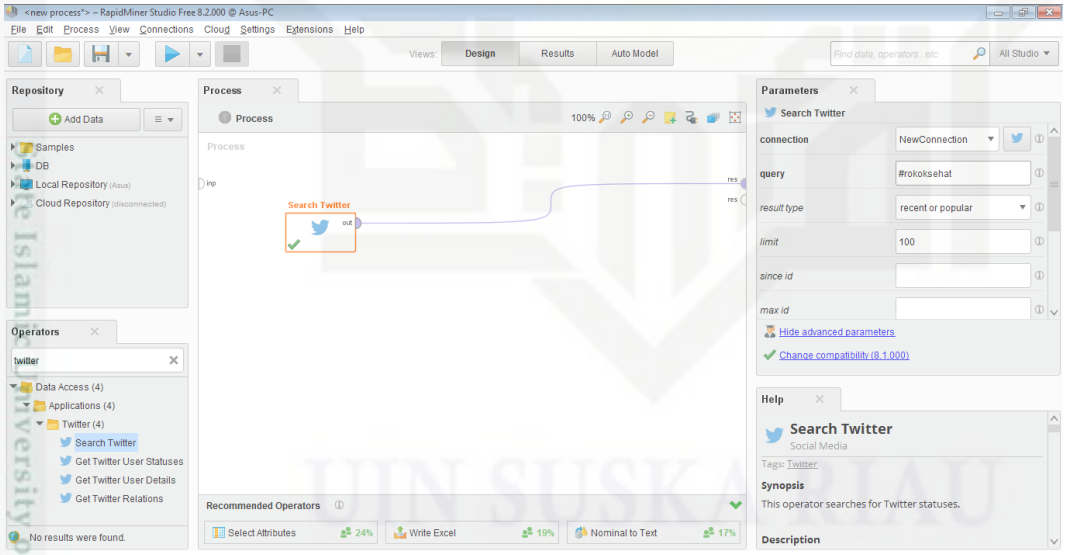
BAB IV
ANALISA DAN PERANCANGAN

4.1 Analisa Data

Pada penelitian ini data yang digunakan merupakan data komentar masyarakat di Twitter yang mengandung kalimat sentimen terhadap rokok yang sudah ditentukan. Jumlah data komentar di Twitter yang akan diunduh adalah 1000 *Tweet*. Data yang telah terkumpul diberi label secara manual dengan label positif dan label negatif. Data yang telah diunduh akan dibagi menjadi 500 label positif dan 500 label negatif dan telah melakukan pemrosesan text.

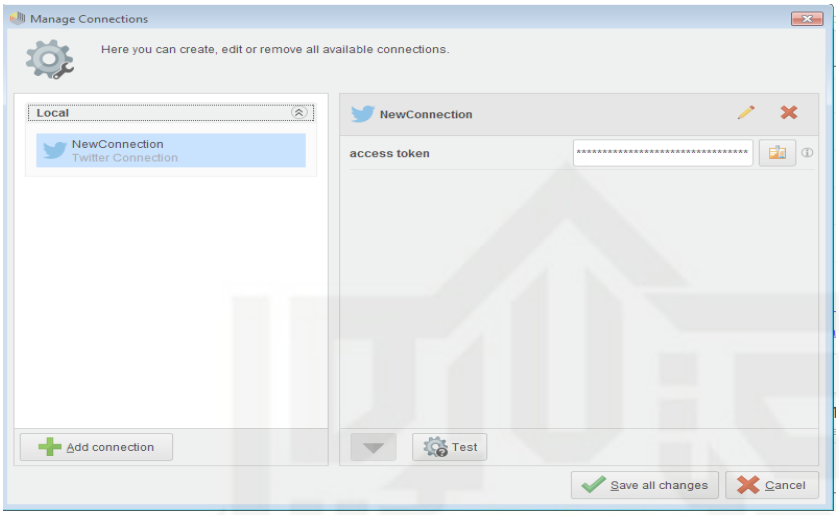
Data komentar masyarakat diperoleh dari *Twitter* dengan menggunakan Twitter API pada *RapidMiner*. Berikut adalah tahapan pengambilan data Twitter API yang digunakan untuk mengambil data komentar pada *RapidMiner* :



Gambar 4.1 Tampilan Awal Rapidminer

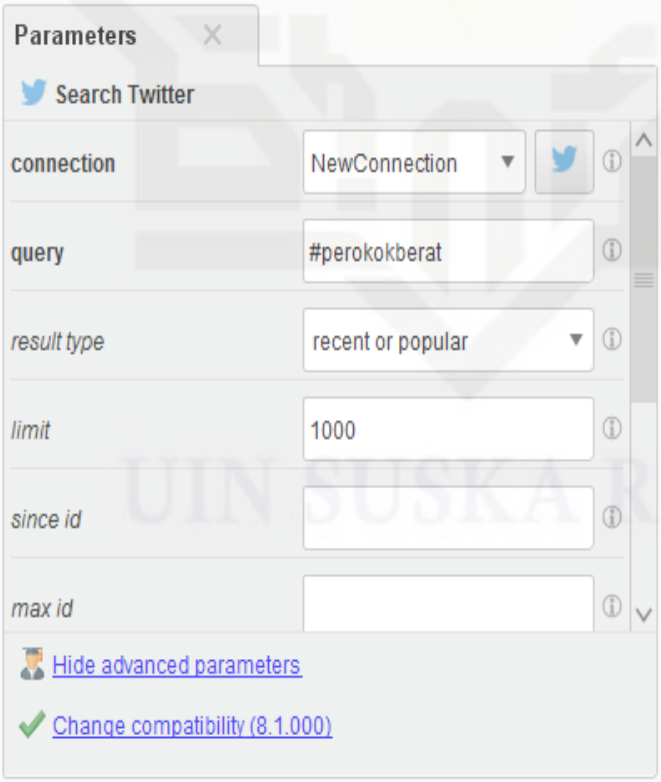
Tampilan diatas ialah awal dari proses pengambilan data *tweet* di *RapidMiner*. Dalam proses tersebut digunakan operator data acces untuk membuat proses pencarian pada *Twitter*.

