KLASIFIKASI TEKS BAHASA BALI DENGAN METODE *SUPERVISED LEARNING NAIVE BAYES CLASSIFIER*

Ida Bagus Gede Widnyana Putra[1], Made Sudarma[2], I Nyoman Satya Kumara[3]

***Abstract*— Increasing availability of Balinese language text documents making the process of finding or classifying information in Balinese text documents is becoming increasingly difficult. Manual classification is inefficient in view of the increase in the number of Balinese written documents. On this paper, application that can classify Balinese text into various document class is presented. The application is developed using Naive Bayes classifier (NBC) method and feature selection using Information Gain (IG) technique. Application is tested using cross validation method. The results shows that average accuracy of 10 fold cross validation is 95.22%.**

***Intisari*— Ketersediaan dokumen teks bahasa Bali yang meningkat jumlahnya membuat proses pencarian informasi pada dokumen teks berbahasa Bali menjadi semakin sulit. Mengklasifikasikanya secara manual menjadi tidak efisien mengingat peningkatan jumlah dokumen yang semakin banyak. Pada penelitian ini dikembangkan sebuah aplikasi yang dapat mengklasifikasikan teks bahasa Bali ke dalam kategori yang ditentukan. Aplikasi ini menggunakan metode *Naive Bayes Classifier* (NBC) dan metode *Information Gain*(IG) untuk seleksi fitur. Aplikasi ini diuji dengan teknik *cross validation*. Hasilnya adalah nilai rata-rata akurasi dari 10 *fold cross validation* sebesar 95,22%.**

Kata Kunci*— *information gain, naive bayes classifier, text mining, cross validation

I. PENDAHULUAN

Kebutuhan informasi pada era globalisasi menuntut penyediaan informasi dapat diperoleh secara mudah, cepat dan tepat. Kemajuan teknologi dan internet membuat penyebaran informasi menjadi lebih mudah dan cepat. Dari hasil riset nasional yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) menunjukkan selama tahun 2014, pengguna internet di Indonesia telah mencapai 88,1 juta atau dengan kata lain tingkat penetrasi internet di Indonesia sebesar 34,9% [1]. Peningkatan penggunaan internet ini mengakibatkan pertumbuhan jumlah informasi yang tersedia meningkat pesat.

Perkembangan teknologi ini membuat berbagai jenis dokumen teks bahasa Bali banyak dipublikasikan secara digital, hal ini memberikan peran yang cukup penting dalam mendukung pelestarian Bahasa Bali dalam menyediakan sumber bacaan berbahasa Bali yang dapat diakses dengan mudah. Namun peningkatan jumlah bacaan bahasa Bali yang tersedia menimbulkan masalah baru dalam menemukan informasi yang relevan sesuai dengan yang diinginkan secara cepat dan tepat.

Dengan jumlah dokumen yang sangat besar, untuk mencari sebuah dokumen akan lebih mudah apabila kumpulan dokumen yang dimiliki terorganisir dan telah dikelompokan sesuai kategorinya masing-masing [2]. Namun proses pengklasifikasian secara manual menjadi tidak efisien mengingat jumlah dokumen yang ada setiap hari bertambah dengan cepat jumlahnya. *Text mining* adalah suatu teknik atau proses yang saat ini sering digunakan untuk melakukan klasifikasi dokumen teks secara otomatis. *Text mining* dapat didefinisikan sebagai suatu proses menggali informasi yang dilakukan seorang pengguna saat berinteraksi dengan sekumpulan dokumen dengan menggunakan alat analisa [3]. Metode *Naïve BayesClassifier* (NBC) sering digunakan dalam penelitian tentang klasifikasi teks karena kesederhanaan dan efektivitasnya yang menggunakan ide dasar probabilitas gabungan dari kata-kata dan kategori untuk memperkirakan probabilitas kategori pada suatu dokumen [4].

