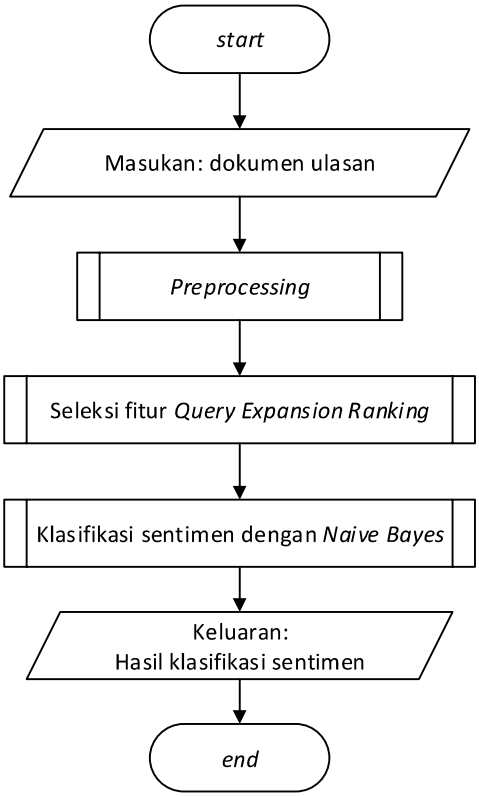
BAB 4 PERANCANGAN

Perancangan adalah perancangan sistem yang meliputi diagram alir, dan manualisasi penghitungan untuk membuat sistem tersebut.

4.1 Deskripsi Sistem

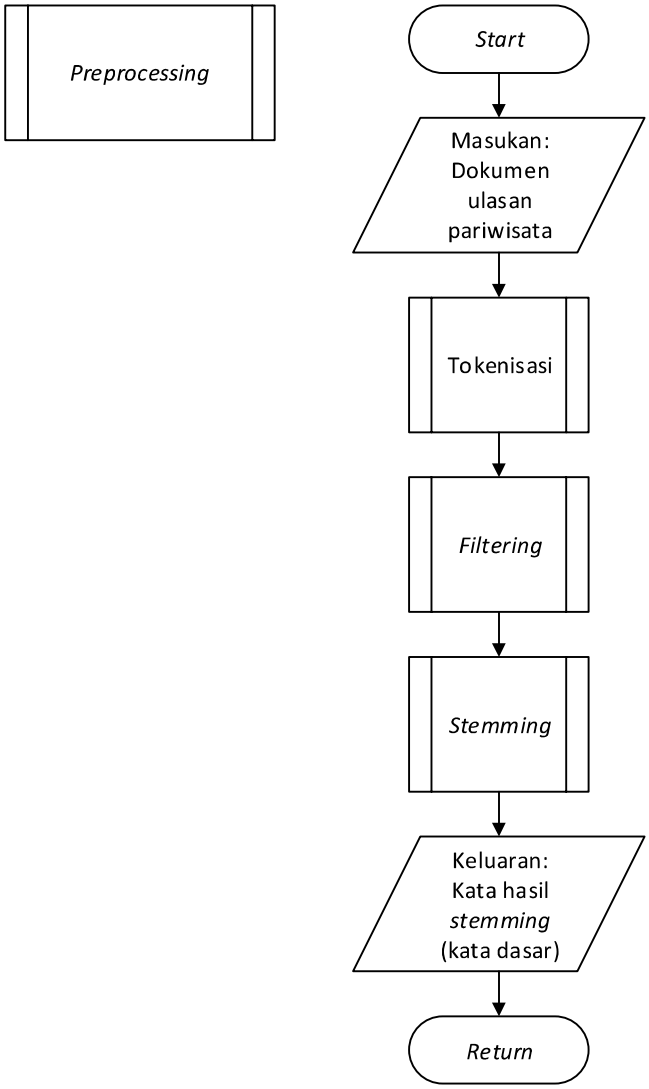
Sistem yang akan dibuat pada penelitian Analisis Sentimen Pariwisata di Kota Malang menggunakan Metode *Naive Bayes* dan Seleksi Fitur *Query Expansion Ranking* adalah sistem yang mampu memuat dokumen dan melakukan serangkaian proses sehingga dokumen tersebut bisa diklasifikasikan menjadi dua kelas, positif dan negatif. Sistem ini memiliki beberapa tahap dari awal hingga selesai, yaitu *preprocessing*, seleksi fitur¸ dan klasifikasi. Diagram alir dari sistem ini ditunjukkan dengan Gambar 4.1.



Gambar 4.1 Diagram alir sistem

4.2 Preprocessing

Preprocessing adalah tahap untuk memperoleh dokumen yang siap untuk diolah ke proses selanjutnya. Proses ini terbagi menjadi tiga tahap yaitu tokenisasi, *filtering,* dan *stemming.* Setelah proses ini selesai maka dokumen tersebut hanya akan terdapat *term* atau fitur yang berisi kata dasar saja yang merupakan representasi dokumen. Gambar 4.2 menunjukkan diagram alir dari *preprocessing.*

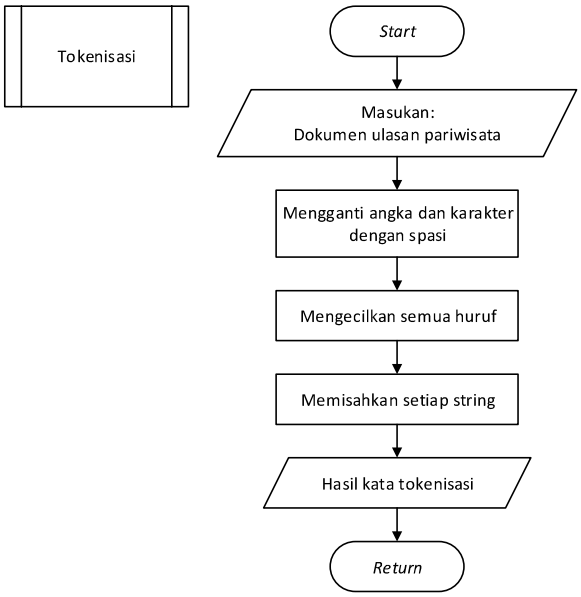


Gambar 4.2 Diagram alir *prepocessing*

Saat melakukan proses *preprocessing* maka sistem memasukkan data berupa dokumen ulasan pariwisata lalu dokumen tersebut diproses tokensasi, *filtering*, dan *stemming* sehingga menghasilkan dokumen yang siap diproses.