Selanjutnya pada parameter pencarian *twitter*, masukan akses token Twitter API yang telah didaftar pada *twitter*.



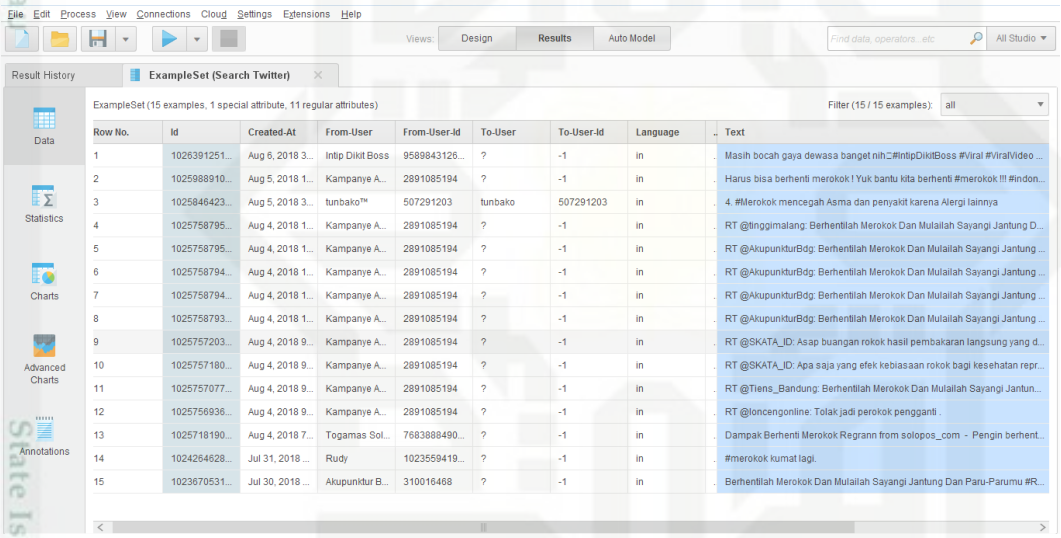
Gambar 4.2 Masukan Token Twitter API

Kemudian masukan kata kunci pencarian data *tweet* yang akan digunakan pada menu query di parameter.



Gambar 4.3 Contoh Query Pencarian Data

Data komentar pada Twitter API menggunakan beberapa kata kunci (keyword) yang berkaitan dengan sentimen masyarakat terhadap rokok. Kata kunci yang digunakan antara lain: #rokok, #rokoksehat, #rokokmahal, #antirokok, #saverokok, #hargarokoknaik, #bebasrokok, #puasarokok, #rokoktidaksehat, #peraturanrokok, #cukairokok, #hargarokok, #kenaikanhargarokok, #sebatangrokok, #perokok, #stopmerokok, #rokokmurah, #antiasaprokok, #berhetimerokok, #melawanrokok, #kawasantanparokok, #kawasanmerokok, #mantanperokok, #iklanrokok, #guabukanperokok, #perokokpasif, #perokokaktif, #bebastembakau, #perokokberat.



Gambar 4.4 Contoh Data Twitter API

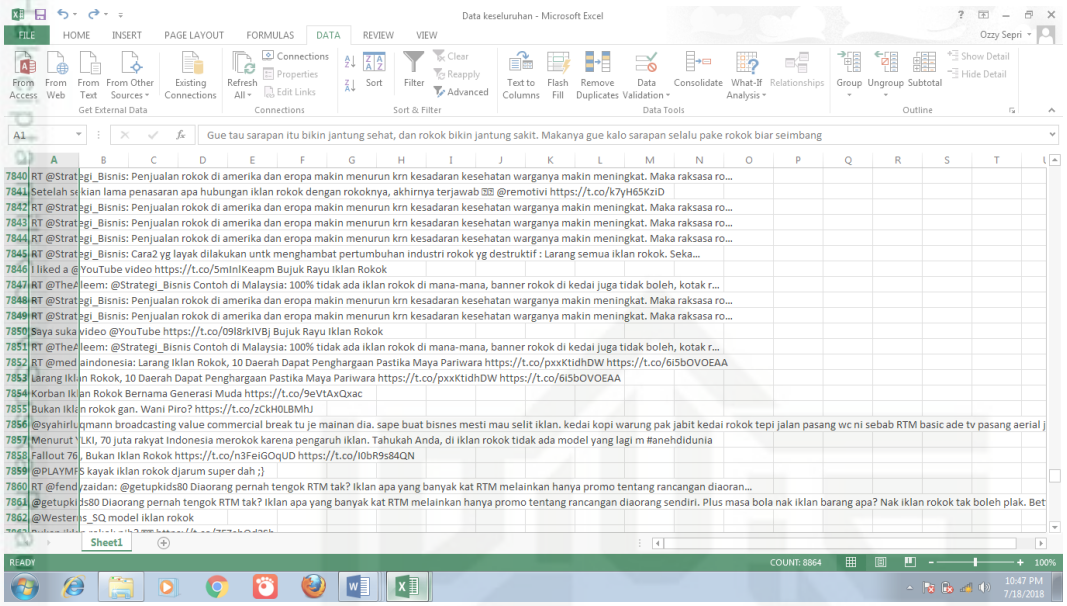
Berikut ini adalah table data rill dari data komentar yang telah di salin menggunakan Twitter API berdasarkan kata kunci (keyword) :

