Sistem Deteksi Bahasa pada Dokumen menggunakan N-Gram

Badrus Zaman, Eva Hariyanti, Endah Purwanti
Program Studi Sistem Informasi, Fakultas Sains dan Teknologi Universitas Airlangga
Jalan Mulyorejo Kampus C, Surabaya 60115
badruszaman@fst.unair.ac.id, eva.hariyanti@gmail.com, endah.purwanti@gmail.com

Diterima: 11 September 2015. Disetujui: 10 Oktober 2015. Dipublikasikan: November 2015

Abstract - Language detection on a very large collection of documents can be done to increasing performance of information retrieval system. One of popular method on language detection is N-Grams, based on pieces of n-characters taken from a string. This research is developed language detection system based on N-Gram that performs by Indonesian or English language. In general, the steps being taken there were 3 phases, namely creating profile of each language, system testing, and system evaluation. Fifty documents were used to creating profile of each language, i.e. 25 Indonesian and 25 English. Sixty documents were used for system testing. System performance was evaluated using F-measures. Based on the test, obtained F-measures for unigram, bigram, and unigram respectively 0.933, 0.917, and 0.933.

Keywords: information retrieval, language detection, N-Gram, Indonesia language, English.

I. PENDAHULUAN

Meningkatnya jumlah dokumen mengakibatkan tingkat kompleksitas dalam penemuan kembali informasi menjadi meningkat. Dewasa ini diperkirakan ada sekitar 3,3 milyar dokumen ter-indeks mesin pencari, yang terdiri dari berbagai bahasa [1]. Meningkatnya jumlah dokumen dapat mengakibatkan penurunan terhadap kinerja *search engine*, apalagi dokumen-dokumen tersebut melibatkan berbagai macam bahasa, sehingga tingkat relevansi hasil pencarian menjadi rendah. Oleh karena itu, perlu dilakukan upaya untuk memberikan fitur tertentu sehingga tingkat relevansi hasil pencarian dapat meningkat. Salah satu upaya yang dapat dilakukan dengan mendeteksi bahasa pada dokumen, sehingga hasil pencarian menjadi lebih relevan dengan biaya komputasi rendah.

Deteksi bahasa (*language detection*), biasa juga disebut identifikasi bahasa (*language identification*) adalah usaha untuk menentukan jenis bahasa secara otomatis dari suatu teks atau dokumen berdasarkan kriteria-kriteria tertentu [2]. Pendeteksian bahasa bertujuan untuk mengklasifikasikan bahasa suatu dokumen berdasarkan *training* yang dilakukan menggunakan koleksi dokumen atau corpus. Menurut Grothe dkk. deteksi bahasa dilakukan sebagai saringan awal pada corpus, sehingga dapat menghasilkan *input* data yang berkualitas [3].

Salah satu pendekatan dalam pendeteksian bahasa menggunakan N-Gram. N-Gram adalah potongan N-karakter yang diambil dari suatu *string*. Ahmed dkk. [2] dan Grothe dkk. [3] mengidentifikasi salah satu keunggulan N-Gram dalam identifikasi bahasa suatu dokumen adalah sifatnya yang resistan pada berbagai macam kesalahan dalam penulisan, sehingga kesalahan pada sebagian *string* hanya berakibat perbedaan pada sebagian N-Gram. Kenyataan bahwa bahasa manusia selalu memiliki beberapa kata yang frekuensi kemunculannya lebih tinggi dibandingkan dengan kata lainnya, juga menjadi dasar digunakannya N-Gram [4]. Akibatnya, frekuensi masing-masing huruf pun dapat bervariatif, misalnya huruf "a" frekuensi munculnya paling tinggi pada bahasa Indonesia, sedangkan untuk bahasa Inggris, vokal "e" merupakan huruf yang frekuensinya paling tinggi [1]. Perbedaan frekuensi kemunculan huruf atau kata, menandakan bahwa N-Gram dari suatu bahasa bersifat unik, sehingga dapat dijadikan sebagai profil pada tiap-tiap bahasa.