4.2.1 Tokenisasi

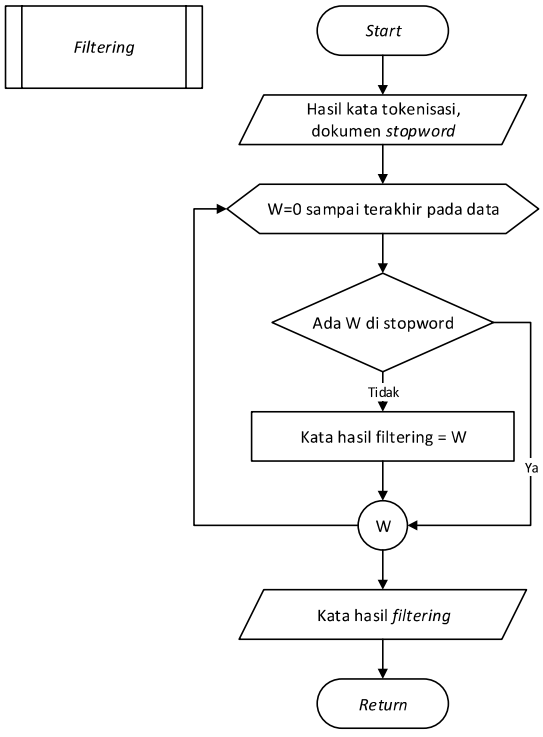
Proses ini adalah proses untuk menghilangkan hal-hal yang dianggap dari data seperti karakter dan angka, serta membuat data menjadi huruf kecil semua supaya data yang akan diolah bisa lebih mudah untuk digunakan. Proses dari tokenisasi digambarkan dengan diagram alir pada Gambar 4.3.



Gambar 4.3 Diagram alir tokenisasi

4.2.2 *Filtering*

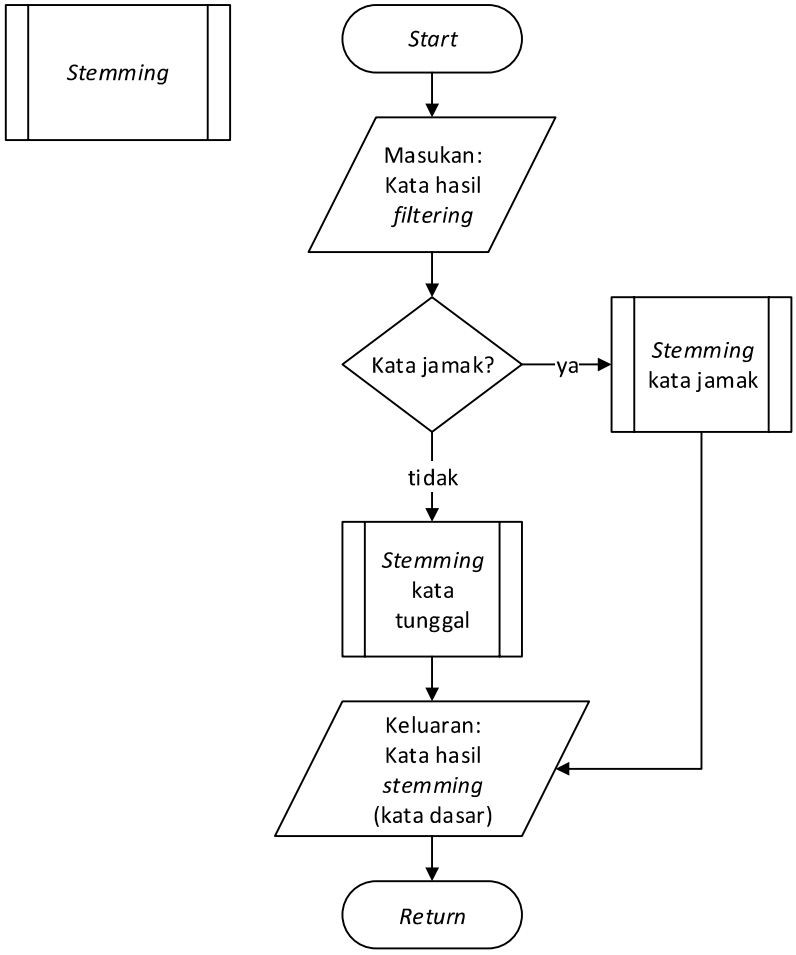
Filtering pada tahap ini merupakan penghilangan kata-kata *stopword* atau kata-kata yang sudah tidak berguna bagi sebuah dokumen. Proses filtering pada sistem ini digambarkan dengan diagram alir pada Gambar 4.4.