Banyaknya kata yang dapat mendefinisikan dan membentuk suatu dokumen menimbulkan masalah pada proses pengklasifikasian teks yaitu tingginya dimensi fitur. Dimensi fitur terdiri dari puluhan atau ratusan ribu fitur unik yang diambil dari dokumen *input* yang dapat tidak saling berhubungan. Permasalahan yang muncul akibat dimensi fitur yang besar pada kategorisasi teks dapat mengurangi kinerja klasifikasi. Untuk mencegah situasi ini, fitur yang diekstrak harus di *filter* sebelum fase klasifikasi untuk menyeleksi fitur yang paling relevan dan yang terbaik untuk mewakili suatu dokumen. Hal ini dilakukan dengan menghapus fitur *noninformative* dan membangun fitur set baru menggunakan metode seleksi fitur. Metode seleksi fitur yang digunakan dalam penelitian ini adalah *Information Gain* (IG). IG merupakan algoritma seleksi fitur yang efisien dalam mengukur jumlah bit informasi yang diperoleh untuk melakukan klasifikasi dengan mengetahui ada atau tidaknya fitur pada dokumen kemudian memilih subset optimal [5].

Melihat permasalahan yang ada, pada penelitian ini akan dibangun suatu aplikasi yang dapat mengklasifikasikan text bahasa Bali dengan metode *Information Gain* dan *Naïve Bayes Classifier.*

[1]*Staff UPT. Pusdalops PB BPBD Provinsi Bali, Mahasiswa Magister Teknik Elektro Bidang Studi Manajemen Sistem Informasi dan Komputer Universitas Udayana, Kampus Sudirman Denpasar Bali (telp.0361-239599; fax: 0361-239599; e-mail: guswid83@gmail.com)*

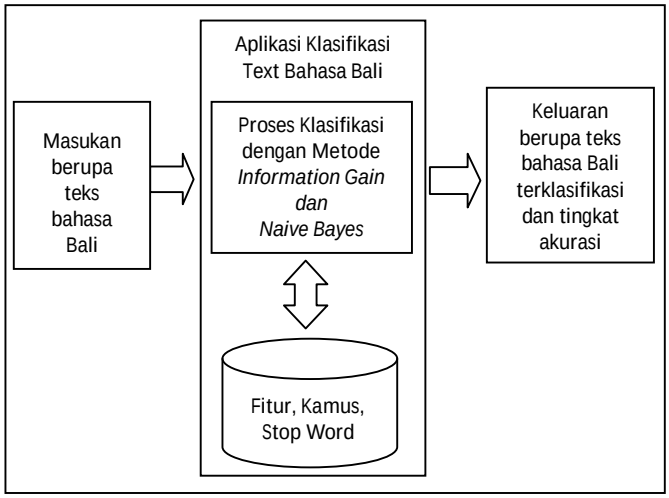
[2]*Staff Pengajar, Magister Teknik ElektroUniversitas Udayana, Kampus Sudirman Denpasar Bali (telp.0361-239599; fax: 0361-239599; e-mail: Msudarma@unud.ac.id)*

[3]*Staff Pengajar, Magister Teknik ElektroUniversitas Udayana, Kampus Sudirman Denpasar Bali (telp.0361-239599; fax: 0361-239599; e-mail: satya.kumara@unud.ac.id)*

