Jurnal CoreIT, Vol.3, No.2, Desember 2017
ISSN 2460-738X (Print)
ISSN 2599-3321 (Online)

Penerapan Metode *Winnowing Fingerprint* dan *Naive Bayes* untuk Pengelompokan Dokumen

Adi Radili[1], Suwanto Sanjaya[2]
[1,2]Teknik Informatika UIN Sultan Syarif Kasim Riau
Jl. H.R. Soebrantas no. 155 KM. 18 Simpang Baru, Pekanbaru 28293
radiliadi@gmail.com[1], suwantosanjaya@uin-suska.ac.id[2]

***Abstrak* – Keanekaragaman dokumen teks serta jumlahnya saat ini terus bertambah yang menyebabkan penumpukan dokumen. Dokumen yang tersebar dan tidak terkoordinasi dengan baik akan menyulitkan pencari informasi dalam mendapatkan informasi yang diinginkan, maka perlu dibuatnya suatu sistem yang dapat mengelompokkan dokumen. Penelitian ini menerapkan metode *winnowing* untuk pemilihan fitur yaitu *fingerprint* dan *naive bayes* untuk pengelompokan. Pengelompokan dokumen dengan menggunakan *winnowing fingerprint* dan *naive bayes* mempunyai 8 bidang keahlian dengan menggunakan 1050 dokumen abstrak dengan 90% data latih dan 10% data uji. Pengujian menghasilkan akurasi 40% (*k-gram*=3, bilangan prima=2 dan jumlah *window*=8), 49,52% (*k-gram*=5, bilangan prima=2 dan jumlah *window*=8), 84,76% (*k-gram*=8, bilangan prima=2 dan jumlah *window*=8) dan 67,61% (*k-gram*=12, bilangan prima=2 dan jumlah *window*=8). Sedangkan pengujian menggunakan data yang seimbang, yaitu 400 data latih (masing-masing kelas memiliki 50 dokumen) menghasilkan akurasi 20% (*k-gram*=3, bilangan prima=2 dan jumlah *window*=8), 27,5% (*k-gram*=5, bilangan prima=2 dan jumlah *window*=8), 70% (*k-gram*=8, bilangan prima=2 dan jumlah *window*=8) dan 47,5% (*k-gram*=12, bilangan prima=2 dan jumlah *window*=8). Konfigurasi *winnowing* dengan nilai *k-gram*=8, bilangan prima=2 dan jumlah *window*=8 akan menghasilkan ciri dokumen yang terbaik untuk pengelompokan dokumen.**

Kata kunci – *Text Mining*, *Winnowing*, *Naive Bayes*, *Fingerprint*, Pengelompokan Dokumen

PENDAHULUAN

Saat ini, perkembangan teknologi cukup pesat sehingga menyebabkan banyak informasi tidak lagi disimpan dalam bentuk dokumen yang diletakkan di dalam lemari (*hardcopy*). Informasi tersebut kini disimpan dalam bentuk digital (*softcopy*). Hal ini mempermudah dalam penyimpanan dan dalam memperbanyak suatu dokumen.

Keanekaragaman dokumen teks serta jumlahnya saat ini terus bertambah yang menyebabkan penumpukan dokumen. Dokumen yang tersebar dan tidak terkoordinasi dengan baik akan menyulitkan pencari informasi dalam mendapatkan informasi yang diinginkan. Sebagai contoh, pencari informasi ingin mencari materi tugas akhir yang berhubungan dengan tema penelitian. Lalu, pencari informasi melakukan pencarian berdasarkan nama dokumen yang diduga memiliki keterkaitan dengan tema penelitian tersebut. Setelah melakukan pencarian, beberapa dokumen yang didapat tidak sesuai dengan yang diharapkan. Beberapa dokumen tersebut berisi hal yang tidak memiliki keterkaitan dengan tema penelitian yang dicari. Salah satu solusi yang dapat digunakan untuk mengatasi masalah ini adalah dengan menggunakan metode yang mampu mengklasifikasikan dokumen teks berdasarkan kesamaan isi dokumen.