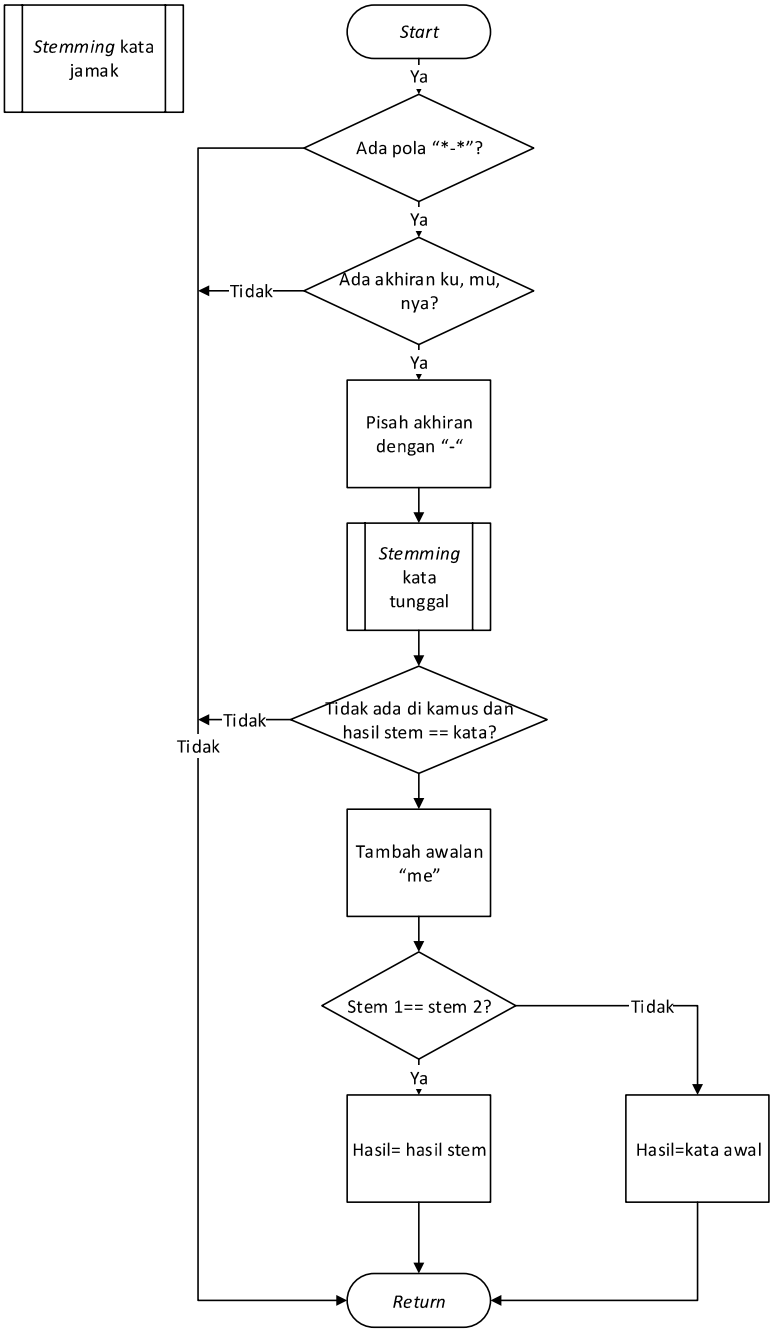
Gambar 4.4 Diagram alir *filtering*

4.2.3 *Stemming*

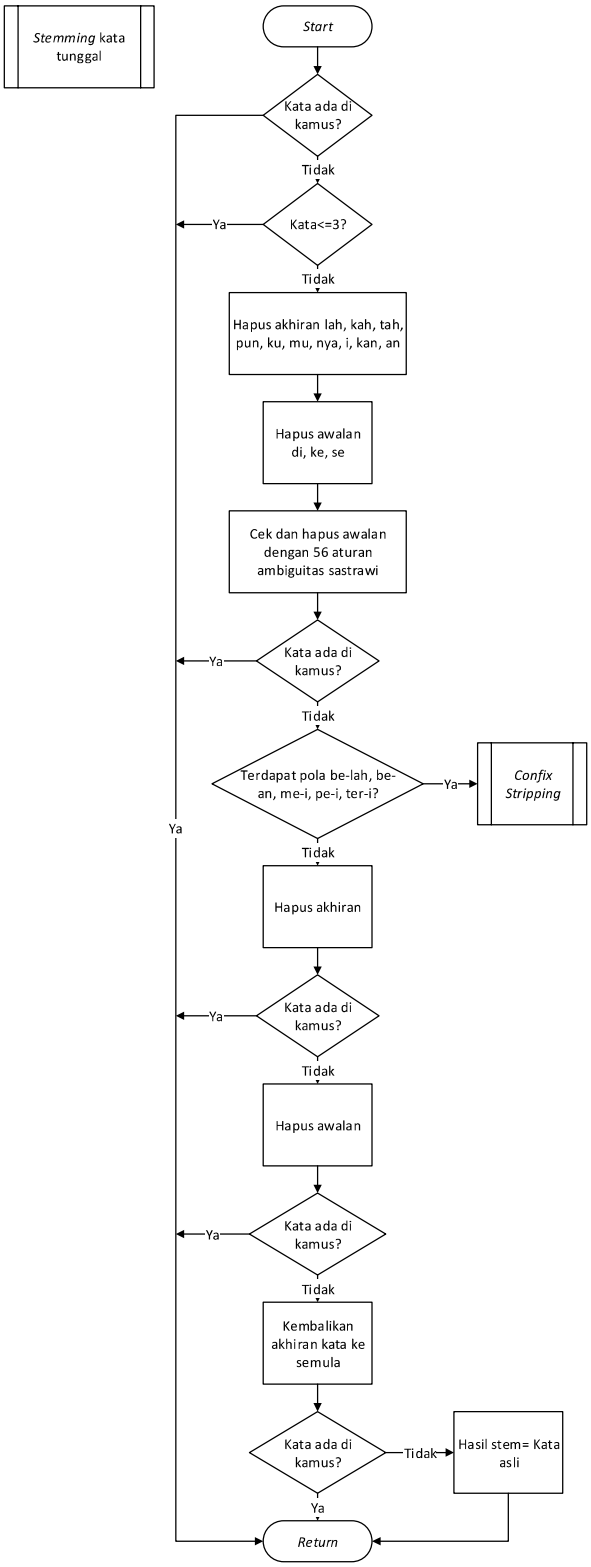
Stemming merupakan proses untuk mengubah kata hasil proses *filtering* ke kata dasarnya dengan menghilangkan imbuhan dan sisipan. Pada sistem ini, proses *stemming* dilakukan dengan menggunakan *library* Sastrawi yang didapatkan dari website GitHub. Proses pada *stemming* dengan library ini menggunakan algoritme Nazief dan Adriani yang ditingkatkan lagi dengan algoritme *Confix Stripping, Enhanced Confix Stripping*, dan algoritme *Modified Enhanced Confix Stripping* untuk Bahasa Indonesia (Librian, 2015). Proses dari *stemming* ini digambarkan dengan diagram alir pada Gambar 4.5.