Berdasarkan uraian di atas, maka fokus pada penelitian ini adalah implementasi sistem deteksi bahasa pada dokumen multi bahasa menggunakan N-Gram. Variasi dokumen bahasa yang digunakan dalam penelitian ini adalah bahasa Indonesia dan Inggris.

II. TINJAUAN PUSTAKA

A. Deteksi Bahasa

Algoritma deteksi bahasa (*language detection*), biasa juga disebut identifikasi bahasa (*language identification*) usaha untuk menentukan jenis bahasa secara otomatis misalnya dengan

program komputer dari suatu teks atau dokumen berdasarkan kriteria-kriteria tertentu yang harus dipenuhi [1]. Identifikasi bahasa adalah salah satu yang paling dasar langkah yang harus diambil dalam banyak sistem yang melibatkan NLP (*Nature Language Processing*) seperti *summarization*, *question answering*, *translation* dan sebagainya [5]. Identifikasi bahasa dapat digunakan sebagai teknik *filtering* untuk mendukung pengguna sistem pencarian informasi yang hanya tertarik pada dokumen dalam bahasa tertentu. Selain itu, identifikasi bahasa penting sebagai langkah pre-*processing* untuk teknik pengolahan bahasa lain seperti *stemming* atau mesin terjemahan yang hanya dapat diterapkan pada dokumen ketika bahasa tersebut dokumen sudah diketahui [3].

B. N-Gram

N-Gram adalah suatu urutan *n* unit yang pada umumnya berupa karakter tunggal atau *string* yang dipisahkan oleh spasi [3]. N-Gram adalah potongan N-karakter yang diambilkan dari suatu *string*. *Blank* ditambahkan pada awal dan akhir suatu *string* untuk mengetahui batas awal dan akhir suatu *string*. Misalnya suatu *string* "TEXT" setelah ditambah awal dan akhir dengan "_" sebagai pengganti *blank* akan didapat N-Gram sebagai berikut :

Unigram : T,E,X,T
Bigram : _T, TE, EX, XT, T_
Trigram : _TE, TEX, EXT, XT_, T_ _

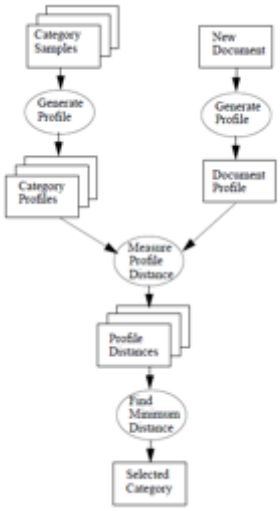
Dapat disimpulkan bahwa untuk *string* berukuran n akan dimiliki n unigram dan n+1 bigram, n+1 trigram dan seterusnya. Penggunaan N-Gram untuk *matching* kata memiliki keuntungan sehingga dapat diterapkan pada *recovery* pada *input* karakter ASCII yang terkena *noise*, *interpretasi* kode pos, *information retrieval* dan berbagai aplikasi dalam pemrosesan bahasa alami [4].

Keuntungan N-Gram dalam pencocokan *string* karena karakteristik N-Gram sebagai bagian dari suatu *string*, sehingga kesalahan pada sebagian *string* hanya akan berakibat perbedaan pada sebagian N-Gram. Sebagai contoh jika N-Gram dari dua *string* dibandingkan, kemudian menghitung cacah N-Gram yang sama dari dua *string* tersebut maka akan didapatkan nilai kesamaan dua *string* tersebut yang bersifat resistan terhadap kesalahan tekstual [4].

