BAB 4
IMPLEMENTASI DAN PENGUJIAN

4.1 Implementasi

Tahap implementasi merupakan tahap pembuatan perangkat lunak, tahap lanjut dari tahap perancangan. Tahap yang dilakukan untuk menerjemahkan perancangan berdasarkan hasil analisis dalam bahasa yang dimengerti oleh komputer serta penerapan perangkat lunak pada keadaan yang sebenarnya.

4.1.1 Implementasi Perangkat Keras

Spesifikasi perangkat keras (*hardware*) yang digunakan untuk membangun sistem pendeteksi similaritas dokumen ini dapat dilihat pada tabel 4.1.

Tabel 4.1 Spesifikasi Perangkat Keras

No	Perangkat Keras	Spesifikasi
1	Prosesor	Kecepatan Minimal 2.2 GHz
2	Monitor	Monitor 14.1''
3	Memori	1024 MB

4.1.2 Implementasi Perangkat Lunak

Spesifikasi perangkat lunak (*software*) yang digunakan untuk membangun sistem pendeteksi similaritas dokumen ini dapat dilihat pada tabel 4.2.

Tabel 4.2 Spesifikasi Perangkat Lunak

No	Perangkat Lunak
1	Sistem operasi Windows XP/ 7

2	Visual Studio 2010
3	Notepad++
4	Wampserver
5	Mysql

4.1.3 Implementasi Database

Pembuatan basis data (*database*), dilakukan dengan menggunakan MySQL. Implementasi basis data (*database*) dalam bahasa SQL adalah seperti dapat dilihat pada tabel 4.3.

Tabel 4.3 Tabel Implementasi Database

No	Proses	SQL
1	Proses Pembuatan Database	CREATE DATABASE data_dokumen;
2	Tabel "tb_dokumen"	CREATE TABLE 'tb_dokumen'(id_dok int(10) NOT NULL AUTO_INCREAMENT, nama_dok varchar(10) NOT NULL, tanggal datetime NOT NULL, gram int(10) NOT NULL, window int(10) NOT NULL, file varbinary(50000) NOT NULL, fingerprints mediumblob NOT NULL, PRIMERY KEY('id_dok'))ENGINE=InnoDB;
3	Tabel "histori"	CREATE TABLE 'histori' (id_proses int (10) NOT NULL AUTO_INCREAMENT,

		nama_dokumen varchar(50) NOT NULL, gram int(10) NOT NULL, window int(10) NOT NULL, algoritma varchar(30) NOT NULL, persentasi varchar(30) NOT NULL, waktu_proses varchar(20) NOT NULL, PRIMERY KEY(‘id_proses’))ENGINE=InnoDB;
4	Tabel “tb_stopword”	CREATE TABLE ‘tb_stopword’ (id_stopword int(10) NOT NULL AUTO_INCREAMENT, stopword varchar(30) NOT NULL, PRIMERY KEY (‘id_stopword’))ENGINE=InnoDB;

4.1.4 Implementasi Antarmuka

Implementasi antarmuka dilakukan untuk mengimplementasikan setiap tampilan yang terdapat pada tahap perancangan yang telah dilakukan. Implementasi antarmuka dari sistem yang telah dibuat dapat dilihat pada tabel 4.4.

Tabel 4.4 Implementasi Antarmuka

No	Nama Antramuka	Proses	Nama File	Keterangan
1	Halaman Utama	Tampilan Awal	Form3.cs	Tampilan awal sistem
2	Algoritma Winnowing	Proses Algoritma Winnowing	Form1.cs	Tampilan dari seluruh proses Algoritma Winnowing

3	Algoritma Manber	Proses Algoritma Manber	Form2.cs	Tampilan dari seluruh proses Algoritma Manber
4	Histori	Proses Pengambilan Data Riwayat	Form4.cs	Tampilan dari riwayat proses yang pernah dilakukan Sistem
5	Form Tentang	Tampilan Teks	Form5.cs	Tampilan yang berisis teks sejarah pembuatan aplikasi
6	Form Bantuan	Tampilan Teks	Form6.cs	Tampilan yang berisi teks panduan penggunaan aplikasi

4.2 Pengujian Metode

Tahapan pengujian yang akan dilakukan terdiri dari pengujian dua metode yang telah diteliti yaitu pengujian metode Winnowing dan pengujian Manber. Pengujian dilakukan bertujuan untuk mengetahui performansi dari kedua algoritma pada sistem yang telah dibuat, pengujian meliputi pengujian jumlah gram, jumlah window, ukuran p, basis bilangan prima, waktu proses dan jumlah langkah yang dilakukan untuk menghasilkan output.