II. METODE PENELITIAN

A. Gambaran Umum Sistem

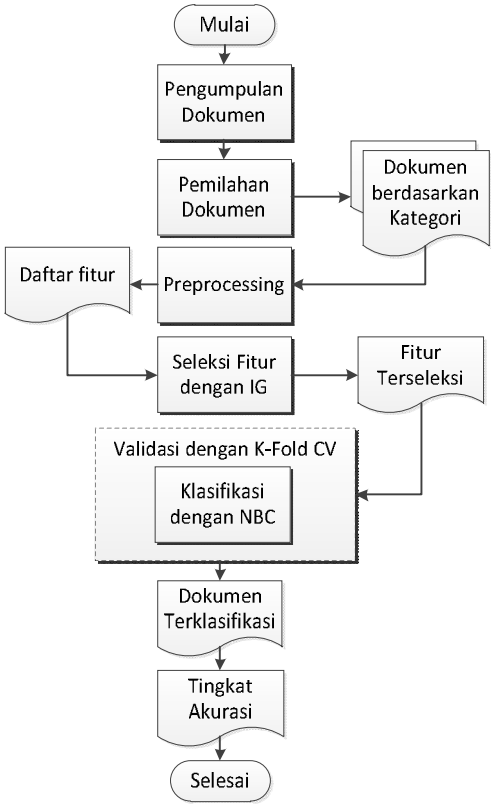
Gambaran umum sistem dari penelitian ini digambarkan dengan *Block Diagram* yang menggambarkan alur sistem secara umum yang berupa masukan, proses dan keluaran seperti terlihat pada gambar 1.



Gambar 1: Gambaran Umum Sistem

B. Alur Penelitian

Diagram alur penelitian secara umum dapat dilihat pada gambar 2.



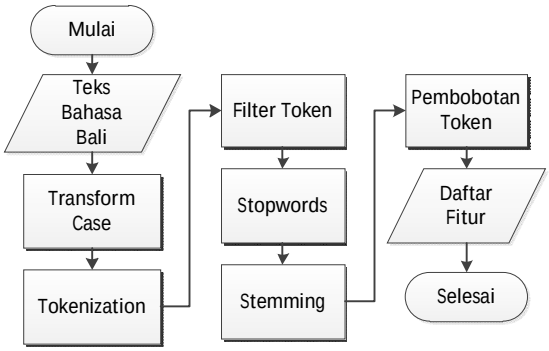
Gambar2: Alur Rancangan Penelitian

Penjelasan dari masing-masing tahapan pada alur penelitian adalah sebagai berikut:

1) Pengumpulan Dokumen: Pada tahap ini dilakukan pengumpulan teks bahasa Bali. Data teks yang digunakan diperoleh dari berita bahasa Bali pada Bali Orti Bali Post edisi minggu dari tahun 2010 sampai 2015.

2) Pemilahan Dokumen: Teks bahasa Bali yang diperoleh kemudian dipilah terlebih dahulu ke dalam kategori yang sudah ditentukan sebelumnya. Dalam penelitian ini ada dua kategori yang sudah ditentukan yaitu seni budaya dan upacara. Jumlah data yang digunakan sebanyak 100 untuk masing-masing ketegori. Setelah itu dokumen dalam kategori tersebut kembali dibagi menjadi dua yaitu sebagai data training dan data uji.

3) Preprocessing: Dalam proses *preprocessing* ini ada beberapa langkah yang harus dilakukan seperti terlihat pada gambar 3.



Gambar 3: Alur *Processing*

4) Seleksi Fitur dengan Information Gain: Hasil pembobotan token pada tahap *preprocessing* akan menjadi masukan untuk tahap seleksi fitur dengan IG. Perhitungan IG didefinisikan dengan rumus (1) [6].

$$IG(t) = -\sum_{i=1}^{m} P_\gamma(c_i) log\, P_\gamma(c_i) + P_\gamma(t) \sum_{i=1}^{m} P_\gamma(c_i|t) log P_\gamma(c_i|t) + P_\gamma(\bar{t}) \sum_{i=1}^{m} P_\gamma(c_i|\bar{t}) log P_\gamma(c_i|\bar{t}) \quad (1)$$

Di mana $P_\gamma(c_i)$ adalah probabilitas dari sebuah dokumen yang berada di label kelas, $P_\gamma(t)$ adalah probabilitas term *t* yang muncul didokumen, $P_\gamma(c_i|t)$ adalah probabilitas dari sebuah dokumen yang berada di label kelas mengingat bahwa term *t* yang muncul di dalam dokumen dan $P_\gamma(c_i|\bar{t})$ adalah probabilitas dokumen yang berada di label kelas mengingat bahwa term *t* tidakmunculdalamdokumen.