C. Profil Frekuensi N-Gram

Setiap bahasa yang digunakan manusia memiliki beberapa kata yang muncul lebih sering dari kata lainnya, sehingga frekuesi kemunculan N-Gram berbeda antara satu bahasa dengan bahasa lainnya karena karakteristik N-Gram berdasarkan bagian dari suatu *string*, hal ini juga menyebabkan perbedaan pada frekuensi profil N-Gram yang dimiliki dokumen dengan bahasa yang berbeda [4]. Profil yang dimaksud adalah yang menggambarkan ciri-ciri yang khas suatu dokumen yang dibentuk dari N-Gram yang dimiliki oleh dokumen suatu bahasa. Dalam menghasilkan profil frekuensi N-Gram sistem membaca dokumen yang menjadi masukan kemudian menghitung jumlah kemunculan setiap N-Gram pada dokumen tersebut.

Gambar 1 menggambarkan proses umum yang dilakukan dalam mendeteksi suatu bahasa menggunakan N-Gram, yang dimulai dengan seperangkat teks yang terdiri dari berbagai macam kategori bahasa yang digunakan untuk data *training* masing-masing model bahasa, yang masing-masing kategori memiliki ribuan cacah kata. Dari seperangkat teks tersebut akan dihitung kemunculan tiap N-Gram untuk menghasilkan satu set profil frekuensi N-Gram untuk mewakili masing-masing kategori bahasa. Ketika dokumen baru akan dideteksi bahasanya maka akan dihitung kemunculan tiap N-Gram yang ada pada dokumen baru tersebut untuk menghasilkan satu set profil frekuensi N-Gram untuk dokumen baru tersebut. Kemudian sistem akan membandingkan profil frekuensi N-Gram dokumen baru dengan profil frekuensi N-Gram masing-masing kategori bahasa dengan menggunakan ukuran jarak. Sistem akan mengklasifikasikan dokumen ke dalam kategori bahasa tertentu yang ukuran jarak profilnya paling kecil [1][6].



Gambar 1. Proses Deteksi Bahasa Menggunakan N-Gram

D. Penentuan Profil Unigram

Unigram adalah sebuah N-Gram yang terdiri dari satu item dari *sequence*. Persamaan untuk menghitung statistik unigram pada sebuah dokumen ditunjukkan pada persamaan 1 [1].

$$simQ = \sum_{i=a}^{z} \frac{(f_i - fR_i)^2}{fR_i} \quad (1)$$

fi = probabilitas unigram ke-i dari dokumen yang terdeteksi
fRi = probabilitas unigram ke-i dari dokumen *training* set
i = unigram ke i, yaitu a,b,c…. z

Nilai simQ digunakan sebagai ukuran jarak antara profil frekuensi unigram dokumen baru dan profil frekuensi unigram pada *training* set, kemudian nilai simQ akan dibandingkan dengan nilai *threshold* untuk menetapkan bahasa dokumen dengan ketentuan nilai simQ lebih kecil dari nilai *threshold*.

E. Penentuan profil bigram dan trigram

Bigram adalah sebuah N-Gram yang terdiri dari 2 item dari *sequence* sedangkan Trigram adalah N-Gram yang terdiri dari 3 item dari *sequence*. Persamaan untuk menghitung statistik dari Bigram dan Trigram pada dokumen ditunjukkan pada persamaan 2 [1].

$$W_{i,d} = \begin{cases} \frac{freq}{N}, & \text{jika bigram, trigram ada dalam } training \text{ set} \\ -\frac{N_{i,d}}{\sum|word_{j,d}|}, & \text{jika bigram, trigram tidak ada dalam } training \text{ set} \end{cases} \quad (2)$$

Wi,d = bobot bigram/trigram i pada dokumen d
Freq = frekuensi kemunculan bigram/trigram i dalam *training* set
N = jumlah bigram/trigram dalam *training* set
Ni,d = banyak kata dalam dokumen d yang mengandung bigram/trigram i
Wordj,d = kata ke-j dalam dokumen d yang mengandung bigram/trigram i
|wordj,d| = panjang dari kata wordj,d

Persamaan untuk menentukan nilai h(d) dokumen dengan Bigram dan Trigram ditunjukkan pada persamaan 3 [1].

$$h(d) = \sum_i \left(\frac{f_{i,d}}{N_d} w_{i,d}\right) \quad (3)$$

Fi,d = frekuensi bigram/trigram I dalam dokumen d
Nd = banyaknya bigram/trigram dalam dokumen d
Wi,d = bobot bigram/trigram i pada dokumen d dari persamaan 2.