4.2.1 Pengujian Metode Winnowing

Pada pengujian metode winnowing dilakukan pengujian jumlah gram, jumlah window, basis bilangan prima ,waktu proses, perbandingan waktu proses menggunakan *tokenizing* dan penghapusan *stopword* dan jumlah langkah

4.2.1.1 Pengujian Jumlah Gram

Analisis jumlah gram dilakukan agar diketahui jumlah gram yang optimal untuk ketepatan dari perhitungan sistem. Gram yang akan dianalisis merupakan jumlah yang diinputkan oleh user. Berikut merupakan tabel hasil analisis dari jumlah gram yang telah dilakukan:

Jumlah window : 2
Basis : 3
Dokumen 1 : Sejarah Google.txt (5633 karakter)
Dokumen 2 : Sejarah Google 2.txt (4862 karakter)

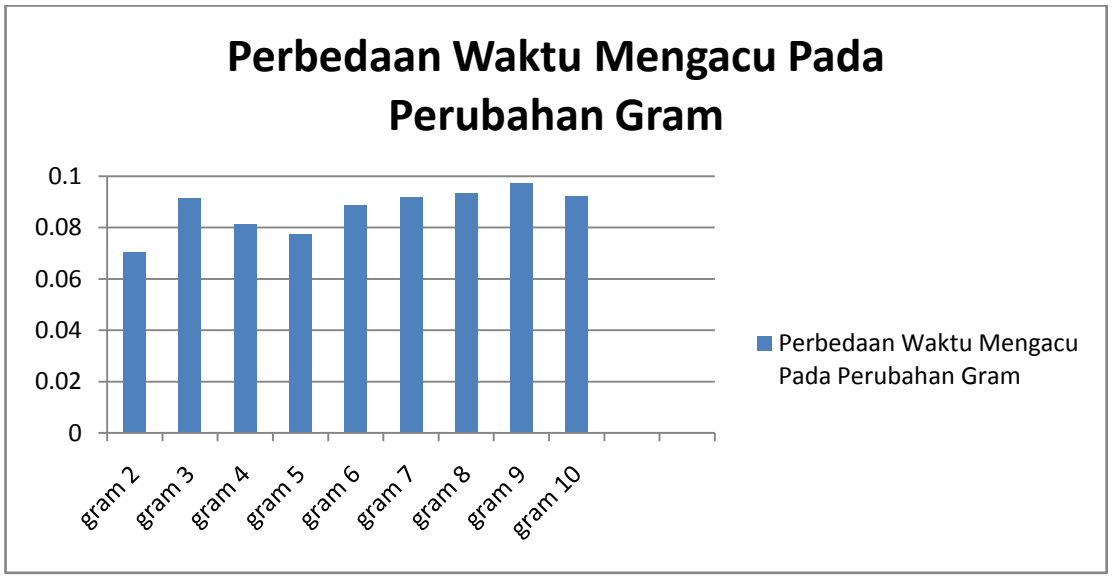
Dokumen 1 memiliki total 5633 karakter dan Dokumen 2 adalah dokumen yang sama namun dihilangkan tiga paragraph untuk membedakan isi dari dokumen sehingga Dokumen 2 memiliki total 4862 karakter. Menurut perhitungan manual persentasi kemiripan seharusnya menghasilkan hasil $\frac{5633}{4862} \times 100 = 86{,}31\ \%$, dan yang mendekati hasil tersebut adalah jumlah gram 7-9 namun gram = 9 lebih mendekati hasil yaitu sebesar 86,93 % dengan selisis 0,62 %. Pengujian jumlah gram dapat dilihat pada table 4.5:

Tabel 4.5 Analisis Gram Winnowing

Jumlah Gram	Waktu proses/second	Persentasi Kemiripan	Selisih
2	0.0703	100 %	13,69 %

3	0.0913	99,16 %	12,85 %
4	0.0814	98,10 %	11,79 %
5	0.0773	95,55 %	9,24 %
6	0.0888	91,77 %	5,46 %
7	0.0919	89,55 %	3,24 %
8	0.0933	88,27 %	1,96 %
9	**0.0973**	**86,93 %**	**0,62 %**
10	0.0923	100 %	13,69 %

Dari hasil analisis jumlah gram pada table 4.4 maka dapat disimpulkan bahwa semakin besar jumlah gram yang diinputkan maka waktu proses semakin meningkat dan persentasi kemiripan menjadi menurun. Input gram dengan nilai >= 10 dengan window = 2 akan menghasilkan nilai yang cukup besar (INF) pada sistem yang mengakibatkan kesalahan perhitungan yang menghasilkan output 100%. Perbedaan waktu proses ketika gram bertambah dapat dilihat pada gambar 4.1