5) Klasifikasi dengan Naive Bayes Classifier: Pada tahap pelatihan, daftar fitur yang sudah terseleksi dan dokumen pelatihan digunakan sebagai masukan, kemudian dilakukan proses preprosesing pada dokumen pelatihan untuk mendapatkan frekuensi kemunculan kata pada daftar fitur dalam dokumen pelatihan. Hasil pembobotan ini akan digunakan untuk melakukan perhitungan statistik untuk mengetahui probabilitas sebuah fitur masuk kedalam kategori tertentu dengan menggunakan rumus (2) dan (3) [7].

$$P(v_j) = \frac{|Doc_j|}{|Contoh|} \quad (2)$$

$P(v_j)$ adalah nilai probabilitas ketegori *j*, dimana *Docj* adalah banyaknya dokumen yang memiliki kategori j dalam pelatihan, sedangkan Contoh banyaknya dokumen dalam contoh yang digunakan untuk pelatihan. Untuk nilai $P(a_i|v_j)$ yaitu probabilitas kata a_i dalam kategori *j* ditentukan dengan:

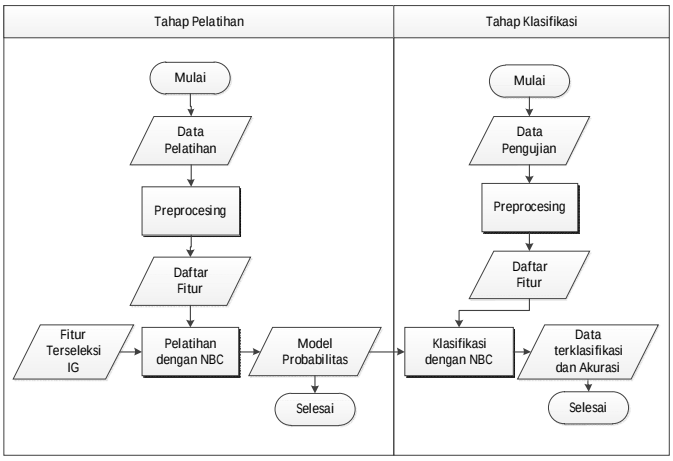
$$P(a_i|v_j) = \frac{|n_k + 1|}{|n + |vocabulary|} \quad (3)$$

Dimana n_k adalah frekuensi munculnya kata a_i dalam dokumen yang berkategori v_j, sedangkan nilai n adalah banyaknya seluruh kata dalam dokumen berkategori v_j, dan *vocabulary* adalah banyaknya kata dalam contoh pelatihan.

Model probabilitas ini akan digunakan sebagai masukan perhitunganpada tahap klasifikasi. Pada tahap klasifikasi dokumen uji dijadikan masukan untuk diklasifikasikan berdasarkan model probabilitas yang sudah dibuat sebelumnya pada tahap pelatihan menggunakan rumus (4) [7].

$$V_{MAP} = \underset{v_j \in V}{argmax}\, P(v_j) \prod_i P(a_i \mid v_j) \quad (4)$$

Keluaran dari tahap ini adalah dokumen yang sudah terklasifikasi dan tingkat akurasi klasifikasi. Alur proses pengklasifikasian menggunakan algoritma NBC disajikan pada Gambar 4.



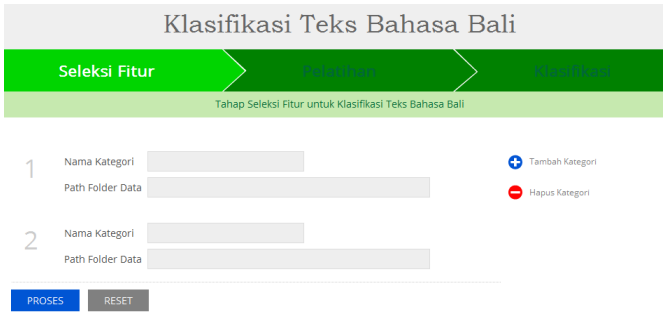
Gambar4: Alur Proses Klasifikasi dengan NBC

III. PENGEMBANGAN APLIKASI

Berikut ini adalah hasilpengembangan aplikasi yang dilakukan pada penelitian ini.