Gambar 4.5 Diagram alir *stemming*



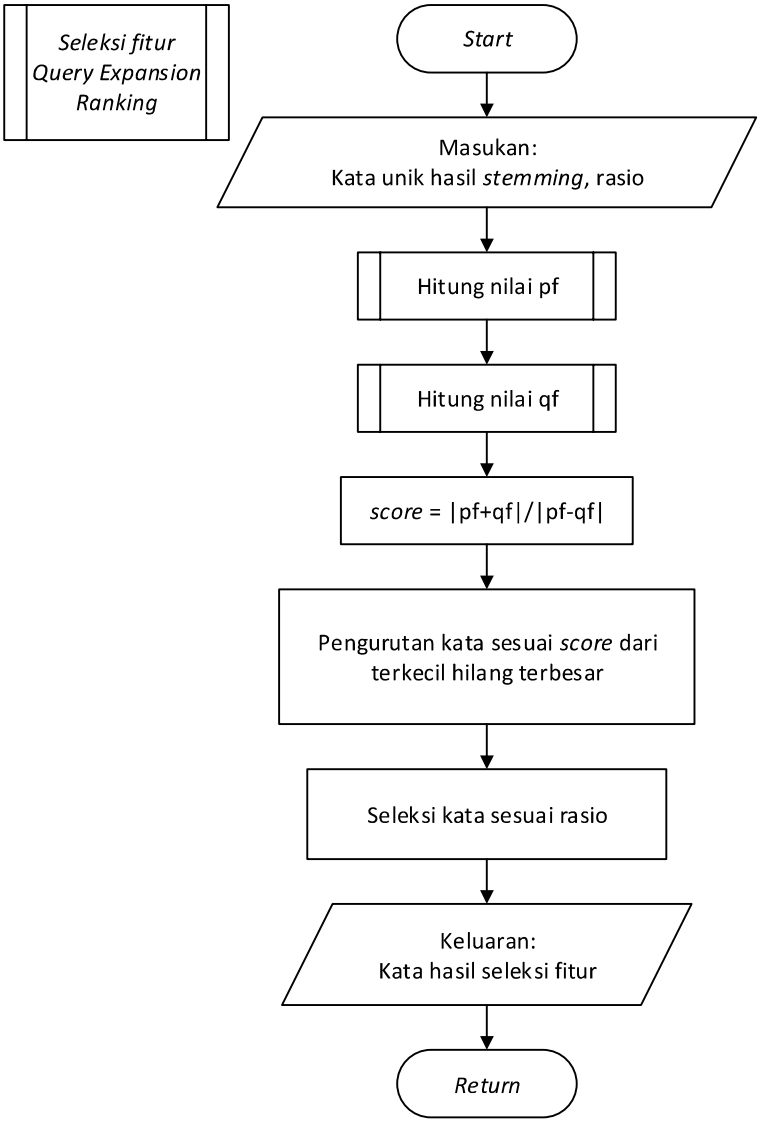
Gambar 4.6 Diagram alir *stemming* kata jamak pada *library* sastrawi



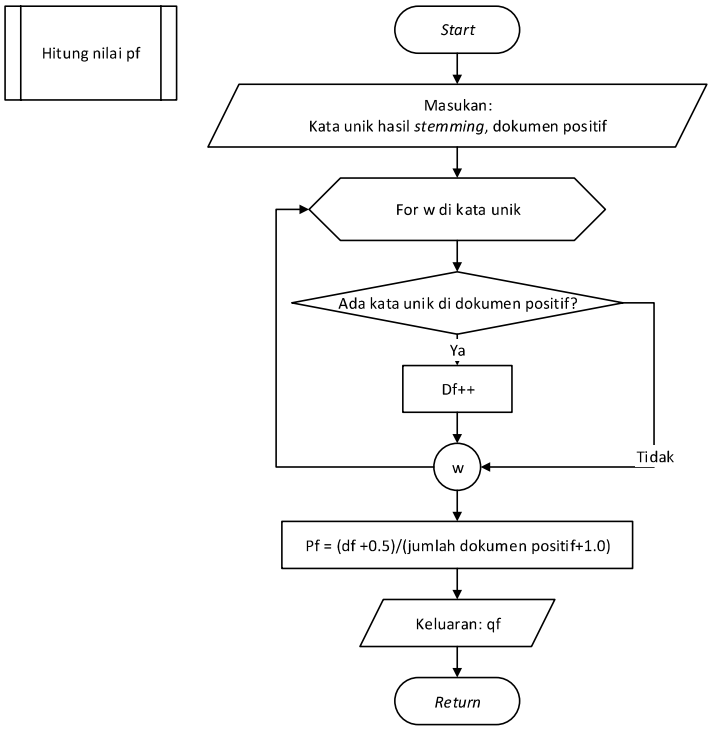
Gambar 4.7 Diagram alir *stemming* kata tunggal

4.3 Seleksi Fitur *Query Expansion Ranking*

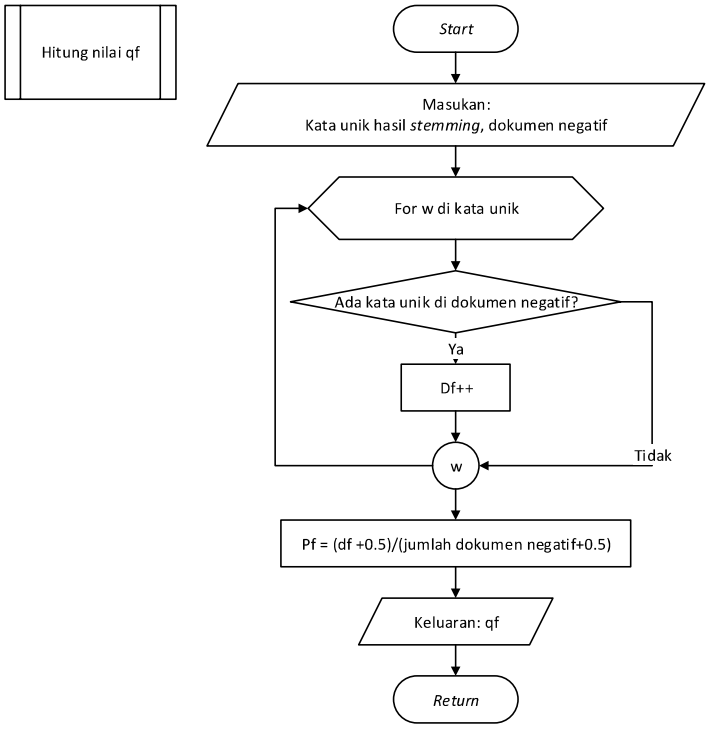
Seleksi fitur *Query Expansion Ranking* bekerja dengan cara menghitung jumlah dokumen yang mengandung kata terpilih. Caranya dengan menginisialisasi menghitung nilai pf, qf, lalu menghitung *score*-nya. Proses dari *Query Expansion Ranking* ditunjukkan dengan diagram alir pada Gambar 4.8, 4.9, dan 4.10.