Nilai h(d) tersebut akan digunakan sebagai ukuran jarak antara profil frekuensi bigram dan trigram dokumen baru dan profil frekuensi bigram dan trigram pada *training* set, kemudian nilai h(d) akan dibandingkan dengan nilai *threshold* untuk menetapkan bahasa dokumen dengan ketentuan nilai h(d) lebih besar dari nilai *threshold*.

III. METODE PENELITIAN

Secara umum, metode penelitian dilakukan dengan 4 tahap, yaitu pengumpulan data, penentuan profil N-Gram, pengujian dokumen, dan evaluasi sistem.

A. Pengumpulan Data

Data yang dikumpulkan dalam penelitian ini adalah data dokumen online dari layanan portal berita. Data yang dikumpulkan berupa dokumen bahasa Indonesia dan bahasa Ingggris.

B. Penentuan Profil N-Gram

Penentuan profil N-Gram dilakukan dari masukan berupa dokumen multi bahasa, kemudian membuat profil untuk masing-masing bahasa pada tiap N-Gram, yakni unigram, bigram, dan trigram. Penentuan profil perlu dilakukan agar dapat digunakan untuk mendeteksi suatu bahasa dengan cara mencocokan kemiripan profil sesuai N-Gram.

C. Uji Coba Sistem

Pengujian dokumen dilakukan dengan masukan dokumen suatu bahasa, kemudian sistem menentukan bahasa berdasarkan masing-masing kedekatan profil pada tiap-tiap N-Gram.

D. Evaluasi sistem

Evaluasi terhadap sistem dilakukan dengan menghitung *Recall* dan *Precision*. *Recall* dan *Precision* berguna untuk mengetahui kemampuan sistem dalam melakukan temu-kembali dokumen yang ingin ditemu-kembalikan sesuai dengan jumlah dokumen yang ada dan keakuratan sistem dalam menghasilkan output. Hasil *Recall* dan *Precision* kemudian dikombinasikan menjadi F-Measures yang merupakan sebuah teknik untuk mengukur tingkat keberhasilan suatu sistem temu kembali. Persamaan *recall*, *precision*, dan F-*measures* ditunjukkan pada persamaan 4, 5, dan 6.

$$Recall = \frac{Tp}{Tp + Fn} \quad (4)$$

$$Precision = \frac{Tp}{Tp + Fp} \quad (5)$$

$$F - Measures = \frac{2xPrecisionxrecall}{Precision + Recall} \quad (6)$$

Untuk menentukan kategori tersebut, dapat disesuaikan dengan tabel *contingency* pada Tabel 1.

TABEL 1.CONTINGENCY TABLE

	Relevant	*Nonrelevant*
Retrieved	*True Positive (Tp)*	*False Positive (Fp)*
Not retrieved	*False Negative (Fn)*	*True Negative (Tn)*

Tp : dokumen relevan yang ditemukan
Fp : dokumen tidak relevan yang ditemukan
Fn : dokumen relevan yang tidak ditemukan
Tn : dokumen yang tidak relevan dan tidak ditemukan

IV. HASIL DAN PEMBAHASAN

Hasil dan pembahasan disesuaikan dengan metode penelitan, sehingga penjelasannya dibagi menjadi empat bagian, yaitu pengumpulan data, penentuan profil N-Gram, pengujian dokumen, dan evaluasi sistem.

A. Pengumpulan Data

Data yang dikumpulkan berupa dokumen bahasa Indonesia dan Inggris sejumlah 110 dokumen, 50 dokumen untuk data *training*, dan 60 dokumen untuk data *testing*. Rincian bahasa dokumen tersebut ditunjukkan pada Tabel 2.