Teknik klasifikasi adalah suatu proses untuk mengelompokkan sejumlah data ke dalam kategori tertentu yang sudah diberikan berdasarkan kesamaan sifat dan pola yang terdapat dalam data tersebut. Dokumen dapat dikelompokkan berdasarkan kesamaan isi yang terdapat dalam dokumen tersebut dengan membandingkan dengan kelompok-kelompok dokumen yang telah ada.

Penelitian yang berhubungan tentang pengelompokan dokumen sebelumnya oleh [8] menggunakan algoritma *winnowing fingerprint* dan *k-nearest neighbour* sebagai metode pengelompokan dokumen. Algoritma *winnowing* telah memenuhi kebutuhan sebuah algoritma pendeteksian kesamaan dokumen yaitu, dalam melakukan pencocokan terhadap dokumen tidak terpengaruh oleh spasi, jenis huruf, tanda baca dan karakter lainnya, atau biasa disebut dengan *whitespace insensitivity*. Algoritma *winnowing* digunakan untuk mencari *fingerprint* yang dihasilkan sistem, sehingga apabila karakter kata yang muncul sebagai *fingerprint* antar dokumen tidak sama, maka proses klasifikasi tidak relevan, persentase akurasi yang didapat adalah 80%. Pada penelitiannya [5] menyimpulkan bahwa dalam mendeteksi kesamaan dokumen, Algoritma Winnowing lebih baik daripada Algoritma Manber karena memberikan jaminan terdeteksinya dokumen yang sama dan terdapatnya informasi posisi

fingerprint pada dokumen. Penelitian yang dilakukan oleh [7] berhasil mendapatkan nilai bilangan prima terbaik adalah bilangan terkecil yaitu 2, sedangkan nilai terbaik untuk window adalah 8.

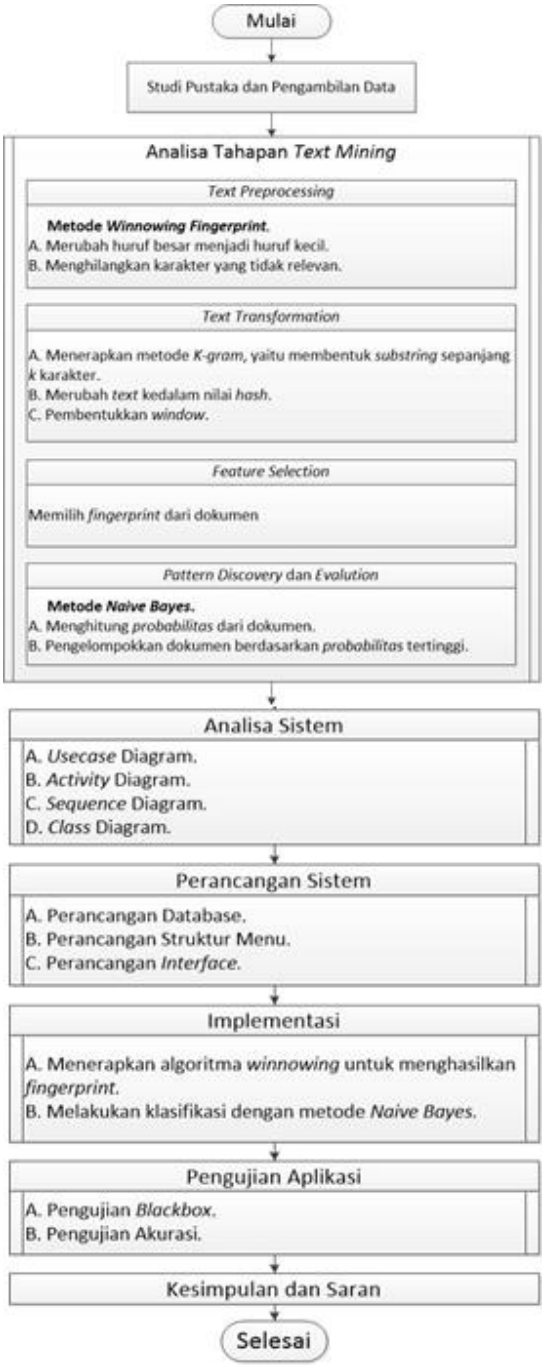
Sedangkan metode *naive bayes* telah diteliti sebelumnya oleh [13] yang berhubungan dengan pengelompokan dokumen teks berita berbahasa indonesia. Semakin besar ukuran data latih atau pembelajaran yang digunakan dengan fitur kata yang semakin banyak, cenderung makin tinggi nilai akurasi dalam pengklasifikasian dokumen, persentase yang didapat adalah 31,25% untuk dokumen tanpa *sub parent* dan *parent category* maksimal 34,37% untuk klasifikasi dokumen menggunakan *sub parent* dan *parent category*. Penelitian juga dilakukan oleh [1] yang menyimpulkan bahwa metode *naive bayes* lebih baik daripada metode *k-nearest neighbor* dengan perbandingan *f-measure* 89% dan 63%. Berdasarkan penelitian yang telah disebutkan diatas, maka disimpulkan bahwa metode *naive bayes* mempunyai tingkat akurasi yang lebih baik daripada *k-nearest neighbor*. Berdasarkan permasalahan dan penelitian terkait, maka pada penelitian ini akan menerapkan algoritma *winnowing* sebagai penghasil *fingerprint* dari setiap dokumen dan metode *naive bayes* untuk mengelompokkan dokumen.