Gambar 4.8 Diagram alir seleksi fitur *query expansion ranking*



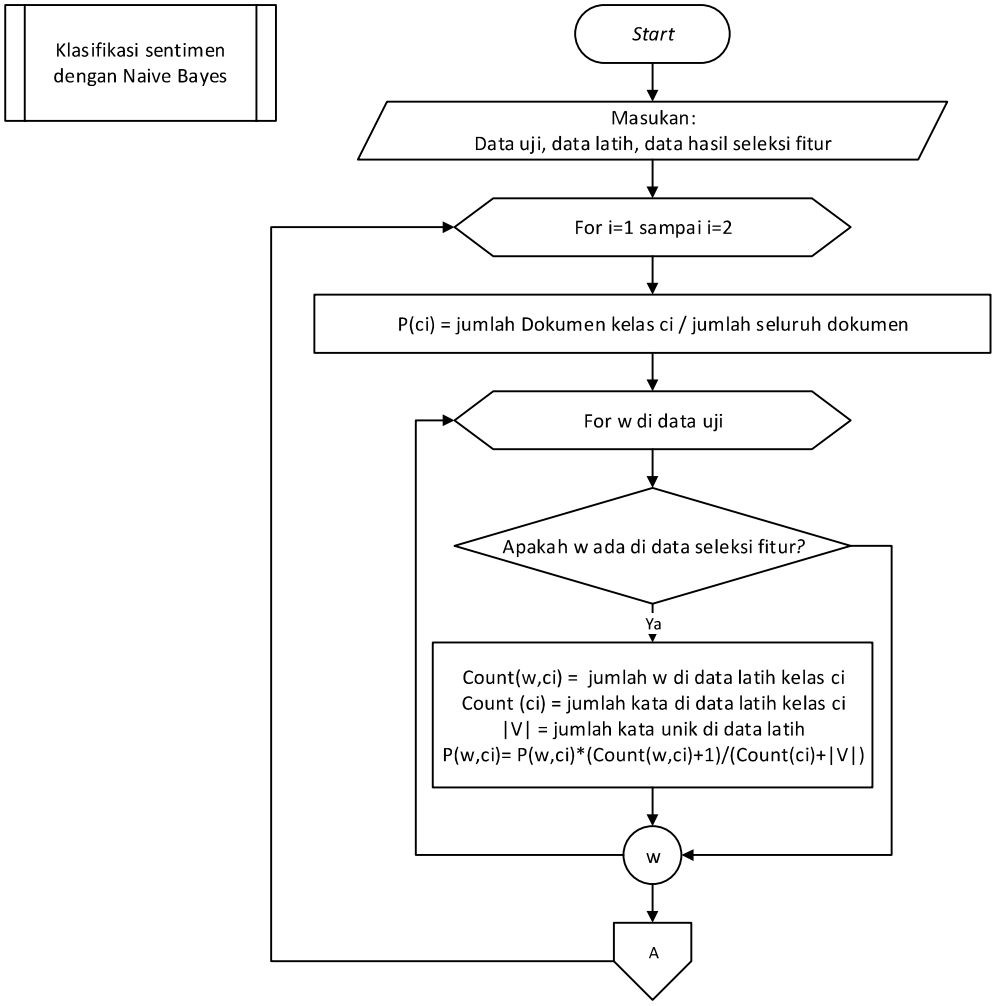
Gambar 4.9 Hitung nilai pf

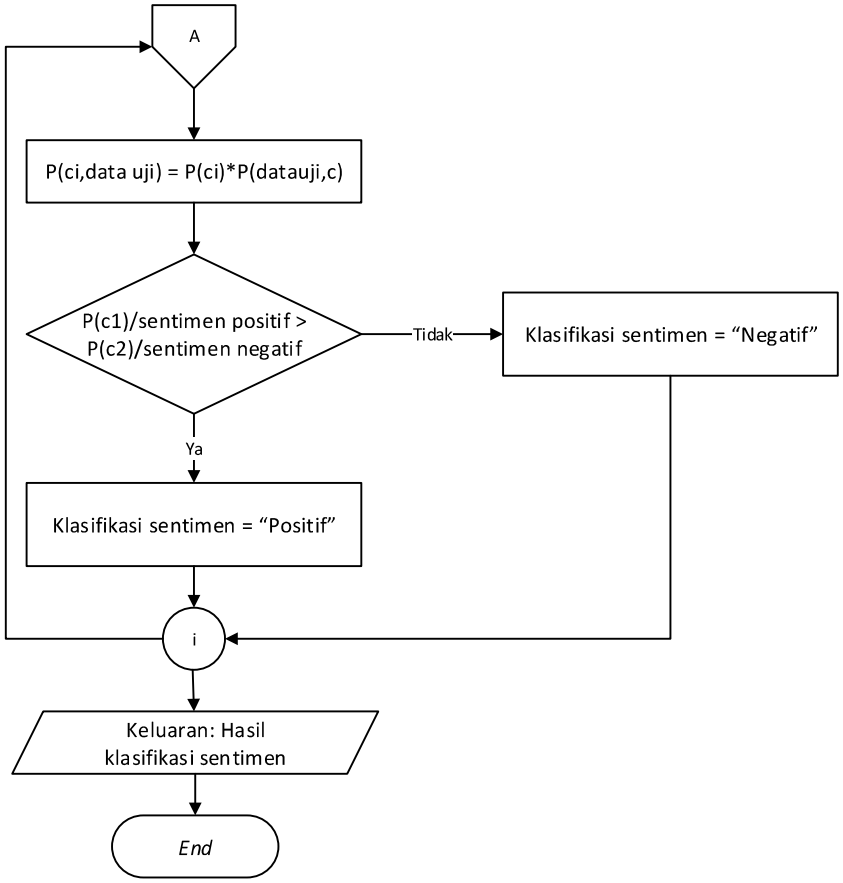


Gambar 4.10 Hitung nilai qf

4.4 Klasifikasi Sentimen menggunakan *Naive Bayes*

Tahap ini hanya dilakukan di fase uji dimana hanya data uji yang akan diklasifikasi. Data uji telah melalui tahap preprocessing dan hanya term dari seleksi fitur yang akan diklasifikasi. Untuk proses klasifikasi dengan *Naive Bayes* ini dapat dilihat pada Gambar 4.11.





Gambar 4.11 Diagram alir klasifikasi sentimen dengan *naive bayes*

4.5 Manualisasi

Manualisasi adalah bagian dimana perhitungan matematika pada sistem dicoba secara manual menggunakan aplikasi komputer *Microsoft Excel.*

4.5.1 Data

Untuk melakukan manualisasi maka dibutuhkan beberapa data berupa ulasan pariwisata di Kota Malang yang sudah ditentukan sentimennya, data tersebut akan digunakan sebagai data latih untuk melakukan analisis sentimen. Adapun data digambarkan melalui Tabel 4.1.

Tabel 4.1 Data latih manualisasi

Dokumen	Ulasan	Sentimen
1	Saya suka di tempat ini	Positif
2	tempatnya bagus bagus	Positif
3	pemandangan yang indah	Positif

Tabel 4.1 (lanjutan)

Dokumen	Ulasan	Sentimen
4	cocok untuk bersantai bersama keluarga	Positif
5	makanannya enak lezat	Positif
6	tempatnya kotor, banyak sampah	Negatif
7	nggak mau kesini lagi. Kapok	Negatif
8	nggak ada rasanya	Negatif
9	pemandangannya nggak sebagus di instagram	Negatif
10	bagus tapi banyak sampah	Negatif

4.5.2 Preprocessing

Tahap ini dibagi menjadi tiga yaitu tokenisasi, *filtering,* dan *stemming.*

4.5.2.1 Tokenisasi

Tokenisasi pada bagian ini menghapus semua karakter dan angka, serta pengubahan semua huruf ke huruf kecil. Hasil dari tokenisasi ditunjukkan oleh Tabel 4.2.

Tabel 4.2 Hasil manualisasi tokenisasi

Dokumen	Ulasan	Sentimen
1	'saya', 'suka', 'di', 'tempat', 'ini'	Positif
2	'tempatnya', 'bagus', 'bagus'	Positif
3	'pemandangan', 'yang', 'indah'	Positif