TABEL 2. REKAPITULASI DOKUMEN

Dokumen *Training*		Dokumen *testing*	
Bahasa	Jml	Bahasa	Jml
Bin	25	Bin	20
Eng	25	Eng	20
		Lainnya	20
Total	50	Total	60

B. Penentuan Profil N-Gram

Langkah-langkah yang dilakukan dalam penentuan profil frekuensi N-Gram dari sebuah dokumen sebagai berikut :

1. Dokumen dibagi menjadi kata-kata. *Blank* ditambahkan pada awal dan akhir kata.
2. Tiap kata di-*scan*, untuk menghasilkan kemungkinan semua N-Gram, untuk N=1 sampai 3.
3. Masukkan ke dalam tabel untuk menemukan nilai untuk setiap N-Gram dan menjumlahkan nilai dalam tabel dengan nilai hasil *scan*.
4. Profil frekuensi tiap N-Gram didapatkan dengan cara membagi jumlah kemunculan tiap N-Gram dengan jumlah keseluruhan N-Gram pada tiap kategori unigram, bigram dan trigram.

Dari langkah ini didapatkan sebanyak sebanyak 6 profil. Profil untuk bahasa Indonesia ada 3, yaitu untuk unigram, bigram, dan trigram, begitu juga untuk bahasa Inggris ada 3, yaitu unigram, bigram, dan trigram. Profil unigram untuk bahasa Indonesia dan bahasa Inggris masing-masing ditunjukkan pada Tabel 3 dan 4.

TABEL 3. PROFIL UNIGRAM UNTUK BAHASA INDONESIA

Ug	*Frek.*	Ug	*Frek.*	Ug.	*Frek.*
a	61895	i	31852	r	18862
b	9174	j	2604	s	17540
c	2463	k	17246	t	20899
d	14267	l	13056	u	16774
e	109328	m	15660	w	1770
f	2011	n	35438	y	4933
g	13492	o	10167		
h	7488	p	12215		
Total *Unigram*				439134	

TABEL 4. PROFIL UNIGRAM UNTUK BAHASA INGGRIS

Ug	*Frek.*	Ug	*Frek.*	Ug	*Frek.*
a	74679	i	63648	r	52616
b	14669	j	1554	s	59191
c	27956	k	7362	t	83086
d	33645	l	38371	u	24575
e	30441	m	22051	w	17346
f	18870	n	63167	y	16999
g	18830	o	67096		
h	46223	p	17635		
Total *Unigram*				800010	

C. Uji Coba Sistem

Pengujian dokumen dibagi atas 2 jenis, yaitu pengujian dokumen untuk unigram dan bigram /trigram. Langkah-langkah pengujian dokumen baru dengan unigram, sebagai berikut:

a. Menentukan profil dokumen uji menggunakan persamaan 1. Semakin kecil nilai simQ maka dokumen semakin mirip dengan profil.
b. Menetapkan nilai *threshold* (ambang batas) untuk masing-masing dokumen bahasa Indonesia dan bahasa Inggris.
c. Menentukan kategori bahasa dokumen baru dengan mengambil nilai simQ yang paling kecil dari nilai simQ tiap bahasa, kemudian membandingkan nilai simQ tersebut dengan nilai *threshold*. Jika nilai simQ lebih kecil dari nilai *threshold* maka ditetapkan bahwa dokumen tersebut memiliki bahasa yang sama dengan *training* set, sebaliknya apabila nilai simQ lebih besar dari nilai *threshold* maka dokumen tersebut tidak masuk dalam kategori bahasa Indonesia maupun Inggris.