Gambar 4.1 Grafik Perbedaan Waktu Proses Terhadap Gram Winnowing

4.2.1.2 Pengujian Jumlah Window

Analisis jumlah window dilakukan agar diketahui jumlah window yang optimal untuk ketepatan dari perhitungan sistem. Jumlah window yang akan dianalisis merupakan jumlah yang diinputkan oleh user. Karena jumlah gram optimal telah ditemukan pada proses analisis sebelumnya yaitu gram=9 maka jumlah gram yang akan dipakai untuk analisis window adalah gram=9 Berikut merupakan tabel hasil pengujian dari jumlah window yang telah dilakukan:

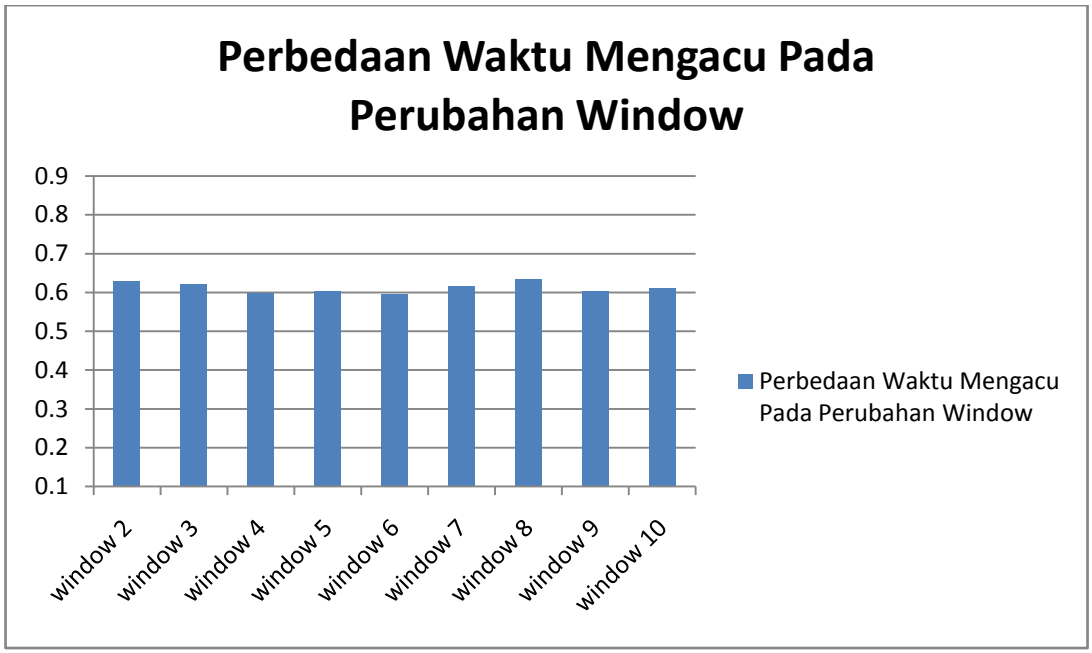
Jumlah gram : 9
Basis : 3
Dokumen 1 : Sejarah Google.txt (5633 karakter)
Dokumen 2 : Sejarah Google 2.txt (4862 karakter)

Tabel 4.6 Analisis Jumlah Window Winnowing

Jumlah Window	Waktu proses/second	Persentasi Kemiripan	Selisih
2	0,6306	86,93 %	0,62 %
3	0,6212	87,09 %	0,78 %
4	0,5970	87,09 %	0,78 %
5	0,6028	86,78 %	0,47 %
6	0,5967	87,00 %	0,69 %
7	0,6168	87,32 %	1,01 %
8	0,6351	86,99 %	0,68 %
9	**0,6034**	**86,72 %**	**0,41 %**
10	0,6116	86,79 %	0,48 %

Dokumen 1 memiliki total 5633 karakter dan Dokumen 2 adalah dokumen yang sama namun dihilangkan tiga paragraph untuk membedakan isi dari dokumen sehingga Dokumen 2 memiliki total 4862 karakter. Dari hasil analisis dapat

disimpulkan bahwa jumlah gram=9 dan window=9 menghasilkan 86,72 %. Persentasi tersebut adalah paling mendekati dengan jumlah persentasi perhitungan manual yaitu 86,31 %. Perbedaan waktu proses ketika jumlah window bertambah dapat dilihat pada gambar 4.2



Gambar 4.2 Grafik Perbedaan Waktu Proses Terhadap Window Winnowing

Dari gambar 4.2 dapat dilihat bahwa pembagian window tidak begitu mempengaruhi waktu proses dikarenakan pembagian window hanya membagi nilai-nilai *fingerprints* saja kedalam window-window, berbeda ketika pembagian gram, pembagian gram membagi dan melakukan perhitungan pada gram yang terbagi-bagi sehingga perbedaan gram mempengaruhi waktu proses.