4	'cocok', 'untuk', 'bersantai', 'bersama', 'keluarga'	Positif
5	'makanannya', 'enak', 'lezat'	Positif
6	'tempatnya', 'kotor', 'banyak', 'sampah'	Negatif
7	'nggak', 'mau', 'kesini', 'lagi', 'kapok'	Negatif
8	'nggak', 'ada', 'rasanya'	Negatif
9	'pemandangannya', 'nggak', 'sebagus', 'di', 'instagram'	Negatif
10	'bagus', 'tapi', 'banyak', 'sampah'	Negatif

4.5.2.2 Filtering

Filtering adalah proses penghilangkan kata yang ada pada *stopword*, proses ini berkerja dengan mencocokkan kata-kata pada hasil tokenisasi, jika terdapat kata pada stopword maka kata tersebut dihapus. Hasil dari filtering ditampilkan pada Tabel 4.3.

Tabel 4.3 Hasil manualisasi *filtering*

Dokumen	Ulasan	Sentimen
1	'suka'	Positif
2	'tempatnya', 'bagus', 'bagus'	Positif
3	'pemandangan', 'indah'	Positif
4	'cocok', 'bersantai', 'keluarga'	Positif
5	'makanannya', 'enak', 'lezat'	Positif
6	'tempatnya', 'kotor', 'sampah'	Negatif
7	'nggak', 'kesini', 'kapok'	Negatif
8	'nggak'	Negatif
9	'pemandangannya', 'nggak', 'sebagus', 'instagram'	Negatif
10	'bagus', 'sampah'	Negatif

4.5.2.3 Stemming

Stemming adalah proses perubahan semua kata ke kata dasar yang biasa disebut fitur, sehingga kata yang berimbuhan akan menjadi kata yang sama, hasil manualisasi stemming ditunjukkan pada Tabel 4.4.

Tabel 4.4 Hasil manualisasi *stemming*

Dokumen	Ulasan	Sentimen
1	'suka'	Positif
2	'tempat', 'bagus', 'bagus'	Positif
3	'pandang', 'indah'	Positif
4	'cocok', 'santa', 'keluarga'	Positif
5	'makan', 'enak', 'lezat'	Positif
6	'tempat', 'kotor', 'sampah'	Negatif
7	'nggak', 'kesini', 'kapok'	Negatif
8	'nggak'	Negatif
9	'pandang', 'nggak', 'bagus', 'instagram'	Negatif
10	'bagus', 'sampah'	Negatif

4.5.3 Seleksi Fitur *Query Expansion Ranking*

Proses seleksi fitur *Query Expansion Ranking* ini adalah proses untuk mendapatkan *score* setiap fitur menggunakan rumus *Query Expansion Ranking* berdasarkan persamaan 2.6, 2.7, dan 2.8 kemudian dirutkan nilainya dari terkecil hingga terbesar. Proses penghitungan ditunjukkan oleh Tabel 4.5.