A. *Antar Muka Sistem*

Antar muka ini berfungsi untuk memudahkan pengguna untuk memasukkan inputan dan melihat hasil dari setiap tahapan proses yang dilakukan pada penelitian ini. Antar muka dari aplikasi ini dapat dilihat pada Gambar 5.



Gambar 5: Antar Muka Sistem

B. *Preprocessing*

Setiap tahapan dalam penelitian ini harus melewati proses *preprocessing* dengan tahapan sebagai berikut:

1) Transform Case: pada tahap ini aplikasi akan mengubah huruf besar yang ada pada dokumen menjadi huruf kecil.

2) Tokenization: pada tahap ini karakter selain huruf seperti tanda baca, angka dihilangkan selanjutnya dilakukan pemecahan kalimat menjadi potongan kata-kata atau token yang terpisah.

3) Filter Token (by Length): pada tahap ini akan dihilangkan kata-kata yang kurang dari 4 huruf dan lebih dari 20 huruf.

4) Stopwords: tahap ini merupakan tahap lanjutan untuk membuang kata-kata yang tidak berhubungan dengan dokumen. Kata tersebut bias berupa kata sambung atau keterangan seperti kata *"anggén","sané", "ring", "miwah", "puniki", "olih"*.

5) Stemming: pada tahap ini kata yang dihasilkan pada tahap sebelumnya akan diubah ke dalam bentuk kata dasarnya. Proses *stemming* pada aplikasi ini menggunakan algoritma *porter stemming* yang telah diadaptasi. Dari percobaan yang dilakukan dengan melakukan *preprocessing* pada 200 dokumen teks bahasa Bali, tanpa menggunakan proses *stemming* dihasilkan 8.410 kata sementara dengan menggunakan proses *stemming* dihasilkan 6.279 kata. Jadi proses *stemming* mampu mengurangi dimensi fitur sebanyak 2.113 kata.

6) Pembobotan Token: pada tahap ini akan dihitung frekuensi kemunculan kata dan kemunculannya dalam kumpulan dokumen yang digunakan.

C. *Tahap Seleksi Fitur*

Pada tahapan seleksi fitur semua kata akandiberi bobot berdasarkan frekuensi kemunculan kata dan jumlah kemunculan kata pada kumpulan dokumen, kemudian dihitung nilai bobot kata berdasarkan perhitungan TFIDF dan IG. Setelah melakukan perhitungan akan ditentukan atau di seleksi kata mana saja yang akan digunakan untuk dijadikan fitur. Tahapan ini diawali dengan mengisi nama kategori dan path folder dari data yang akan digunakan seperti terlihat pada Gambar 6.

Gambar 6: Form Input Seleksi Fitur

Masing-masing kategori terdiri dari 100 dokumen teks bahasa Bali. Hasil dari tahapan seleksi fitur ini dapat dilihat pada Gambar 7. Jumlah fitur yang dihasilkan terdiri dari 6.297 kata yang sudah dihitung nilai bobotnya berdasarkan perhitungan TFIDF dan IG.

NO	KATA	TF	DF	TFIDF	Seni Budaya	Upacara	IG
1	karya	110	39	0.94	80	30	0.034
2	seni	353	61	2.675	338	15	0.207
3	ulangun	14	9	0.164	13	1	0.023
4	asri	21	15	0.223	13	8	0.002
5	lemah	74	25	0.704	53	21	0.021
6	pulo	13	10	0.15	10	3	0.028
7	rupa	112	65	0.833	61	51	0.002
8	batik	21	1	0.347	21	0	0.005
9	lukis	61	9	0.716	60	1	0.023
10	rumasuk	48	34	0.425	39	9	0.054

Gambar 7: Hasil Seleksi Fitur

Keterangan :