4.2.1.3 Pengujian Basis Bilangan Prima

Analisis bilangan prima dilakukan untuk mendapatkan informasi jumlah dari maksimal bilangan prima yang dapat digunakan oleh Algoritma Winnowing menurut rumus yang dilakukan pada proses *Rolling Hash* yaitu:

$$C_1 * b^{k-1} + C_2 * b^{k-2} + \cdots + c_{k-1} * b + c_k$$

Keterangan:

c : nilai ascii karakter

b : basis (bilangan prima)

k : banyak karakter

Bilangan prima dalam rumus *Rolling Hash* tidak ditentukan oleh user tapi ditentukan oleh *programmer* atau pembuat sistem. Maka dilakukan analisis untuk mendapatkan nilai bilangan prima yang optimal untuk dapat diterapkan pada rumus tersebut. Berdasarkan analisis yang telah dilakukan bilangan prima yang ditentukan maka diperoleh basis bilangan prima yang paling optimal untuk jumlah gram k=9 dan window w=9 yaitu basis bilangan prima b=3. Seperti diketahui bahwa terdapat banyak bilangan prima yaitu 2,3,5,7,11,13,17,19,31 dst. Pemilihan basis bilangan prima = 3 dikarenakan jumlah k yang paling optimal pada pengujian sebelumnya adalah k=9 dapat menghasilkan jumlah yang sangat besar jika diterapkan pada rumus *Rolling Hash* sehingga sistem tidak dapat menampung nilai yang besar tersebut dan menghasilkan nilai INF yang akan mempengaruhi proses perhitungan.

4.2.1.4 Pengujian Winnowing Menggunakan Tokenizing dan Penghapusan Stopword

Analisis performansi algoritma Winnowing mengggunakan *tokenizing* dan penghapusan *stopword* (kata yang sering digunakan namun tidak memiliki makna) dilakukan untuk mendapatkan informasi perbedaan waktu dan ketepatan dari hasil analisis yang dilakukan tanpa menggunakan *tokenizing* dan penghapusan *stopword* sebelumnya. Penghapusan *stopword* berpengaruh pada kecepatan algoritma dalam mengolah teks karena kata-kata yang tidak memiliki makna dihapus sehingga waktu

yang diperlukan untuk pembagian gram dan window relatif lebih sedikit. Berikut ini merupakan hasil analisis yang telah dilakukan.

Dokumen 1 memiliki total 5633 karakter lalu dilakukan proses penghapusan *stopword* menjadi 3148 karakter dan Dokumen 2 adalah dokumen yang sama namun dihilangkan tiga paragraph untuk membedakan isi dari dokumen sehingga Dokumen 2 memiliki total 4862 karakter dan dilakukan penghapusan *stopword* menjadi 3654 karakter. Pengujian jumlah gram menggunakan *tokenizing* dan penghapusan *stopword* dapat dilihat pada tabel 4.7.

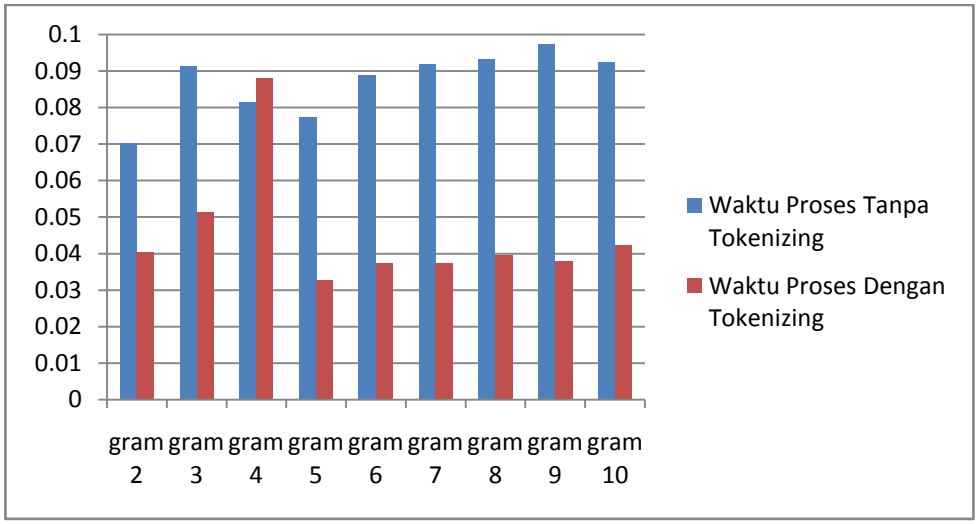
Jumlah window : 2
Basis : 3
Dokumen 1 : Sejarah Google.txt (3654 karakter)
Dokumen 2 : Sejarah Google 2.txt (3148 karakter)