Tabel 4.1 Data Komentar Setiap *Keyword*

No	Hastag	Jumlah data
1	#rokok	76
2	#rokoksehat	253
3	#rokokmahal	310
4	#antirokok	500
5	#saverokok	9
6	#hargarokoknaik	89

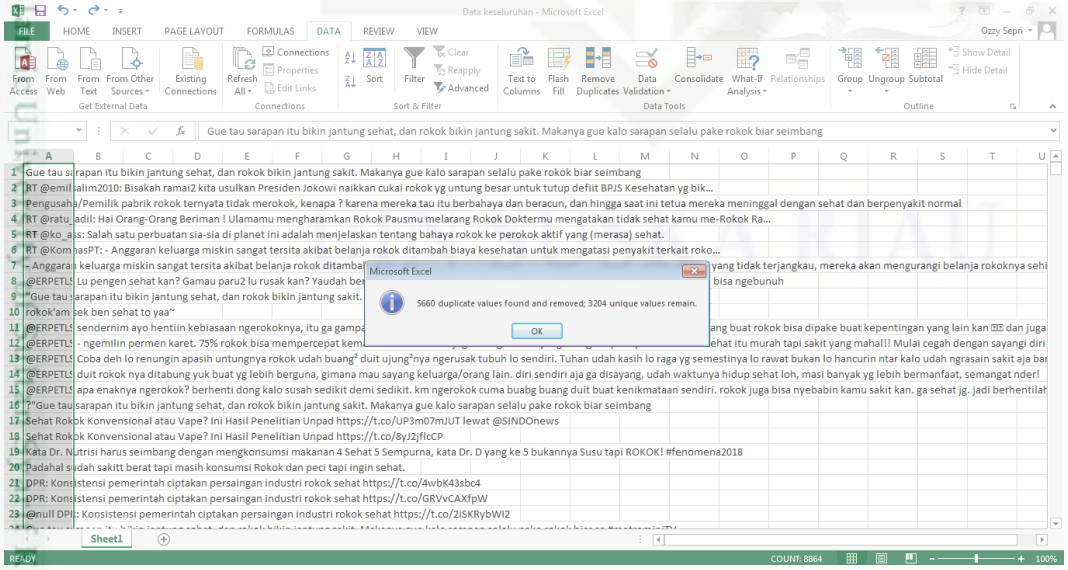
7	#bebasrokok	500
8	#puasarokok	500
9	#rokoktidaksehat	54
10	#peraturanrokok	27
11	#cukairokok	362
12	#hargarokok	1000
13	#kenaikanhargarokok	11
14	#sebatangrokok	14
15	#perokok	1000
16	#stopmerokok	621
17	#rokokmurah	273
18	#antiasaprokok	29
19	#berhentimerokok	1000
20	#melawanrokok	18
21	#kawasantanparokok	182
22	#kawasanmerokok	117
23	#mantanperokok	16
24	#iklanrokok	394
25	#guabukanperokok	14
26	#rokokpasif	193
27	#rokokaktif	343
28	#bebastembakau	47
29	#perokokberat	122

Jumlah data keseluruhan yang didapat pada API Twitter sebanyak 8.864 data *tweet*. Data komentar *tweet* yang telah didapat di export lalu disimpan dalam bentuk format .CSV. Berikut ini adalah contoh data *tweet* yang telah disimpan dalam bentuk format .CSV :



Gambar 4.5 Data Komentar dalam format CSV

Kemudian data *tweet* tersebut dilakukan pengecekan apakah terdapat data *tweet* yang sama untuk menghindari redundancy pada data *tweet*. Untuk melakukan pengecekan data yang redundancy dilakukan di micrososft excel. Dengan cara blok seluruh data *tweet* yang akan dilakukan redundancy. Kemudian pilih pada remove duplicates pada Microsoft excel. Dari proses redudancy tadi telah dibuang data yang sama sebanyak 5.660 sehingga didapat hasil *tweet* menjadi 3.204 dari *tweet* awal 8.864. Berikut adalah gambar redudancy data *tweet*