Penelitian ini dibatasi pada:

a. Data yang digunakan berupa dokumen teks digital yang mempunyai format *plaintext* (txt).
b. Data latih yang digunakan berasal dari situs ITS (Institut Teknologi Sepuluh Nopember) *Digital Repository*, Fakultas Teknologi Informasi, Jurusan Teknik Informatika.
c. Bagian dalam abstrak yang digunakan berupa judul, konten abstrak, dan kata kunci (*keyword*).

METODOLOGI PENELITIAN

Tahapan penelitian yang dilakukan dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

HASIL DAN PEMBAHASAN

A. Analisis Masalah

Dokumen digital saat ini terus bertambah dari sisi jumlah maupun keanekaragaman yang menyebabkan sulit untuk mencari dokumen yang mempunyai kesamaan isi ataupun tema yang sama. Pencarian dokumen secara manual dengan menggunakan nama *file* dokumen sering mendapatkan hasil yang tidak berhubungan atau tidak relevan. Hal itu dikarenakan nama *file*

dokumen belum tentu menggambarkan isi dari dokumen, sehingga . Oleh karena itu, sangat penting untuk mengelola dan mengelompokkan dokumen. Dokumen yang telah diberi label atau dikelompokkan berdasarkan isi karakter yang ada di dalamnya akan mempermudah dalam proses pencarian dokumen.

B. Analisis Kebutuhan Data

Pada Pada tahapan analisa akan kebutuhan data penelitian dilakukan analisa terhadap data yang dibutuhkan pada pembuatan aplikasi klasifikasi dokumen. Berikut adalah data-data yang dibutuhkan :

1. Dokumen Latih.
   Merupakan data koleksi yang berisi kumpulan dokumen teks berekstensi .txt yang telah mempunyai kelompok atau telah diberi label. Setiap dokumen latih akan diproses ke dalam tahapan *text mining* sampai menghasilkan *output* berupa *fingerprint*, yang akan dijadikan acuan sebagai klasifikasi dokumen.
2. Dokumen Uji.
   Merupakan dokumen yang akan ditentukan jenis kelompok atau labelnya berdasarkan *fingerprint* dokumen latih.
3. Konfigurasi *Winnowing*.
   Merupakan penentuan nilai yang digunakan dalam algoritma *winnowing*. Berdasarkan penelitian yang dilakukan oleh (Ridho, 2013), bilangan basis prima=2 dan jumlah *window*=8 akan menghasilkan ciri dokumen yang lebih baik, sehingga seluruh proses algoritma *winnowing* yang ada akan menggunakan konfigurasi tersebut. Penentuan nilai *k-gram* dilakukan beberapa pengujian dengan nilai *k-gram*=3, *k-gram*=5, *k-gram*=8 dan *k-gram*=12.