Tabel 4.7 Analisis Gram Winnowing menggunakan Tokenizing

Jumlah Gram	Waktu proses/second	Persentasi Kemiripan
2	0.0404	98,33 %
3	0,0512	97,29 %
4	0,0880	95,57 %
5	0,0328	92,52 %
6	0,0375	89,58 %
7	0,0374	88,94 %
8	0,0396	87,94 %
9	**0,0380**	**85,95 %**
10	0,0423	100 %

Menurut perhitungan manual persentasi kemiripan seharusnya menghasilkan hasil $\frac{3148}{3654} \times 100 = 86{,}15\ \%$, dan yang mendekati hasil tersebut adalah jumlah

gram 7-9 namun gram =9 lebih mendekati hasil yaitu sebesar 85,95 %. Perbedaan waktu proses ketika menggunakan tokenizing terhadap perubahan gram dapat dilihat pada gambar 4.3



Gambar 4.3 Grafik Pengaruh Gram dan Tokenizing Terhadap Kecepatan

Begitu pula pada analisis window yang telah dilakukan dengan mengambil jumlah gram = 9 dan window = 5 mendapatkan hasil 86,12%. Hasil ini sangat mendekati nilai perhitungan manual. Pengujian dapat dilihat pada tabel 4.8.

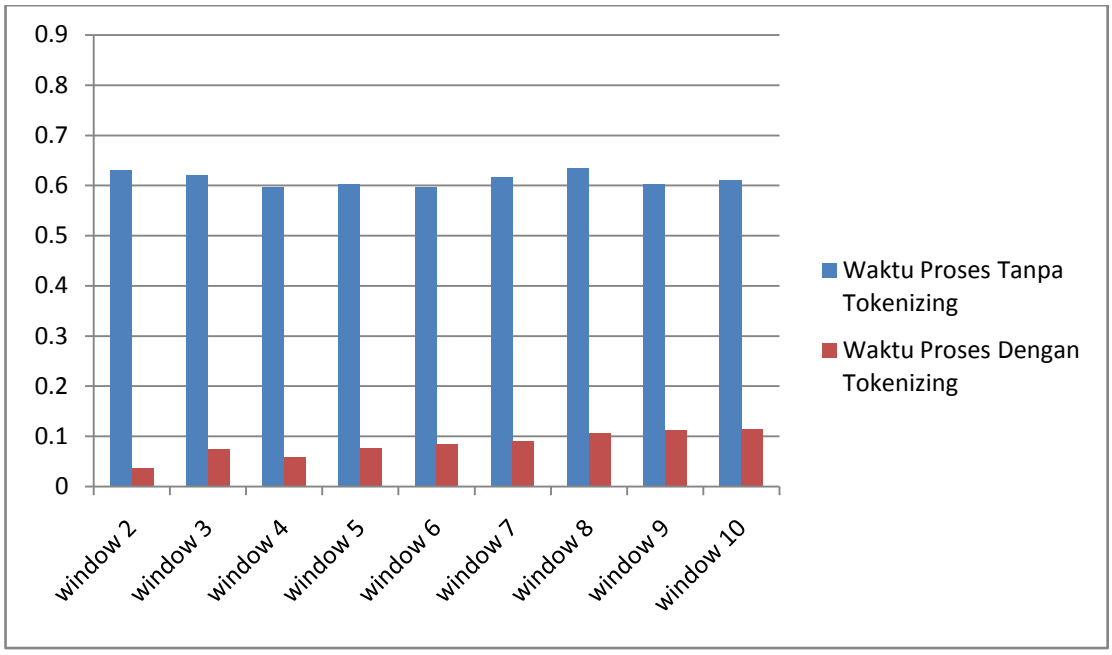
Jumlah gram : 9
Basis : 3
Dokumen 1 : Sejarah Google.txt (3654 karakter)
Dokumen 2 : Sejarah Google 2.txt (3148 karakter)

Tabel 4.8 Analisis Window Winnowing menggunakan Tokenizing

Jumlah window	Waktu proses/second	Persentasi Kemiripan
2	0.0380	85,95 %
3	0,0758	86,22 %

4	0,0592	86,24 %
5	**0,0759**	**86,12 %**
6	0,0844	86,29 %
7	0,0903	86,28 %
8	0,1076	86,44 %
9	0,1137	86,44 %
10	0,1145	86,46 %

Pengujian waktu proses terhadap perubahan window menggunakan tokenizing dapat dilihat pada gambar 4.4



Gambar 4.4 Grafik Pengaruh Window dan Tokenizing Terhadap Kecepatan

Dari hasil pengujian ini dapat disimpulkan bahwa waktu proses yang dilakukan semakin cepat dibandingkan dengan proses tanpa penghapusan *stopword*.