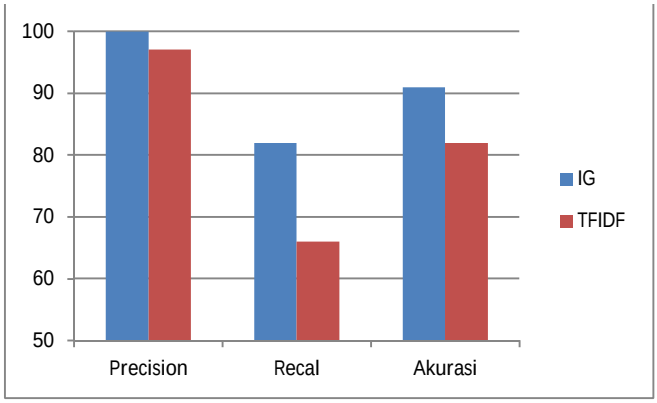
TF : Frekuensi kemunculan kata.
DF : Frekuensi dokumen dimana kata tersebut muncul.
TFIDF : Nilai bobot kata berdasarkan perhitungan TFIDF.
Seni Budaya : Frekuensi kemunculan kata pada kategori seni budaya.
Upacara : Frekuensi kemunculan kata pada kategori upacara.
IG : Nilai bobot kata berdasarkan perhitungan IG.

Untuk mengetahui hasil optimasi yang dihasilkan menggunakan metode seleksi fitur IG maka dilakukan pengujian untuk membandingkan hasil klasifikasi antara perhitungan TFIDF dan IG. Pengujian dilakukan dengan menggunakan 30 fitur dengan nilai bobot terbesar pada bobot IG dan TFIDF. Hasil dari pengujian ini diperlihatkan pada Tabel 1.

TABEL I.
HASIL PERBANDINGAN KLASIFIKASI DENGAN TFIDF DAN IG

Metode	Precision	Recall	Akurasi
TFIDF	97,059 %	66 %	82 %
IG	100 %	82 %	91 %

Grafik perbandingan hasil seleksi fitur menggunakan TFIDF dan IG ditunjukkan oleh Gambar 8.



Gambar 8: Grafik Hasil Perbandingan Klasifikasi dengan TFIDF dan IG

D. Tahap Pelatihan

Tahap pelatihan diawali dengan memilih metode pembobotan dan jumlah fitur yang akan digunakan. Setelah itu diinputkan path folder dokumen pelatihan untuk masing-masing kategori. Form input dari tahapan pelatihan dapat dilihat pada Gambar 9.

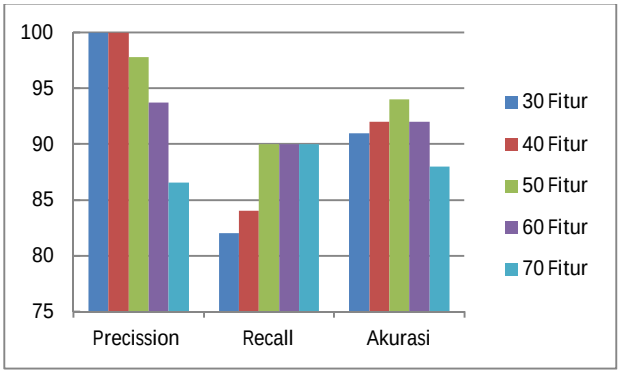
Gambar 9: Form Input Pelatihan

Pada penelitian ini jumlah fitur atau kata yang akan digunakan pada tahap pelatihan adalah 50 kata yang memiliki nilai bobot IG terbesar. Pemilihan jumlah fitur ini berdasarkan percobaan yang telah dilakukan sebelumnya dengan hasil seperti pada Tabel 2.