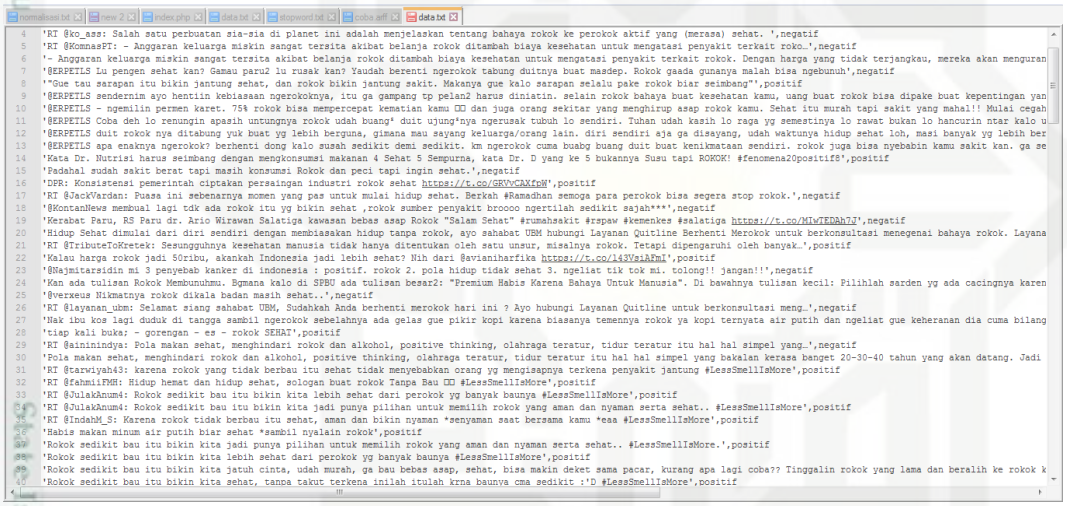


Gambar 4.6 hasil pembuangan data yang sama

Data *tweet* yang telah di redundancy dilakukan perlabelan, yaitu positif dan negatif secara manual. Dari 3.204 data *tweet* hasil redundancy maka diambil 1000 data *tweet* untuk penelitian ini. 500 data positif dan 500 data negatif. Kemudian data komentar disimpan kedalam format .txt. Berikut adalah contoh data komentar yang dicopy kedalam format .txt. setelah data melalui proses seleksi. Setelah diubah data diubah menjadi format text (*.txt) kemudian dilakukan proses pelabelan dengan format perbaris sebagai berikut.

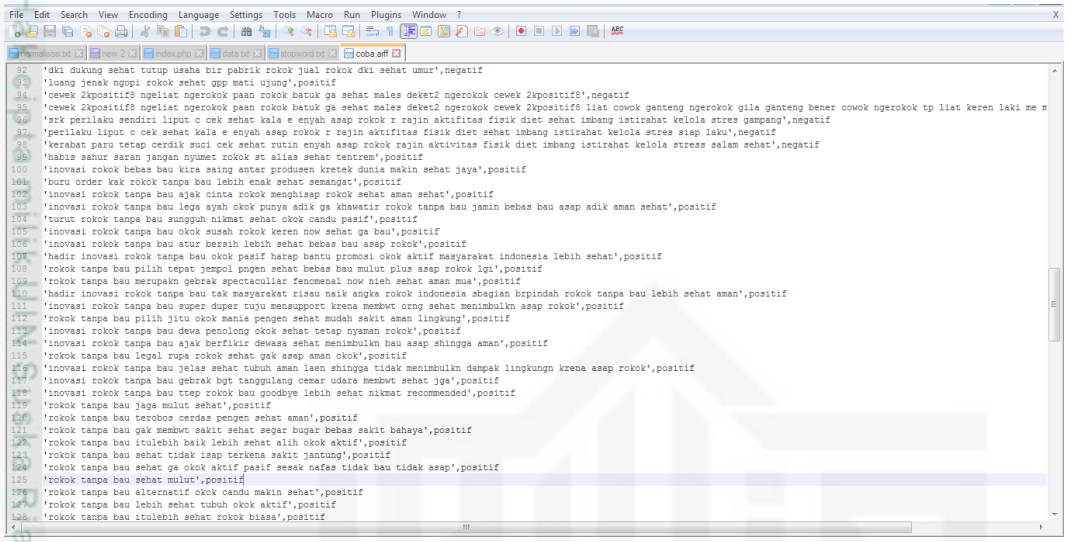
'isi *tweet*', kelas sentimen (positif atau negatif)

Untuk lebih jelasnya dapat dilihat pada gambar 4.7 dibawah ini:



Gambar 4.7 Format data tweet *.txt

Data yang telah melalui proses seleksi, pelabelan dan diubah menjadi text kemudian melalui pemrosesan text menggunakan coding PHP. Coding PHP lebih lengkap dilampirkan pada Lampiran A. Setelah melalui pemrosesan teks data yang didapat diubah kembali menjadi format *Attribute relation file format* (Arff) untuk menyesuaikan format pada Weka seperti terlihat pada gambar dibawah ini Gambar 4.8 :