C. Batasan Implementasi

Batasan implementasi pada aplikasi pengelompokan dokumen teks ini adalah sebagai berikut:

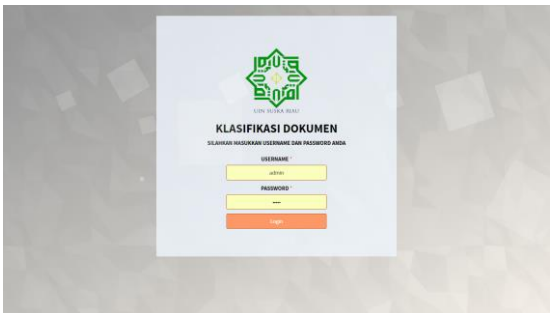
a. Proses pelabelan atau penentuan kelas pada data latih dilakukan secara manual berdasarkan subjek yang tertera di situs *Digital Repository* ITS dan ciri utama bidang keahlian di situs Jurusan Teknik Informatika Fakultas Teknologi Informasi ITS.
b. Pengambilan data uji dilakukan secara acak dan manual.

D. Implementasi Antarmuka Aplikasi

Setelah tahap analisa dan perancangan selesai dibuat, selanjutnya adalah tahap implementasi terhadap rancangan *interface* yang telah dibuat.

1. Halaman *Login*

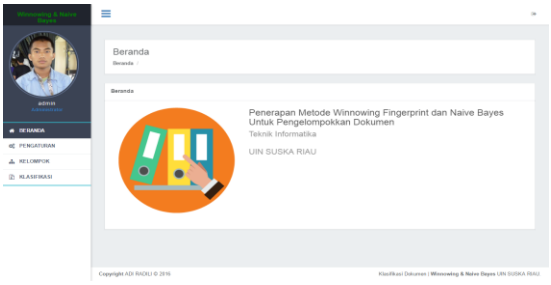
Berikut ini adalah tampilan dari halaman *login* dari sistem klasifikasi dokumen yang merupakan tampilan awal sebelum melakukan akses ke sistem, pada halaman ini terdapat *username* dan *password* sebagai syarat untuk melakukan akses ke sistem untuk tampilan dapat dilihat pada Gambar 2.



Gambar 2 Halaman *Login*

2. Halaman Beranda

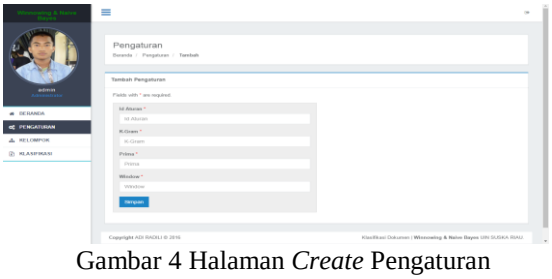
Berikut ini adalah tampilan dari halaman beranda dari sistem klasifikasi dokumen yang merupakan tampilan dari sistem setelah berhasil *login* dengan *username* dan *password* yang benar, dapat dilihat pada Gambar 3.



Gambar 3 Halaman Beranda

3. Halaman *Create* Pengaturan

Berikut ini adalah tampilan dari halaman *create* pengaturan dari sistem klasifikasi dokumen yang merupakan tampilan dari sistem untuk menambahkan data pengaturan, pada halaman ini terdapat kolom id aturan, k-gram, prima, window dan tombol Simpan untuk menyimpan data pengaturan ke dalam sistem. Untuk tampilan dapat dilihat pada Gambar 4.

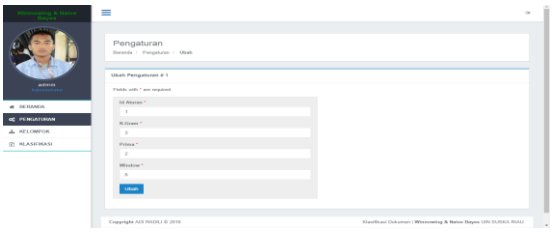


Gambar 4 Halaman *Create* Pengaturan

4. Halaman *Update* Pengaturan

Berikut ini adalah tampilan dari halaman *update* pengaturan dari sistem klasifikasi dokumen yang merupakan tampilan dari sistem untuk mengubah