4.2.2 Pengujian Metode Manber

Pada pengujian metode manber dilakukan pengujian jumlah gram, jumlah ukuran p, basis bilangan prima ,waktu proses, perbandingan waktu proses menggunakan *tokenizing* dan penghapusan *stopword* dan jumlah langkah yang dilakukan

4.2.2.1 Pengujian Jumlah Gram Manber

Analisis jumlah gram dilakukan untuk mengetahui jumlah gram optimal yang dapat dipakai oleh Algoritma Manber. Berdasarkan pengujian yang telah dilakukan diperoleh data sebagai berikut:

Jumlah p : 3
Basis : 3
Dokumen 1 : Sejarah Google.txt (3654 karakter)
Dokumen 2 : Sejarah Google 2.txt (3148 karakter)

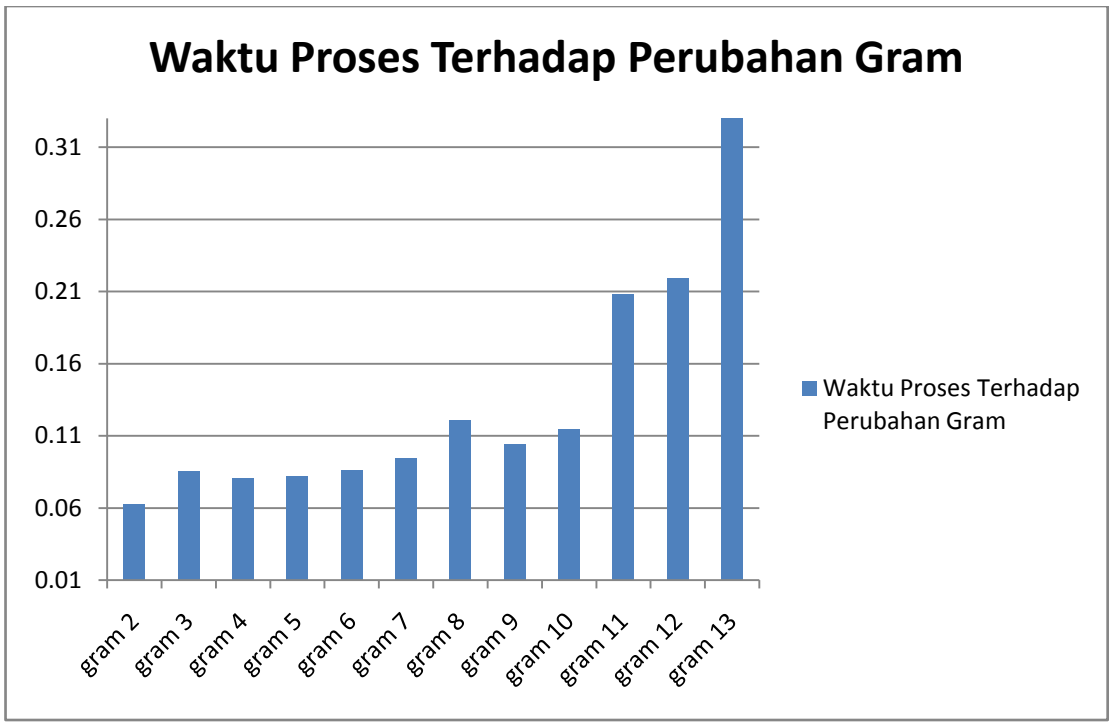
Menurut perhitungan manual persentasi kemiripan seharusnya menghasilkan hasil $\frac{3148}{3654} \times 100 = 86{,}15\ \%$, dan yang mendekati hasil tersebut adalah jumlah gram 8-13 namun yang paling mendekati adalah jumlah gram = 13

Tabel 4.9 Analisis Gram Manber

Jumlah gram	Waktu proses/second	Persentasi Kemiripan	Selisih
2	0,0627	100 %	13,85 %
3	0,0851	96,55 %	10,40 %
4	0,0807	95,37 %	9,22 %
5	0,0817	91,75 %	5,60 %
6	0,0859	90.07 %	3,92 %
7	0,0940	88,46 %	2,31 %

8	0,1208	87,60 %	1,45 %
9	0,1037	87,36 %	1,21 %
10	0,1142	87,32 %	1,17 %
11	0,2080	87,17 %	1,02 %
12	0,2190	87,07 %	0,92 %
13	**0,3336**	**86,98 %**	**0,83 %**

Dokumen 1 setelah melalui tahap preprocessing memiliki total 3654 karakter dan Dokumen 2 adalah dokumen yang sama namun dihilangkan tiga paragraph untuk membedakan isi dari dokumen sehingga Dokumen 2 memiliki total 3148 karakter. Perubahan waktu proses dari perubahan gram dapat dilihat pada gambar 4.5



Gambar 4.5 Grafik Waktu Proses Terhadap Gram Manber

Dilihat dari gambar 4.5 pada metode Manber menunjukan ketika gram bertambah maka waktu proses secara visual bertambah.

