topik

Nilai accuracy untuk identifikasi topik diperoleh sebesar 0.9778 atau 97,78%. Pada identifikasi topik diperoleh hasil yang maksimal saat pemilihan kata kunci sebesar 20.

Jika kita lihat pada [2], nilai akurasi yang dihasilkan untuk klasifikasi kategori dan identifikasi topik masing – masing adalah 93,84 % dan 97,26%. Sedangkan pada tugas akhir ini, dihasilkan nilai akurasi yang lebih tinggi yaitu 95,56% untuk klasifikasi kategori dan 97,78% untuk identifikasi topik.

Selain hasil diatas, pada Gambar 4.1 dapat kita lihat bahwa terjadi peningkatan rata-rata akurasi sebesar 2,22% pada jumlah kata kunci sebesar 10, hal ini dikarenakan terjadi perubahan jumlah dokumen yang diidentifikasi benar pada 4 kategori yaitu penurunan 10 % pada kategori

Regional dan peningkatan 10% pada Nasional, Bisnis Ekonomi dan Pariwisata. Sedangkan pada jumlah kata kunci 10 ke jumlah kata kunci 15 hingga 25, terjadi peningkatan rata-rata akurasi yang konstan yaitu sebesar 1,11%, hal tersebut dikarenakan adanya perubahan jumlah dokumen yang diidentifikasi benar pada 2 kategori yaitu penurunan akurasi 10% pada kategori Internasional dan peningkatan akurasi 10% pada kategori Metropolitan atau terjadi peningkatan 20% pada kategori Internasional namun terjadi penurunan 10% pada kategori Edukasi. Selain itu pada Gambar 4.2 terjadi peningkatan yang konstan sebesar 1,11% dari pemilihan kata kunci sejumlah 5-20, sedangkan dari pemilihan jumlah kata kunci 20-40 tidak terjadi peningkatan, melainkan hasil maksimal diperoleh untuk pemilihan jumlah kata kunci sebesar 20.

V. KESIMPULAN

1. Program telah selesai dibuat menggunakan Algoritma Peningkatan Porter Stemmer dan Likelihood serta diuji mampu melakukan proses klasifikasi kategori serta identifkasi topik pada artikel berita berbahasa Indonesia
2. Berdasarkan hasil uji coba, proses klasifikasi kategori mendapatkan hasil yang optimal saat menggunakan jumlah kata kunci sebanyak 25, sedangkan untuk identifikasi topik diperoleh hasil yang maksimal dengan jumlah kata kunci sebanyak 20.
3. Nilai *accuracy* untuk klasifikasi kategori diperoleh sebesar 95,56 %, sedangkan untuk identifikasi topik sebesar 97,78 %. Kedua nilai tersebut tampak lebih baik daripada nilai *accuracy* yang dihasilkan pada penelitian sebelumnya.

VI. SARAN

Sebagai evaluasi dan pengembangan selanjutnya diharapkan dapat dilakukan beberapa saran berikut :

1. Riset lebih lanjut dalam hal *running time*, karena membutuhkan waktu yang cukup lama saat identifikasi topik.
2. Program disediakan fungsi *download* dokumen agar secara otomatis disimpan mengikuti format Corpus.

DAFTAR PUSTAKA

[1] Bracewell D, Jiajun Yan, Fuji Ren dan Shingo Kuroiwa.2009. “Category Classification and Topic Discovery of Japanese and English News Article,” Electronic Notes in Theoretical Computer Science 225(2009) 51-65.

[2] Fuddoly, Aini Rachmania Kusumaagama, Agus Zainal Arifin.2011. “Klasifikasi Kategori dan Identifikasi Topik pada Artikel Berita Bahasa Indonesia,” ITS.Surabaya

[3] Karaa,Wahiba Ben Abdessalem, “A New Stemmer to Improve Information Retrieval,” International Journal of Network Security & Its Applications (IJNSA), Vol.5, No.4, July 2013

[4] DR.E. Garcia,2006. The Classic Vector Space Model, <URL:http://www.miislita.com/term-vector/term-vector-3.html>

[5] Wiguna,Putu Bagus Susastra, Bimo Sunarfri Hantono."Peningkatan Algoritma Porter Stemmer Bahasa Indonesia berdasarkan Metode Morfologi dengan Mengaplikasikan 2 Tingkat Morfologi dan Aturan Kombinasi Awalan dan Akhiran," JNTETI, Vol.2, No.2,2013

[6] Agusta Ledy, "Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia", Konferensi Nasional Sistem dan Informatika 2009; Bali, November 14, 2009

[7] Nadirman Firnas, 2006. Sistem Temu-Kembali Informasi Dengan Metode Vector Space Model Pada Pencarian File Dokumen Berbasis Teks, <URL:http://kabulkurniawan.web.ugm.ac.id/wp-content/uploads/SKRIPSI.pdf>

[8]<URL:https://dataq.wordpress.com/2013/06/16/perbedaan-precision-recall-accuracy/