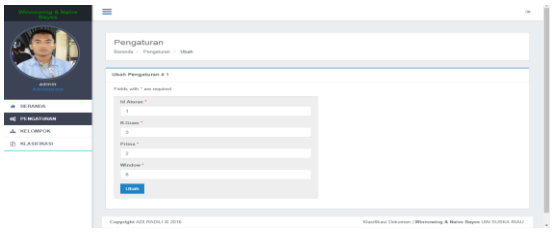
data pengaturan, pada halaman ini terdapat kolom aturan, k-gram, prima, window dan tombol Ubah untuk menyimpan data pengaturan yang telah diubah ke dalam sistem. Untuk tampilan dapat dilihat pada Gambar 5.



Gambar 5 Halaman *Update* Pengaturan

5. Halaman Kelompok

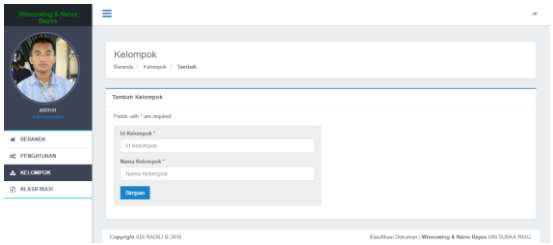
Berikut ini adalah tampilan dari halaman kelompok dari sistem klasifikasi dokumen, pada halaman ini dapat dilihat data kelompok yang telah disimpan dalam database dan untuk tampilan dapat dilihat pada Gambar 6.



Gambar 6 Halaman Kelompok

6. Halaman *Create* Kelompok

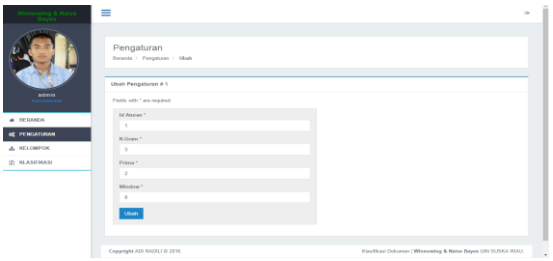
Berikut ini adalah tampilan dari halaman *create* kelompok dari sistem klasifikasi dokumen yang merupakan tampilan dari sistem untuk menambahkan data kelompok, pada halaman ini terdapat kolom id kelompok, nama kelompok dan tombol Simpan untuk menyimpan data kelompok ke dalam sistem. Untuk tampilan dapat dilihat pada Gambar 7.



Gambar 7 Halaman *Create* Kelompok

7. Halaman *Update* Kelompok

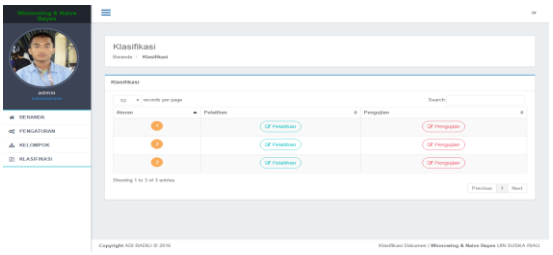
Berikut ini adalah tampilan dari halaman *update* kelompok dari sistem klasifikasi dokumen yang merupakan tampilan dari sistem untuk mengubah data kelompok, pada halaman ini terdapat kolom id kelompok, nama kelompok dan tombol Ubah untuk menyimpan data kelompok yang telah diubah ke dalam sistem. Untuk tampilan dapat dilihat pada Gambar 8.



Gambar 8 Halaman *Update* Kelompok

8. Halaman Klasifikasi

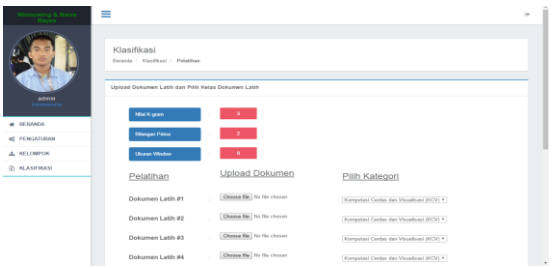
Berikut ini adalah tampilan dari halaman klasifikasi dari sistem klasifikasi dokumen. Pada halaman ini dapat dilihat pengaturan yang telah dibuat untuk proses pelatihan dan pengujian dokumen. Untuk tampilan dapat dilihat pada Gambar 9 dibawah ini.