4.2.2.2 Pengujian Ukuran P Manber

Analisis jumlah ukuran P dilakukan untuk mengetahui jumlah ukuran p optimal yang dapat dipakai oleh Algoritma Manber. Berdasarkan pengujian yang telah dilakukan diperoleh data sebagai berikut:

Jumlah gram : 13
Basis : 3
Dokumen 1 : Sejarah Google.txt (3654 karakter)
Dokumen 2 : Sejarah Google 2.txt (3148 karakter)

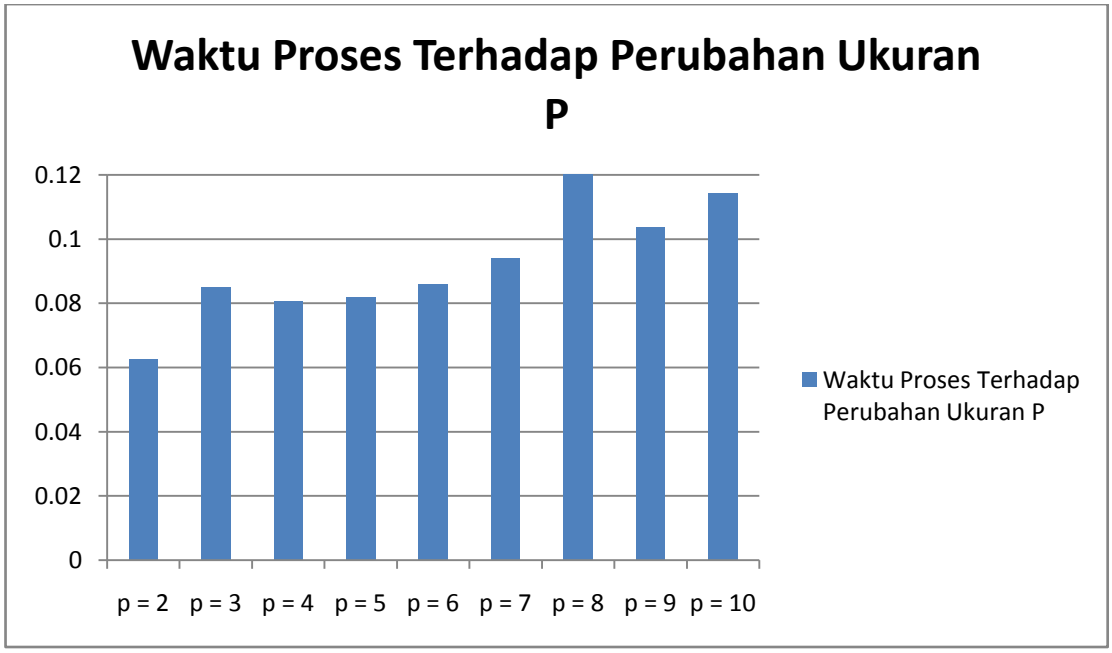
Menurut perhitungan manual persentasi kemiripan seharusnya menghasilkan hasil $\frac{3148}{3654} \times 100 = 86{,}15\ \%$, dan yang mendekati hasil tersebut adalah jumlah gram 10 dan ukuran p= 9 yaitu sebesar 85,71% dengan selisih 0,44 %.

Tabel 4.10 Analisis Ukuran p Manber

Jumlah ukuran p	Waktu proses/second	Persentasi Kemiripan	Selisih
2	0,0627	87,76 %	1,61 %
3	0,0851	87,32 %	1,17 %
4	0,0807	88,99 %	2,84 %
5	0,0817	87,67 %	1,52 %
6	0,0859	87,18 %	1,03 %
7	0,0940	87,98 %	1,83 %
8	0,1208	88,91 %	2,76 %
9	**0,1037**	**85,71 %**	**0.44 %**
10	0,1142	90,18 %	4,03 %

Dokumen 1 setelah melalui tahap preprocessing memiliki total 3654 karakter dan Dokumen 2 adalah dokumen yang sama namun dihilangkan tiga paragraph untuk membedakan isi dari dokumen sehingga Dokumen 2 memiliki total 3148 karakter. Perbandingan persentasi algoritma winnowing dan manber yaitu winnowing dapat menghasilkan 86,12% dengan gram=9 dan window = 5. Winnowing lebih mendekati dengan hasil perhitungan manual.