Langkah-langkah pengujian dokumen baru dengan bigram dan trigram, sebagai berikut :

a. Menentukan nilai h(d) berdasarkan persamaan 3.
b. Jika nilai h(d) cenderung besar maka profil frekuensi bigram dan trigram pada dokumen *training* set dan dokumen yang diuji memiliki nilai profil frekuensi yang mirip, sebaliknya dianggap memiliki nilai profil yang tidak mirip.
c. Menetapkan nilai *threshold* (ambang batas) untuk masing-masing dokumen bahasa Indonesia dan bahasa Inggris.
d. Menentukan kategori bahasa dokumen baru dengan cara membandingkan nilai h(d) terkecil tiap bahasa dengan nilai *threshold*. Jika nilai h(d) lebih besar dari *threshold*, maka dokumen tersebut memiliki bahasa yang sama dengan *training* set, sebaliknya dokumen tersebut tidak termasuk dalam data *training*.

1. Penentuan nilai simQ dan h(d)

Nilai simQ didapatkan dari perhitungan dokumen *training* dan profil unigram untuk masing-masing bahasa, pada tiap-tiap unigram dengan menggunakan persamaan1. Sedangkan nilai h(d) untuk bigram dan trigram dihasilkan dengan menghitung nilai frekuensi relatif untuk bigram dan trigram kalimat yang diuji dengan nilai frekuensi relatif untuk bigram dan trigram dari data *training* set bahasa Indonesia dan bahasa Inggris dengan menggunakan persamaan 3. Hasil perhitungan nilai SimQ dan h(d) ditunjukkan pada Tabel 5.

TABEL 5. HASIL PERHITUNGAN NILAI SIMQ DAN H(D) KALIMAT YANG DIUJI DENGAN DATA *TRAINING* SET BAHASA INDONESIA DAN BAHASA INGGRIS.

Nilai	Bahasa Indonesia	Bahasa Inggris
simQ	0,349866	0,780255613
h(d) *Bigram*	0,014618	0,010368
h(d) *Trigram*	0,011041	0,007679371

2. Penentuan nilai *threshold*

Nilai *threshold* berfungsi untuk menentukan nilai jarak kedekatan apakah nilai jarak kedekatan itu memenuhi syarat bahwa dokumen tersebut termasuk kategori bahasa yang sama. Nilai *threshold* ditentukan dengan cara melakukan percobaan yang berulang pada sistem. Percobaan yang dilakukan adalah dengan melakukan uji coba pada dokumen-dokumen bahasa Inggris dan bahasa Indonesia dengan data *training* set bahasa yang sama dengan dokumen yang diuji untuk mengetahui batas nilai jarak kedekatan yang dimiliki dokumen-dokumen tersebut. Tabel 6 merupakan hasil penentuan nilai *threshold* untuk masing-masing bahasa berdasarkan hasil percobaan.

TABEL 6. *THRESHOLD* UNTUK BAHASA INDONESIA DAN BAHASA INGGRIS

Bahasa	*Unigram* (simQ)	*Bigram* (h(d))	*Trigram* (h(d))
Bahasa Indonesia	0,07000	0,00800	0,00120
Bahasa Inggris	0,03000	0,00560	0,00100

3. Penentuan bahasa dokumen

Penentuan bahasa dokumen dilakukan dengan cara membandingkan nilai simQ dan h(d) dokumen yang diuji dengan *threshold* bahasa yang cenderung dimiliki oleh dokumen yang diuji. Kecenderungan bahasa dokumen yang diuji dilakukan berdasarkan Tabel 6. Untuk unigram nilai yang akan digunakan adalah simQ. Semakin kecil nilai yang dimiliki oleh simQ berarti semakin kecil pula jarak kedekatan antara kalimat yang diuji dengan data *training* set suatu bahasa. Sedangkan untuk bigram dan trigram nilai yang akan digunakan adalah h(d). Semakin besar nilai yang dimiliki oleh h(d) berarti semakin kecil pula jarak kedekatan antara kalimat yang diuji dengan data *training* set suatu bahasa. Selanjutnya adalah membandingkan nilai simQ dan h(d) kalimat yang telah memiliki kecenderungan pada suatu bahasa dengan nilai *threshold* bahasa tersebut.