Gambar 9 Halaman Klasifikasi

9. Halaman Pelatihan

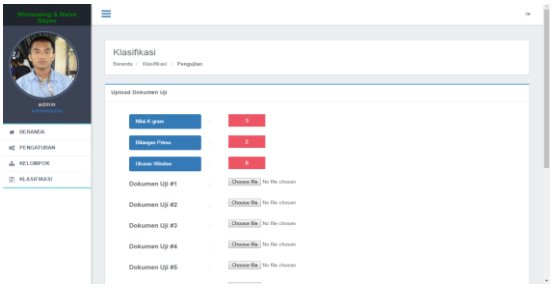
Berikut ini adalah tampilan dari halaman pelatihan dari sistem klasifikasi dokumen yang berisikan informasi nilai k-gram, bilangan prima, ukuran window. Pada halaman ini terdapat kolom *upload file* dokumen, jenis kelompok dokumen dan tombol Latih. Untuk tampilan dapat dilihat pada Gambar 10 dibawah ini.



Gambar 10 Halaman Pelatihan

10. Halaman Pengujian

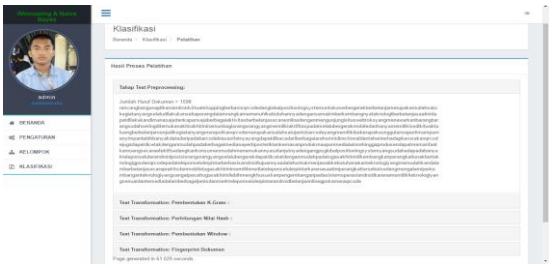
Berikut ini adalah tampilan dari halaman pelatihan dari sistem klasifikasi dokumen yang berisikan informasi nilai k-gram, bilangan prima, ukuran window. Pada halaman ini terdapat kolom *upload file* dokumen yang akan diuji dan tombol Uji. Untuk tampilan dapat dilihat pada Gambar 11.



Gambar 11 Halaman Pengujian

11. Halaman Hasil Pelatihan

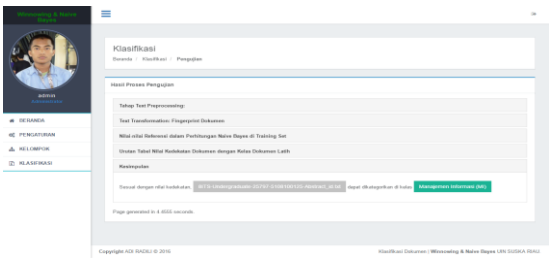
Berikut ini adalah tampilan dari halaman hasil pelatihan dari sistem klasifikasi dokumen. Informasi yang ditampilkan adalah tahap *preprocessing*, pembentukan *k-gram*, perhitungan nilai *hash*, pembentukan *window* dan *fingerprint* dokumen. Untuk tampilan dapat dilihat pada Gambar 12.



Gambar 12 Halaman Hasil Pelatihan

12. Halaman Hasil Pengujian

Berikut ini adalah tampilan dari halaman hasil pengujian dari sistem klasifikasi dokumen. Informasi yang ditampilkan adalah tahap *preprocessing*, text *transformation fingerprint* dokumen, nilai-nilai referensi dalam perhitungan *naive bayes* di data latih, urutan tabel nilai kedekatan dokumen dengan kelas dokumen latih dan kesimpulan. Untuk tampilan dapat dilihat pada Gambar 13.



Gambar 13 Halaman Hasil Pengujian

F. Pengujian

Pengujian sistem dilakukan untuk memeriksa tingkat akurasi kecocokan pengelompokan dokumen uji dengan prediksi awal kelompok pada dokumen uji tersebut. Misalkan dokumen uji 1 termasuk dalam kelompok Komputasi Cerdas Visualisasi (KCV), setelah dilakukan proses ekstraksi fitur dengan *winnowing fingerprint* dan klasifikasi dengan *naive bayes* maka hasil yang muncul adalah dokumen tersebut termasuk kedalam kelompok KCV juga.