Waktu proses terhadap perubahan ukuran P dalam program menghasilkan grafik yang dapat dilihat pada gambar 4.6



Gambar 4.6 Grafik Waktu Proses Terhadap Gram Manber

Dari hasil pengujian ukuran p maka dapat disimpulkan bahwa perbedaan ukuran p mempengaruhi proses waktu dari sistem yang secara visual dapat dilihat ketika p bertambah maka waktu proses pun ikut bertambah

4.2.2.3 Pengujian Basis Bilangan Prima

Analisis bilangan prima dilakukan untuk mendapatkan informasi jumlah dari maksimal bilangan prima yang dapat digunakan oleh Algoritma Manber menurut rumus yang dilakukan pada proses *Rolling Hash* yaitu:

$$C_1 * b^{k-1} + C_2 * b^{k-2} + \cdots + c_{k-1} * b + c_k$$

Keterangan:
c : nilai ascii karakter
b : basis (bilangan prima)
k : banyak karakter

Bilangan prima dalam rumus Rolling Hash tidak ditentukan oleh user tapi ditentukan oleh programmer atau pembuat sistem. Maka dilakukan analisis untuk mendapatkan nilai bilangan prima yang optimal untuk dapat diterapkan pada rumus tersebut. Berdasarkan analisis yang telah dilakukan bilangan prima yang ditentukan maka diperoleh basis bilangan prima yang paling optimal untuk jumlah gram k=9 dan window w=9 yaitu basis bilangan prima b=3. Seperti kita ketahui bahwa terdapat banyak bilangan prima yaitu 2,3,5,7,11,13,17,19,31 dst. Pemilihan basis bilangan prima = 3 dikarenakan jumlah k yang paling optimal pada analisis sebelumnya adalah k=9 dapat menghasilkan jumlah yang sangat besar jika diterapkan pada rumus Rolling Hash sehingga sistem tidak dapat menampung nilai yang besar tersebut dan menghasilkan nilai INF yang akan mempengaruhi proses perhitungan.

4.3 Kesimpulan Pengujian

Berdasarkan pengujian yang telah dilakukan dapat ditarik kesimpulan bahwa:

1. Dalam rumus *Rolling Hash* jumlah gram mempengaruhi persentasi kemiripan

2. Jika jumlah gram semakin besar mengakibatkan pembagian kata semakin besar sehingga semakin besar gram maka semakin sedikit nilai hash yang dibandingkan
3. Semakin besar jumlah gram maka waktu yang diperlukan sistem semakin bertambah dikarenakan perulangan yang diperlukan untuk mengulang karakter semakin banyak
4. Pada Algoritma Winnowing jumlah window mempengaruhi persentasi kemiripan dikarenakan pembagian window sama seperti pembagian gram, yaitu membagi nilai hash menjadi beberapa window
5. Pada Algoritma Manber pemilihan ukuran *p* sebagai pemilihan *fingerprints* menciptakan kelonggaran data sehingga melewatkan data yang seharusnya bisa lebih baik dipilih seperti proses pembagian window pada winnowing.
6. Penggunaan *tokenizing* dan penghapusan *stopword* pada sistem membuat waktu proses lebih cepat
7. Gram dengan jumlah besar menimbulkan kesalahan persentasi perhitungan dikarenakan sistem tidak dapat menampung data perhitungan yang terlalu besar sehingga dibatasi jumlah gram >=2 dan gram <=10
8. Pemilihan *fingerprints* winnowing lebih mendekati ketepatan dari pada pemilihan *fingerprints* manber
9. Dari hasil analisis jumlah langkah atau eksekusi proses Algoritma Winnowing lebih banyak dari Jumlah langkah Manber sehingga dalam waktu kecepatan winnowing lebih lama dari pada manber.
10. Pengaruh bahasa pada algoritma hanya berpengaruh pada kecepatan dikarenakan pada proses *preprocessing* yaitu terdapat proses penghapusan kata *stopword* yang berbeda dari setiap bahasa.
11. Penggunaan rumus *Rolling Hash I* sudah cukup bisa digunakan tanpa Rumus *Rolling Hash II* dikarenakan dalam bahasa pemrograman terdapat perulangan yang bisa digunakan untuk mengulang seluruh kata yang terdapat pada sebuah

dokumen, sehingga jika satu rumus saja yang digunakan dapat menghemat proses yang dilakukan.

12. Penggunaan basis bilangan prima yang besar juga akan menghasilkan nilai *hash* yang besar sehingga alternatif basis bilangan prima yang digunakan yaitu b = 3
13. Nilai gram tidak boleh sama atau melebihi jumlah seluruh karakter pada dokumen