4. Hasil Uji Coba Sistem

Uji coba sistem dilakukan terhadap 60 dokumen, dengan cara diujicobakan satu per satu. Ke-60 dokumen yang digunakan dibagi menjadi 3 bagian, yaitu 20 dokumen bahasa Indonesia, 20 dokumen bahasa Inggris, dan 20 dokumen lainnya berbahasa diluar bahasa Indonesia dan bahasa Inggris.

Hasil uji coba untuk dokumen *testing* bahasa Indonesia, menunjukkan bahwa sistem dapat mendeteksi secara benar sebanyak 100% untuk masing-masing N-Gram. Begitu juga untuk hasil uji coba dokumen *testing* bahasa Inggris, sistem dapat mendeteksi secara benar sebanyak 100% untuk masing-masing N-Gram. Sedangkan, untuk hasil uji coba untuk dokumen *testing* lainnya, sistem dapat mendeteksi secara benar sebanyak 16 kali dari 20 percobaan untuk unigram dan trigram, serta 15 kali dari 20 percobaan untuk bigram.

D. Evaluasi sistem

Evaluasi sistem dilakukan terhadap 60 dokumen *testing*, dengan cara diuji cobakan satu per satu. Kemudian sistem akan memberikan hasil deteksi bahasa, baik untuk unigram, bigram, maupun trigram. Hasil uji coba dokumen *testing* ini kemudian direkapitulasi menggunakan tabel *contingency* untuk masing-masing N-Gram, dan dihitung nilai *precision*, *recall*, dan F-*measures* sesuai persamaan 3, 4, dan 5.

Dari hasil evaluasi, kinerja sistem sangat baik. Hal ini ditunjukkan dengan hasil F-*measures* yang mendekati 1, yakni rata-rata 0,93 pada Tabel 7. Sedangkan untuk penggunaan N-Gram yang terbaik didapatkan pada unigram dan trigram.

TABEL 7. REKAPITULASI HASIL EVALUASI SISTEM

	unigram	bigram	trigram
precision	0.933	0.917	0.933
recall	0.933	0.917	0.933
F-*measures*	0.933	0.917	0.933

V. KESIMPULAN

Berdasarkan hasil uji coba dan evaluasi, maka didapatkan beberapa kesimpulan, yaitu:

a. Sistem deteksi bahasa menggunakan N-Gram menunjukkan kinerja yang sangat baik dalam mendeteksi bahasa dari dokumen. Hal ini ditunjukkan dengan hasil rata-rata F-*measures* sebesar 0,93.

b. Kinerja terbaik ditunjukkan pada unigram dan trigram, yakni dengan F-*measures* sebesar 0,93.

REFERENSI

[1] Hamzah, A. (2010). *Deteksi bahasa untuk dokumen teks berbahasa Indonesia*. Dalam prosiding Dukungan ICT dalam bidang industry dan manajemen ESDM. Halaman A-5 – A-13.

[2] Ahmed B., Cha, S.H, dan Tappert C., (2004). *Language Identification from Text Using N-Gram Based Cumulative Frequency Addition*. Proceedings of Student/Faculty Research Day, CSIS, Pace University, May 7th, 2004.

[3] Grothe, L., De Luca, E.W., dan N¨urnberger, A. (2008). *A Comparative Study on Language Identification Methods*. Dalam Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). Halaman 980-985.

[4] Padr´o, M.,dan Padr´o, L. (2004).*Comparing methods for language identification*. Dalam prosiding Procesamiento del Lenguaje Natural. Halaman 155–162.

[5] Lui M., Lau J. H., dan Baldwin T. (2014). *Automatic Detection and Language Identification of Multilingual Documents*. Journal of Transactions of the Association for Computational Linguistics, 2 (2014) 27-40.

[6] Ramisch, C., (2008). *N-Gram models for language detection.* M2R Informatique - Double dipl ̂ome ENSIMAG – UJF/UFRIMA.