TABEL II.
HASIL PERCOBAAN UNTUK MENENTUKAN JUMLAH FITUR

Percobaan	Jumlah Fitur	Precision	Recall	Akurasi
1	30	100 %	82 %	91 %
2	40	100 %	84 %	92 %
3	50	97,826 %	90 %	94 %
4	60	93,75 %	90 %	92 %
5	70	87 %	90 %	88 %

Dapat dilihat dari kelima percobaan yang dilakukan, untuk jumlah fitur sebanyak 50 fitur memiliki nilai akurasi paling tinggi yaitu sebesar 94%. Sehingga dapat dijadikan dasar untuk menentukan jumlah fitur yang akan digunakan untuk tahap pelatihan pada aplikasi ini. Untuk gambaran lebih jelas mengenai perbandingan kelima percobaan diatas dapat dilihat dalam grafik pada Gambar 10.



Gambar 10: Grafik Perbandingan Berdasarkan Jumlah Fitur

E. Tahap Klasifikasi

Tahap ini menggunakan nilai probabilitas dari 50 fitur yang sudah ditentukan pada tahapan pelatihan untukdigunakansebagaimasukanperhitunganuntukmengklasifikasikandokumen. Keluaran tahap klasifikasi ini adalah dokumen yang sudah diklasifikasikan dan dilengkapi dengan nilai Precision, Recal dan Akurasi serta lama waktu yang dibutuhkan untuk melakukan klasifikasi. Nama file yang diawali dengan huruf "SB" sudah diklasifikasikan secara manual sebagai dokumen dengan kategori Seni Budaya sedangkan nama file yang diawali dengan "UP" diklasifikasikan sebagai kategori Upacara. Kolom hasil merupakan hasil dari klasifikasi yang dihasilkan oleh sistem. Hasil dari tahap klasifikasi ini dapat dilihat pada Gambar 11.

Gambar 11: Hasil Tahap Klasifikasi

Dari Gambar 11 dapat dilihat hasil klasifikasi dari 50 dokumen yang diuji menghasilkan nilai *Precission* 95,652 %, *Recall* 88 % dan Akurasi 92 %. Untuk dokumen yang tidak mengandung kata yang terdapat dalam daftar fitur yang terseleksi maka hasil klasifikasi dari sistem akan menunjukkan "Dokumen Tidak Termasuk Kategori".

IV. PENGUJIAN

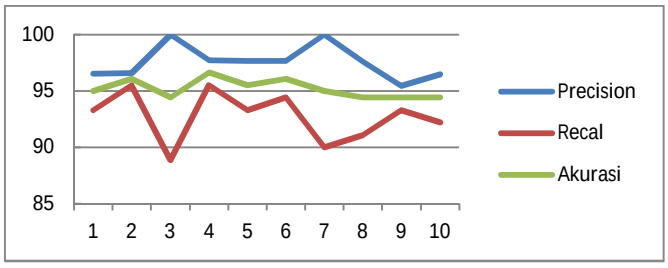
A. Cross Validation

Pengujian dilakukan dengan cross validation sebanyak 10 kali dengan membagi 100 dokumen pada masing-masing kategori menjadi 10 data set. Setiap *fold* akan menggunakan 10 dokumen untuk tahap pelatihan dan 90 dokumen untuk tahap klasifikasi pada masing-masing kategori. Output dari pengujian ini berupa tingkat persentase dari Precision, Recal dan Akurasi. Hasil dari Pengujian mengunakan 10 *fold cross validation* dapat dilihat pada Tabel 3.

TABEL III.
HASIL 10 *FOLD CROSS VALIDATION*

Fold	Precision	Recall	Akurasi
1	96,552 %	93,333 %	95 %
2	96,629 %	95,556 %	96,111 %
3	100 %	88,889 %	94,444 %
4	97,727 %	95,556 %	96,667 %
5	97,674 %	93,333 %	95,556 %
6	97,701 %	94,444 %	96,111 %
7	100 %	90 %	95 %
8	97,619 %	91,111 %	94,444 %
9	95,455 %	93,333 %	94,444 %
10	96,512 %	92,222 %	94,444 %
Rata-rata	**97,587 %**	**92,778 %**	**95,222 %**

.Dapat dilihat pada Tabel 3 nilai precision, recal dan akurasi antara setiap *fold* memiliki nilai yang berdekatan dengan nilai rata-rata precision 97,587 %, recal 92,778 %dan akurasi sebesar 95,222 %. Gambaran grafik perbandingan nilai Precision, Recal dan Akurasi pada setiap *fold* diperlihatkan pada gambar 12.