Tabel 4.5 Manualisasi *query expansion ranking*

FITUR	DOKUMEN										df_+^f	df_-^f	p_f	q_f	$Score_f$
	1	2	3	4	5	6	7	8	9	10					
nggak	0	0	0	0	0	0	1	1	1	0	0	3	0.083	0.636	1.301
sampah	0	0	0	0	0	1	0	0	0	1	0	2	0.083	0.455	1.449
kotor	0	0	0	0	0	1	0	0	0	0	0	1	0.083	0.273	1.880
kesini	0	0	0	0	0	0	1	0	0	0	0	1	0.083	0.273	1.880
kapok	0	0	0	0	0	0	1	0	0	0	0	1	0.083	0.273	1.880
instagram	0	0	0	0	0	0	0	0	1	0	0	1	0.083	0.273	1.880
suka	1	0	0	0	0	0	0	0	0	0	1	0	0.250	0.091	2.143
indah	0	0	1	0	0	0	0	0	0	0	1	0	0.250	0.091	2.143
cocok	0	0	0	1	0	0	0	0	0	0	1	0	0.250	0.091	2.143
santai	0	0	0	1	0	0	0	0	0	0	1	0	0.250	0.091	2.143
keluarga	0	0	0	1	0	0	0	0	0	0	1	0	0.250	0.091	2.143
makan	0	0	0	0	1	0	0	0	0	0	1	0	0.250	0.091	2.143
enak	0	0	0	0	1	0	0	0	0	0	1	0	0.250	0.091	2.143
lezat	0	0	0	0	1	0	0	0	0	0	1	0	0.250	0.091	2.143
bagus	0	2	0	0	0	0	0	0	1	1	1	2	0.250	0.455	3.444
tempat	0	1	0	0	0	1	0	0	0	0	1	1	0.250	0.273	23.000
pandang	0	0	1	0	0	0	0	0	1	0	1	1	0.250	0.273	23.000

Untuk mendapatkan nilai-nilai diatas maka diambil fitur “bagus” sebagai contoh.

- Nilai df_+^f didapatkan dengan menghitung jumlah kemunculan fitur “bagus” pada setiap dokumen kelas positif yaitu dokumen satu hingga lima. Kemunculan fitur “bagus” terdapat pada dokumen 2 sehingga nilai df_+^{bagus} adalah 1 walaupun pada dokumen tersebut fitur “bagus” muncul dua kali.
- Nilai df_-^f didapatkan dengan menghitung jumlah kemunculan fitur “bagus” pada setiap dokumen kelas negatif yaitu dokumen enam hingga sepuluh. Kemunculan fitur “bagus” terdapat pada dokumen sembilan dan sepuluh sehingga nilai df_-^{bagus} adalah 2.
- Nilai p_f didapatkan dengan menggunakan rumus 2.6.

$p_f = \frac{df_+^f + 0.5}{n^+ + 1.0}$; $p_{bagus} = \frac{1+0.5}{5+1.0}$; $p_{bagus} = \frac{1.5}{6}$; $p_{bagus} = 0.250$

- Nilai q_f didapatkan dengan menggunakan rumus 2.7.

$q_f = \frac{df_-^f + 0.5}{n^- + 0.5}$; $q_{bagus} = \frac{2+0.5}{5+0.5}$; $p_{bagus} = \frac{2.5}{5.5}$; $p_{bagus} = 0.455$

- Nilai $Score_f$ didapatkan dengan menggunakan rumus 2.5.

$Score_f = \frac{|p_f + q_f|}{|p_f - q_f|}$; $Score_{bagus} = \frac{|0.250+0.455|}{|0.250-0.455|}$;

$Score_{bagus} = \frac{0.705}{0.205}$; $Score_{bagus} = 0.344$

Setelah *score Query Expansion Ranking* setiap fitur didapatkan dan telah dirutkan dari terkecil hingga terbesar. Fitur-fitur tadi dipilih (dipotong) sesuai variasi rasio yang diinginkan. Pada contoh manualisasi ini rasio yang yang digunakan adalah 70% atau pemotongan 30%. Dari jumlah fitur asli sebanyak tujuh belas, seleksi fitur memilih sepuluh fitur terbaik yang akan digunakan untuk proses selanjutnya. Manualisasi seleksi fitur ditunjukkan oleh Tabel 4.6.

Tabel 4.6 Seleksi fitur

fitur	$Score_f$	fitur qer 70%
nggak	1.24	nggak
kotor	1.666667	instagram
sampah	1.666667	sampah
kesini	1.666667	kotor
kapok	1.666667	kesini
instagram	1.666667	enak
suka	2.6	lezat
indah	2.6	suka
cocok	2.6	keluarga
santai	2.6	makan
keluarga	2.6	
makan	2.6	
enak	2.6	
lezat	2.6	
tempat	7	
bagus	7	
pandang	7	

4.5.4 Klasifikasi Sentimen menggunakan *Naïve Bayes*

Pada tahap ini akan dilakukan klasifikasi dengan *Multinomial Naive Bayes* dengan mencari peluang setiap kata dari data uji. Hasil manualisasi untuk klasifikasi dokumen uji tersebut terbagi menjadi beberapa tahap sebagai berikut:

Tabel 4.7 Data uji

Dokumen	Ulasan	Sentimen
1	tempatnya nyaman dan enak, saya suka berkunjung kemari	?
2	Kotor, nggak mau	?

1. *Preprocessing*

- Tokenisasi

Tabel 4.8 Tokenisasi data uji

Dokumen	Ulasan	sentimen
1	'tempatnya', 'nyaman', 'dan', 'enak', 'saya', 'suka', 'berkunjung', 'kemari'	?
2	'kotor', 'nggak'	?

- *Filtering*

Tabel 4.9 *Filtering* data uji

Dokumen	Ulasan	sentimen
1	'tempatnya', 'nyaman', 'enak', 'suka', 'berkunjung', 'kemari'	?
2	'kotor', 'nggak'	?