Pengujian sistem termasuk juga pengujian program secara menyeluruh. Kesimpulan program yang telah diintegrasikan perlu diuji coba atau dilakukan *testing* sistem untuk melihat apakah sebuah program dapat menerima dengan baik, memproses dan memberikan keluaran program yang baik pula serta sesuai dengan harapan.

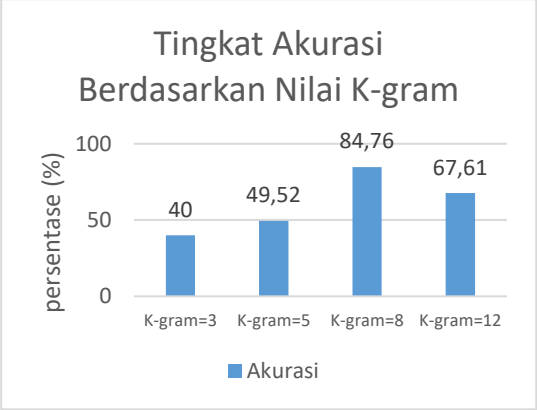
Pada tahap pengujian ini akan dilakukan beberapa hal sebagai berikut.

1. Pengujian dilakukan dengan menggunakan 8 (delapan) kelompok dokumen, yaitu Komputasi Cerdas dan Visualisasi (KCV), Komputasi Berbasis Jaringan (KBJ), Rekayasa Perangkat Lunak (RPL), Interaksi Grafika dan Seni (IGS), Arsitektur dan Jaringan Komputer (AJK), Algoritma dan Pemrograman (AP), Dasar dan Terapan Komputasi (DTK) dan Manajemen Informasi (MI).
2. Dokumen yang diuji adalah dokumen yang bukan termasuk data latih.
3. Pengujian dilakukan dengan 2 macam, yaitu dengan 945 data latih dan 105 data uji serta dengan data yang seimbang yaitu 400 data latih dan 50 data uji.
4. Pengujian dilakukan berdasarkan konfigurasi winnowing (skenario pengujian) pada Tabel 1.
5. Pengujian pada penelitian menggunakan metode *Blackbox* dan perhitungan tingkat akurasi keberhasilan.

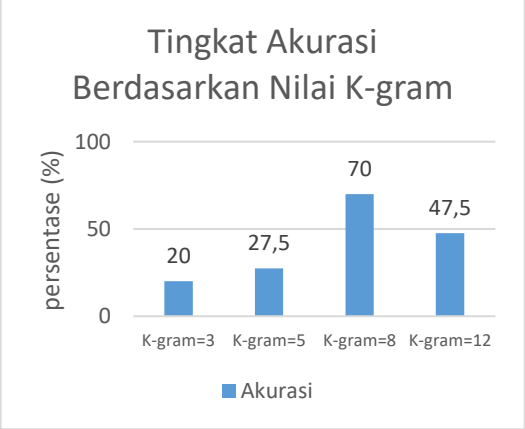
Tabel 1 Skenario Pengujian

Aturan	*K-gram*	Prima	*Window*
1	3	2	8
2	5	2	8
3	8	2	8
4	12	2	8

Pengambilan kesimpulan pengujian akurasi data uji dilakukan dengan membanding pengujian dengan tiga (3) buah nilai k-gram sesuai dengan skenario pengujian seperti pada Tabel 1, yaitu *k-gram*=8, *k-gram*=5 dan *k-gram*=3, tingkat akurasi berdasarkan nilai *k-gram* dapat dilihat pada grafik Gambar 14 dan Gambar 15.