Gambar 12: Grafik Tingkat *Precision*, *Recal* dan Akurasi pada setiap *fold*

B. Kecepatan

Untuk mengetahui kecepatan yang dibutuhkan untuk menyelesaikan semua proses dari awal sampai sebuah dokumen diklasifikasikan maka pada setiap tahap penelitian akan dicatat lama proses yang dibutuhkan. Spesifikasi komputer yang digunakan adalah Notebook ASUSX450C dengan *processor* Intel Core i3, RAM 2 GB dan HDD 500 GB. Proses dilakukan dengan kondisi normal yaitu program aplikasi perkantoran seperti Ms. Word, Ms. Exel dan Internet berjalan seperti biasa. Berdasarkan pengujian yang dilakukan didapat hasil seperti pada Tabel 4.

TABEL IV.
LAMA WAKTU PROSES SETIAP TAHAP

No	Tahap	Jumlah Dokumen	Waktu Proses
1	Seleksi Fitur	100	9 menit 5 detik
2	Pelatihan	50	37 detik
3	Klasifikasi	50	36 detik
Total Waktu			10 menit 18 detik

Dari Tabel 4 dapat dilihat proses seleksi fitur membutuhkan waktu yang paling lama. Hal ini disebabkan

karena jumlah dokumen yang diolah cukup banyak, tahapan perhitungan pembobotan yang rumit dan proses penyimpanan ke database. Tetapi tahapan ini cukup dilakukan satu kali saja dalam setiap pembangunan model klasifikasi yang diinginkan. Sementara waktu yang dibutuhkan untuk tahap pelatihan dan klasifikasi hampir sama dan cukup singkat. Rata-rata waktu yang dibutuhkan pada tahap pelatihan dan klasifikasi kurang dari 1 detik per dokumen. Pada saat aplikasi ini siap digunakan, pengguna cukup menjalankan tahap klasifikasi saja.

V. KESIMPULAN

Penelitian ini telah menghasilkan aplikasi yang dapat mengklasifikasikan dokumen teks bahasa Bali ke dalam ketegori yang ditentukan menggunakan metode *Information Gain* dan *Naive Bayes Classiffier.* Dari hasil pengujian yang dilakukan dengan 10 *fold cross validation* disimpulkan bahwa performa metode NBC dalam mengklasifikasikan cukup akurat dan konsisten. Hal ini ditunjukkan dengan rata-rata nilai *precision, recall* dan akurasi yang dihasilkan diatas 90%.

REFERENSI

[1] APJII, “Profil Pengguna Internet Indonesia 2014”, APJII, 2015.

[2] Rozaq A., Arifin A.Z. dan Purwitasari,“Klasifikasi Dokumen Teks Berbahasa Arab Menggunakan Algoritma Naive Bayes”, ITS Surabaya, 2011.

[3] Feldman R. dan Sanger J.,“The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data”, Cambridge University Press, 2007.

[4] Chy A. N., Seddiqui M.H., dan Das S.,“Bangla news classification using Naive Bayes Classifier”, Computer and Information Technology (ICCIT), 2014, p. 596-615.

[5] Hong Z., Yong R., dan Xue Y, “Research on Text Feature Selection Algorithm Based on Information Gain and Feature Relation Tree”,Web Information System and Application Conference (WISA), 2000, paper 11.3.4, p. 109.

[6] Hatta H. R., Arifin A. Z dan Yuniarti A., “Metode Hibridasi Ant Colony Optimization Dan Information Gain Untuk Seleksi Fitur Pada Dokumen Teks Arab”, SCAN , 2013, Vol 8 no 2.

[7] Hamzah A., “Klasifikasi Teks Dengan Naïve Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita dan Abstract Akademis”, Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST), 2012, Periode 3.