- *Stemming*

Tabel 4.10 *Stemming* data uji

Dokumen	Ulasan	sentimen
1	'tempat', 'nyaman', 'enak', 'suka', 'kunjung', 'kemari'	?
2	'kotor', 'nggak'	?

2. Pencocokan term uji dengan term hasil seleksi fitur. Ini dilakukan untuk mengetahui apakah term tersebut ada atau tidak, jika ada term akan diproses dengan *Naive Bayes*, jika tidak maka term tidak akan diproses. Hasil pencocokan ditampilkan pada Tabel 4.13.

Tabel 4.11 Hasil pencocokan data uji dan seleksi fitur

pencocokan terhadap seleksi fitur			
dokumen	fitur	ada/tidak	proses/tidak
1	tempat	tidak ada	tidak
	nyaman	tidak ada	tidak
	enak	ada	proses
	saya	tidak ada	tidak
	suka	ada	proses
	kunjung	tidak ada	tidak
	kemari	tidak ada	tidak
2	kotor	ada	proses
	nggak	ada	proses

3. Melakukan perhitungan *Naive Bayes* setiap kelas dari fitur terpilih.

- Mencari kelas peluang positif dan negatif data uji berdasarkan persamaan 2.4.

  $P(c) = \frac{N_c}{N}$; $P(positif) = \frac{5}{10}$; $P(positif) = 0.5$

  Peluang positif adalah 0.5 yang didapat dari membagi jumlah data positif yang jumlahnya lima dengan jumlah seluruh data dari data latih yang berjumlah sepuluh.

  $P(c) = \frac{N_c}{N}$; $P(negatif) = \frac{5}{10}$; $P(negatif) = 0.5$

  Peluang negatif adalah 0.5 yang didapat dari membagi jumlah data negatif yang jumlahnya lima dengan jumlah seluruh data dari data latih yang berjumlah sepuluh.

- Mencari peluang setiap fitur berdasarkan persamaan 2.5.

  Untuk mencari peluang setiap fitur dilakukan perhitungan sesuai rumus 2.5. Sebagai contoh untuk mencari peluang "enak" pada kelas positif maka kita harus mengetahui jumlah fitur "enak" di data latih kelas positif, jumlah seluruh fitur pada kelas positif, dan jumlah fitur unik pada seluruh data latih. Berikut perhitungannya:

  $P(w|c) = \frac{count(w,c)+1}{count\ (c)+|V|}$; $P(enak|positif) = \frac{1+1}{12+17}$;

  $P(enak|positif) = \frac{2}{29}$; $P(enak|positif) = 0.068$

  Hasil dari pencarian peluang setiap fitur pada kelas negatif dan positif ditampilkan pada Tabel 4.12.

Tabel 4.12 Hasil Peluang setiap Fitur

Dokumen	Peluang Fitur	Peluang
1	$P(enak\|positif)$	0.06896
	$P(suka\|positif)$	0.03333
	$P(enak\|negatif)$	0.06896
	$P(suka\|negatif)$	0.03333
2	$P(kotor\|positif)$	0.03448
	$P(nggak\|positif)$	0.03448
	$P(kotor\|negatif)$	0.06667
	$P(nggak\|negatif)$	0.13333

- Menghitung nilai *Naive Bayes* setiap dokumen dihitung berdasarkan persamaan 2.3, sebagai contoh berikut perhitungan *Naive Bayes* pada dokumen 1 kelas positif.

  $P(c|d) = P(c) \prod_{i=1}^{n} P(w_i|c)$;

  $P(positif|enak, suka) = \times\ P(positif)\ \prod_{i=0}^{n} P(enak, suka|positif)$;

$$P(positif|enak, suka) = 0.5 \times 0.06896 \times 0.03333$$

$$P(positif|enak, suka) = 0.002378121$$

Nilai *Naive Bayes* kelas positif dan negatif dari setiap dokumen ditunjukkan dengan Tabel 4.13.

Tabel 4.13 Nilai *Naive Bayes*

Dokumen	Peluang	Nilai Naive Bayes
dok 1	P(positif|tempatnya nyaman dan enak, saya suka berkunjung kemari)	0.002642357
	P(negatif|tempatnya nyaman dan enak, saya suka berkunjung kemari)	0.000566893
dok 2	P(positif|kotor nggak mau)	0.000660589
	P(negatif|kotor nggak mau)	0.004535147

Dari tabel diatas kita bandingkan nilai *Naive Bayes* kelas manakah yang lebih besar. Pada dokumen 1 hasil *Naive Bayes* kelas positif lebih besar dari pada kelas negatif maka dokumen 1 masuk ke dalam kelas positif. Sedangkan pada dokumen 2, hasil *Naive Bayes* kelas negatif lebih tinggi dari kelas positif, maka dokumen 2 masuk ke dalam kelas negatif.

4.6 Perancangan Pengujian

Pengujian sistem dilakukan sesuai dengan skenario pengujian untuk mengetahui kinerja sistem. Pada penelitian ini pengujian dilakukan dengan mengamati pengaruh variasi rasio seleksi fitur *Query Expansion Ranking* pada tingkat akurasi sistem sehingga bisa disimpulkan berapa rasio terbaik untuk menghasilkan nilai akurasi yang tinggi. Variasi rasio yang akan diujikan pada penelitian ini ditunjukkan oleh Tabel 4.14.

Tabel 4.14 Perancangan pengujian

No	Fitur	Jumlah Fitur	Akurasi
1	25%		
2	50%		
3	75%		
4	100%		

Akurasi dihitung menggunakan persamaan Akurasi 2.9. Sebagai contoh, terdapat sepuluh data uji dengan jumlah data positif lima dan negatif lima namun pada sistem data uji tersebut terklasifikasi 7 positif dan 3 negatif maka akurasi dapat dihitung dengan cara sebagai berikut:

TP = 5 ; TN= 3 ; FP = 0 ; FN = 2

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$$

$$Akurasi = \frac{5+3}{5+3+0+2} \times 100\%$$

$$Akurasi = \frac{8}{10} \times 100\%$$

$$Akurasi = 80\%$$