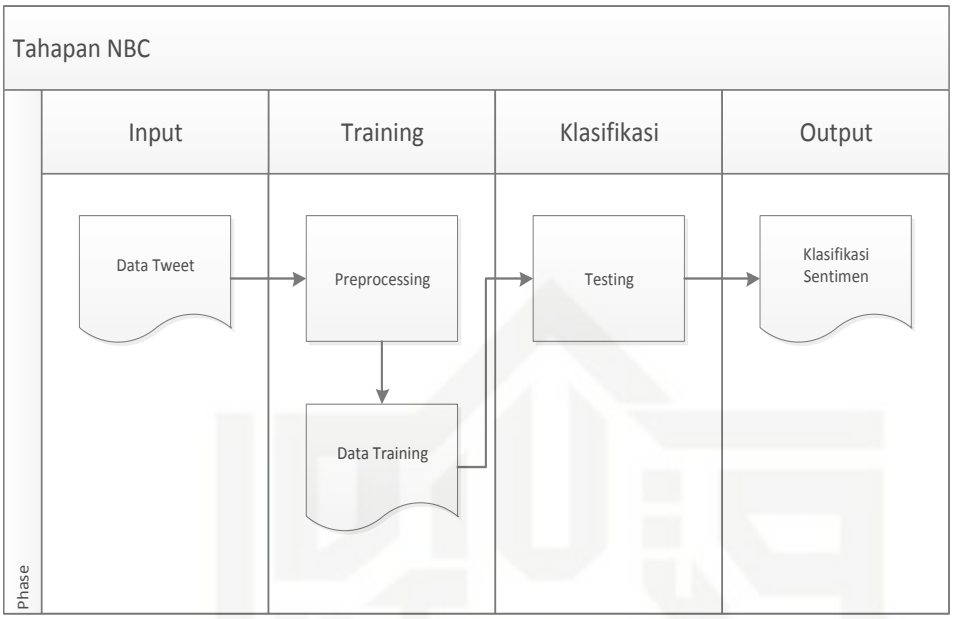


Gambar 4.8 Format data tweet *.arff

Pada gambar terlihat bahwa ada @relation, @attribute dan @data.@relation menunjukan nama relasi pada data, @attribute merupakan keterangan fitur atau atribut dari data *tweet* berupa nama dan tipe data dari fitur dan @data berisikan *tweet* hasil preprocessing serta kelas dari *tweet* tersebut, dimana satu baris menyatakan satu data.

4.2 Analisa Metode Naïve Bayes Classifier

Analisa algoritma NBC adalah tahapan menentukan proses klasifikasi sentimen. Penelitian ini menggunakan tools Weka 3-6-10 yang merupakan salah satu tools data mining yang bisa digunakan untuk klasifikasi *tweet*. Adapun function yang digunakan adalah Bayes multinominal bisa digunakan untuk klasifikasi data dengan NBC. Secara umum tahapan kalsifikasi sentimen *tweet* dapat dilihat pada gambar 4.9 berikut ini.



Gambar 4.9 Tahapan Klasifikasi NBC Secara Umum

Berikut adalah Penjelasan Gambar 4.9 :

1. Input

   Bagian input dari penelitian ini adalah seluruh data *tweet* yang berjumlah 1000 dan sudah dilabel manual serta dibagi menjadi 90% dari 1000 data sebagai data latih dan 10% dari 1000 data sebagai data uji. Adapun data *tweet* tersebut harus sesuai dengan format masukan Weka.

2. Training

   Tahapan selanjutnya adalah preprocessing, stemming dan pembobotan fitur. Untuk pembobotan fitur dapat menggunakan Weka. Dalam penelitian inii untuk NBC pembobotan fiturnya menggunakan TF.

3. Klasifikasi

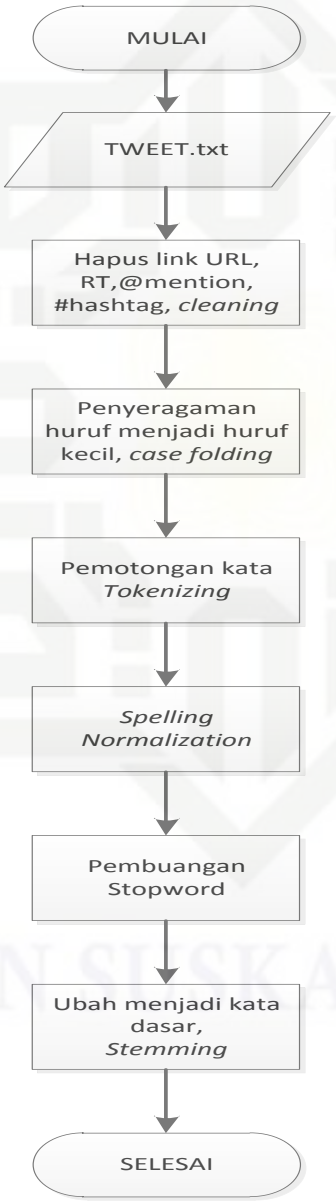
   Tahap ini akan digunakan untuk proses klasifikasidan pengujian. Pada NBC tidak diperlukan pencarian model terbaik dan Pada NBC pula data latih langsung bisa diklasifikasikan.

4. Output

   Hasil keluaran (output) dari klasifikasi adalah nilai pengujian dan akurasi pengklasifikasian teks.

4.3 *Text Prepocessing*

Preprocessing merupakan langkah penting dalam melakukan analisa sentimen yang bertujuan untuk membersihkan data dari unsur-unsur yang ada tetapi tidak di butuhkan.Gambar 4.10 berikut ini merupakan langkah secara umum *preprocessing text*.



Gambar 4.10 Flowchart Text Prepocessing

Pada tabel 4.2 dibawah ini adalah contoh data komentar yang sudah berhasil dikumpulkan dan sudah diberi label sentimen.

Tabel 4.2 Contoh Data Komentar

Tweet (n)	Komentar	Sentimen
Tweet (1)	Demi Kesehatanmu, Berhentilah Merokok! #AntiRokok	Negatif
Tweet (2)	Asap Rokok Membuat Nafas ini Sesak	Negatif
Tweet (3)	katakan tidak pada rokok #**antirokok** #antiperokok	Negatif
Tweet (4)	Batuknya seorang perokok adalah tanda rusaknya bagian dalam tubuh! #**AntiRokok** #HidupSehat #RokokMembunuhmu #AsapnyaMembunuhOrangSekitar	Negatif
Tweet (5)	sayangi diri sendiri & orang yang kita cintai dengan tidak merokok. --- #selamatkananakdariasaprokok #**antirokok** http://fb.me/wDQwXIsp	Negatif
Tweet (6)	Dalam nikotinku terdapat kenyamaan selain darimu. Uwaaaww #saverokok	Positif
Tweet (7)	#SaveRokok kasian nanti petani pabrik tembakau bangkrut	Positif
Tweet (8)	Rokok itu punya ciri khas tersendiri #asbak #aburokok #**saverokok**	Positif