Gambar 14 Grafik Tingkat Akurasi Berdasarkan Nilai K-gram 945 Data Latih



Gambar 15 Grafik Tingkat Akurasi Berdasarkan Nilai K-gram 400 Data Latih

Berdasarkan Gambar 14 dan Gabar 15 dapat dilihat pengujian dengan tingkat akurasi terbaik yaitu pengujian dengan 945 data latih, nilai *k-gram*=8 dengan akurasi 84,76%, *k-gram*=12 dengan akurasi 67,61%, nilai *k-gram*=5 dengan akurasi 49,52% dan akurasi terendah dengan nilai *k-gram*=3 dengan akurasi 40%, sedangkan pengujian dengan 400 data latih, nilai *k-gram*=8 dengan akurasi 70%, *k-gram*=12 dengan akurasi 47,5%, *k-gram*=5 dengan akurasi 27,5% dan akurasi terendah dengan nilai *k-gram*=3 dengan akurasi 20%.

KESIMPULAN DAN SARAN

Kesimpulan penelitian pengelompokan dokumen dengan menggunakan *winnowing fingerprint* dan *naive bayes* menggunakan 1050 dokumen abstrak dengan 90% data latih dan 10% data uji serta dengan data yang seimbang yaitu 400 data latih dan 40 data uji adalah sebagai berikut :

1. Akurasi terbaik adalah 84,76% dengan 945 data latih.
2. Konfigurasi *winnowing* dengan nilai *k-gram*=8, bilangan basis prima=2 dan jumlah *window*=8 menghasilkan ciri dokumen yang terbaik untuk pengelompokan dokumen.

Berdasarkan hasil yang diperoleh selama penelitian, saran pada penelitian selanjutnya yaitu:

1. Menerapkan metode lain untuk klasifikasi *text mining* seperti *logistic regression*, *support-vector machines* dan lain-lain, sehingga dapat diketahui metode mana yang menghasilkan akurasi yang lebih baik.
2. Menerapkan *filtering* pada proses *text preprocessing* untuk menghilangkan kata penghubung, sehingga *output* dari *text preprocessing* lebih relevan.

REFERENSI

[1] Anggono, R., Suryani, A. A., & Kurniati, A. P. (2009). *Analisis Perbandingan Metode K-Nearest Neighbor Dan Naive Bayes Classifier Dalam Klasifikasi Teks*. Universitas Telkom.

[2] Elbegbayan, N. (2005). *Winnowing , a Document Fingerprinting Algorithm*. *TDDC03 Projects*. Linkoping University.

[3] Han, J., Kamber, M., & Pei, J. (2006). *Data Mining. Concepts and Techniques*.

[4] Jurafsky, D., & Martin, J. H. (2015). Speech and Language Processing. In *Classification: Naive Bayes, Logistic Regression, Sentiment*.

[5] Kurniawati, A., & Wicaksana, I. W. S. (2008). Perbandingan Pendekatan Deteksi Plagiarism Dokumen Dalam Bahasa Inggris. In *KOMMIT 2008* (pp. 20–21). Depok.

[6] Kusrini, & Luthfi, Emha. (2009). *Algoritma Data Mining*. Yogyakarta : Penerbit Andi.

[7] Ridho, M. (2013). *Rancang Bangun Aplikasi Pendeteksi Penjiplakan Dokumen Menggunakan Algoritma Biword Winnowing*. Universitas Islam Negeri Sultan Syarif Kasim Riau.

[8] Sanjaya, S., & Absar, E. A. (2015). Pengelompokan Dokumen Menggunakan Winnowing Fingerprint dengan Metode K - Nearest Neighbour. *Jurnal CoreIT*, *1*(2), 50–56.

[9] Sathya, S., & Rajendran, N. (2015). A Review on Text Mining Techniques, *3*(5), 274–284.

[10] Schleimer, S., Wilkerson, D. S., Aiken, A., & Berkeley, U. C. (2003). Winnowing : Local Algorithms for Document Fingerprinting. *SIGMOD 2003*.

[11] Tan, A. (1999). *Text Mining : The state of the art and the challenges Concept-based*. Singapore.

[12] Xhemali, D., Hinde, C. J., & Stone, R. G.

(2009). Naïve Bayes vs . Decision Trees vs . Neural Networks in the Classification of Training Web Pages. *IJCSI International Journal of Computer Science Issues*, *4*(1), 16–23.

[13] Yanti, D. (2013). *Analisis Akurasi Algoritma Naive Bayes Pada Klasifikasi Dokumen Berkategori*. Universitas Sumatera Utara.