Tweet (9)	Turunkan harga rokok... #saveRokok	Negatif

Berikut ini merupakan penjelasan dari gambar 4.10 tahapan dari *text preprocessing* :

1. *Cleaning*

Adapun kata atau karakter yang akan dihilangkan adalah karakter atau simbol, link url (http://situs.com), *hashtag* (#), *username* atau *mention* (@username), emoticon dan RT (tanda *retweet*) serta *emoticon*. Hasil *cleaning* dari contoh *tweet* pada tabel 4.3 adalah sebagai berikut:

Tabel 4.3 Hasil *Cleaning*

***Tweet* (n)**	**Komentar**	**Sentimen**
1	Demi Kesehatanmu, Berhentilah Merokok	Negatif
2	Asap Rokok Membuat Nafas ini Sesak	Negatif
3	katakan tidak pada rokok	Negatif
4	Batuknya seorang perokok adalah tanda rusaknya bagian dalam tubuh	Negatif
5	sayangi diri sendiri & orang yang kita cintai dengan tidak merokok	Negatif
6	Dalam nikotinku terdapat kenyamaan selain darimu	Positif
7	kasian nanti petani pabrik tembakau bangkrut	Positif
8	Rokok itu punya ciri khas tersendiri	Positif
9	Turunkan harga rokok	Negatif

2. *Case Folding*

Proses *Case folding* adalah proses penyeragaman bentuk huruf dengan mengubah semua huruf menjadi huruf kecil, dan juga menghilangkan tanda baca dan angka, dalam hal ini hanya menggunakan huruf antara a sampai z. Tabel 4.4 berikut adalah hasil *tweet* yang telah dilakukan proses *case folding*.

Tabel 4.4 Hasil *Case Folding*

Tweet (n)	Komentar	Sentimen
1	demi kesehatanmu berhentilah merokok	Negatif
2	asap rokok membuat nafas ini sesak	Negatif
3	katakan tidak pada rokok	Negatif
4	batuknya seorang perokok adalah tanda rusaknya bagian dalam tubuh	Negatif
5	sayangi diri sendiri orang yang kita cintai dengan tidak merokok	Negatif
6	dalam nikotinku terdapat kenyamaan selain darimu	Positif
7	kasian nanti petani pabrik tembakau bangkrut	Positif
8	rokok itu punya ciri khas tersendiri	Positif
9	turunkan harga rokok	Negatif

3. *Tokenizing*

Proses *Tokenizing* yaitu proses memecah *tweet* atau kalimat menjadi sebuah kata dengan melakukan analisa terhadap kumpulan kata dengan memisahkan kata tersebut dan menentukan struktur sintaksis dari tiap kata tersebut. Pada

penelitian ini fitur yang digunakan adalah fitur bigram. Berikut ini adalah hasil *tokenizing* untuk contoh *tweet* yang dapat dilihat pada table 4.5 :

Tabel 4.5 Hasil Tokenizing

***Tweet* 1**	***Tweet* 2**	***Tweet* 3**	***Tweet* 4**	***Tweet* 5**	***Tweet* 6**	***Tweet* 7**	***Tweet* 8**	***tweet* 9**
demi	asap	katakan	batuknya	sayangi	dalam	kasian	rokok	turunkan
kesehatanmu	rokok	tidak	seorang	diri	nikotinku	nanti	itu	harga
berhentilah	membuat	pada	perokok	sendiri	terdapat	petani	punya	rokok
merokok	nafas	rokok	adalah	orang	kenyamanan	pabrik	ciri	
	ini		tanda	yang	selain	tembakau	khas	
	sesak		rusaknya	kita	dirimu	bangkrut	tersendiri	
			bagian	cintai				
			dalam	dengan				
			tubuh	tidak				
				merokok				

4. Spelling Normalization

Merupakan proses perbaikan kata yang tidak sesuai dengan penulisan kata yang sebenarnya misalnya “tdk” diubah menjadi “tidak”. Dapat dilihat pada table 4.6.

Tabel 4.6 Hasil Spelling Normalization

sTweet 1	*Tweet 2*	*Tweet 3*	*Tweet 4*	*Tweet 5*	*Tweet 6*	*Tweet 7*	*Tweet 8*	*Tweet 9*
demi	asap	Katakana	batuknya	sayangi	dalam	kasihan	rokok	turunkan
kesehatanmu	rokok	tidak	seorang	diri	nikotinku	nanti	itu	harga
berhentilah	membuat	pada	perokok	sendiri	terdapat	petani	punya	rokok
merokok	nafas	rokok	adalah	orang	kenyamanan	pabrik	ciri	
	ini		tanda	yang	selain	tembakau	khas	
	sesak		rusaknya	kita	dirimu	bangkrut	tersendiri	
			bagian	cintai				
			dalam	dengan				
			tubuh	tidak				
				merokok				

5. *Filtering*

Filtering adalah tahap mengambil kata-kata penting dari hasil token. Biasanya tahap ini menggunakan algoritma *stop-list*(membuang kata-kata kurang penting) atau *word-list (*menyimpan kata penting*).*

Tabel 4.7 Hasil *Filtering*

Tweet 1	*Tweet* 2	*Tweet* 3	*Tweet* 4	*Tweet* 5	*Tweet* 6	*Tweet* 7	*Tweet* 8	*Tweet* 9
demi	asap	Tidak	batuknya	sayangi	nikotinku	kasihan	rokok	turunkan
kesehatanmu	rokok	Rokok	perokok	diri	kenyamanan	petani	khas	harga
berhentilah	membuat		rusaknya	cintai	dirimu	pabrik	tersendiri	rokok
merokok	nafas		tubuh	tidak		tembakau		
	sesak			merokok		bangkrut		

6. *Stemming*

Proses *stemming* merupakan proses mengubah kata-kata yang mengandung imbuhan (affik), awalan (prefix), akhiran (suffix), sisipan (infix), dan awalan akhiran (konfix) menjadi kata dasar dengan menggunakan algoritma *stemming* Nazief dan Adriani. Berdasarkan contoh *tweet* diatas, maka hasil *stemming*-nya dapat dilihat pada tabel 4.8 berikut ini.

Tabel 4.8 Hasil *Stemming*

Tweet 1	*Tweet* 2	*Tweet* 3	*Tweet* 4	*Tweet* 5	*Tweet* 6	*Tweet* 7	*Tweet* 8	*Tweet* 9
demi	asap	tidak	batuk	sayang	nikotin	kasih	rokok	turun
kesehatan	rokok	rokok	rokok	diri	nyaman	petani	khas	harga
henti	buat		rusak	cinta	diri	pabrik	diri	rokok
rokok	nafas		tubuh	tidak		tembakau		

Tweet 1	*Tweet* 2	*Tweet* 3	*Tweet* 4	*Tweet* 5	*Tweet* 6	*Tweet* 7	*Tweet* 8	*Tweet* 9
	sesak			rokok		bangkrut		

4.4 *Text transformation*

Pada penelitian ini digunakan pembobotan dengan menggunakan TF (*term frequency*). Pembobotan adalah proses merubah kata menjadi bentuk vektor. TF adalah jumlah kemunculan kata dalam dokumen.

Dalam penelitian tugas akhir ini fitur yang digunakan adalah *unigram* dengan pembobotan menggunakan TF. Kata direpresentasi ke dalam bentuk vektor, dimana tiap kata dihitung sebagai satu fitur. Adapun perhitungan bobot yang digunakan adalah *Term Frequency* (TF). Pada tabel 4.9 adalah hasil pembobotan selengkapnya berdasarkan contoh komentar di atas.

Tabel 4.9 Text Transformation

Kosa Kata	*tf(D1)*	*tf(D2)*	*tf(D3)*	*tf(D4)*	*tf(D5)*	*tf(D6)*	*tf(D7)*	*tf(D8)*	*tf(D9)*	*tf*
demi	1	0	0	0	0	0	0	0	0	1
kesehatan	1	0	0	0	0	0	0	0	0	1
henti	1	0	0	0	0	0	0	0	0	1
rokok	1	1	1	1	1	0	0	1	1	7
asap	0	1	0	0	0	0	0	0	0	1
buat	0	1	0	0	0	0	0	0	0	1
nafas	0	1	0	0	0	0	0	0	0	1
sesak	0	1	0	0	0	0	0	0	0	1
tidak	0	0	1	0	1	0	0	0	0	2
batuk	0	0	0	1	0	0	0	0	0	1

Kosa Kata	*tf(D1)*	*tf(D2)*	*tf(D3)*	*tf(D4)*	*tf(D5)*	*tf(D6)*	*tf(D7)*	*tf(D8)*	*tf(D9)*	*tf*
rusak	0	0	0	1	0	0	0	0	0	1
tubuh	0	0	0	1	0	0	0	0	0	1
sayang	0	0	0	0	1	0	0	0	0	1
diri	0	0	0	0	1	1	0	1	0	3
cinta	0	0	0	0	1	0	0	0	0	1
nikotin	0	0	0	0	0	1	0	0	0	1
nyaman	0	0	0	0	0	1	0	0	0	1
kasih	0	0	0	0	0	0	1	0	0	1
petani	0	0	0	0	0	0	1	0	0	1
pabrik	0	0	0	0	0	0	1	0	0	1
tembakau	0	0	0	0	0	0	1	0	0	1
bangkrut	0	0	0	0	0	0	1	0	0	1
khas	0	0	0	0	0	0	0	1	0	1
turun	0	0	0	0	0	0	0	0	1	1
harga	0	0	0	0	0	0	0	0	1	1

Kosa kata : Kata tunggal yang telah terpisah dari dokumen query

tf(n) : *Term frequency* (frekuensi kata) pada dokumen ke-*n*

4.5 Proses klasifikasi Naïve Bayes Classifier

1 Tahapan latih/training untuk mencari prior menggunakan persamaan 2.2.

$$\mathbf{P(p/n)} = \frac{\mathbf{N(pos/neg)}}{\mathbf{N}}$$

$$p(positif) = \frac{f_d(positif)}{|D|} = \frac{3}{9} = 0.3333$$

$$p(negatif) = \frac{f_d(negatif)}{|D|} = \frac{6}{9} = 0.6667$$

2 Tahap Uji (*testing*)

Data uji

Komentar	Kelas
Dulu belajar merokok, Kali ini berusaha belajar berhenti merokok!!! #**sehattanparokok** #stopmerokok #bahayamerokok.	?

Setelah di *text mining*

Komentar	Kelas
belajar rokok usaha henti	?

Kemudian hitung probabilitas setiap *term* dari data uji dengan menggunakan persamaan 2.3. Sebelumnya hitung jumlah seluruh *term* yang terdapat pada data latih yang telah di *stemming*. Jumlah *term* pada data latih sebanyak 34, 11 *term* dari kategori positif, dan 23 *term* dari kategori negatif. Banyaknya *term* tergantung pada proses *preprocessing* dan *stemming.* Perhitungan probabilitas dari setiap *term* data uji menggunakan pesamaan 2.3 sebagai berikut :

$$\mathbf{P\ (w|pos/neg)} = \frac{\mathbf{count(w,pos/neg)+1}}{\mathbf{count(pos/neg)+\ |v|}}$$

Diketahui IVI =

Count Positif = 11, *count* negatif = 23

- Probabilitas data uji kelas positif

$$p(\text{"belajar"}|\text{"positif"}) = \frac{(f(\text{"}belajar\text{"}|\text{"positif"}) + 1)}{f(\text{"positif"}) + |V|} = \frac{0+1}{11+25} = 0.0278$$

$$p(\text{"}rokok\text{"}|\text{"positif"}) = \frac{(f(\text{"}rokok\text{"}|\text{"positif"}) + 1)}{f(\text{"positif"}) + |V|} = \frac{1+1}{11+25} = 0.0556$$

$$p(\text{"usaha"}|\text{"positif"}) = \frac{(f(\text{"}usaha\text{"}|\text{"positif"}) + 1)}{f(\text{"positif"}) + |V|} = \frac{0+1}{11+25} = 0.0278$$

$$p(\text{"henti"}|\text{"positif"}) = \frac{(f(\text{"}henti\text{"}|\text{"positif"}) + 1)}{f(\text{"positif"}) + |V|} = \frac{0+1}{11+25} = 0.0278$$

- Probabilitas data uji kelas negatif

$$p(\text{"belajar"}|\text{negatif"}) = \frac{(f(\text{"}belajar\text{"}|\text{"negatif"}) + 1)}{f(\text{"negatif"}) + |V|} = \frac{0+1}{23+25} = 0.0208$$

$$p(\text{"rokok"}|\text{negatif"}) = \frac{(f(\text{"}rokok\text{"}|\text{"negatif"}) + 1)}{f(\text{"negatif"}) + |V|} = \frac{6+1}{23+25} = 0.1458$$

$$p(\text{"usaha"}|\text{negatif"}) = \frac{(f(\text{"}usaha\text{"}|\text{"negatif"}) + 1)}{f(\text{"negatif"}) + |V|} = \frac{0+1}{23+25} = 0.0208$$

$$p(\text{"henti"}|\text{negatif"}) = \frac{(f(\text{"}henti\text{"}|\text{"negatif"}) + 1)}{f(\text{"negatif"}) + |V|} = \frac{0+1}{23+25} = 0.0208$$

Setelah diketahui propabilitas kata terhadap kelas positif dan kelas negatif, maka selanjutnya akan dilakukan pemilihan kelas pada dokumen data uji tersebut. Pada penentuan kelas data uji, menggunakan perhitungan 2.4 yaitu :

$$\mathbf{P\left(p/n \mid dt\right) = P\left(p/n\right) * \Pi p(w|p/n)}$$

Keterangan :

$\mathbf{P\left(p/n \mid dt\right)}$ = Pemilihan kelas

$\mathbf{P\left(p/n\right)}$ = Probabilitas kelas positif/ negatif

$\mathbf{\Pi p(w|p/n)}$ = Total *Conditional Probabilities* kata pada kelas positif dan negatif.

Berdasarkan persamaan 2.4, maka pemilihan kelas positif dan negatif pada data uji adalah sebagai berikut :

$$\boldsymbol{p}(\textbf{test}|\textbf{positif}) = p(positif)\ x\ p(belajar|positif) x\ p(rokok|positif) x\ p(usaha|positif)\ x\ p(henti|positif)$$

$$\boldsymbol{p}(\textbf{test}|\textbf{positif}) = {}^{3}/_{9}\ x\ 0.0278\ x\ 0.0556\ x\ 0.0278\ x\ 0.0278 = \mathbf{3.97}$$

$$\boldsymbol{p}(\textbf{test}|\boldsymbol{negatif}) = p(negatif)\ x\ p(belajar|negatif) x\ p(rokok|negatif) x\ p(usaha|negatif)\ x\ p(henti|negatif)$$

$$\boldsymbol{p}(\textbf{test}|\textbf{negatif}) = {}^{6}/_{9}\ x\ 0.0208\ x\ 0{,}0133\ x\ 0.0208\ x\ 0.0208 = \mathbf{8.80}$$

Nilai probabilitas tertinggi adalah pada kategori Positif yaitu sebesar $\mathbf{8.80}$, sehingga komentar tersebut diklasifikasikan ke dalam kategori komentar negatif.

Tabel 4.12 Hasil penentuan kelas data uji (test)

	Tweet **(n)**	**Komentar**	**Kelas**
Data Latih	***Tweet*** **(1)**	Demi Kesehatanmu, Berhentilah Merokok	Negatif
	Tweet **(2)**	Asap Rokok Membuat Nafas ini Sesak	Negatif
	Tweet **(3)**	katakan tidak pada rokok	Negatif
	Tweet **(4)**	Batuknya seorang perokok adalah tanda rusaknya bagian dalam tubuh	Negatif
	Tweet **(5)**	sayangi diri sendiri & orang yang kita cintai dengan tidak merokok	Negatif
	Tweet **(6)**	Dalam nikotinku terdapat kenyamaan selain darimu	Positif

	***Tweet* (7)**	kasian nanti petani pabrik tembakau bangkrut	Positif
	***Tweet* (8)**	Rokok itu punya ciri khas tersendiri	Positif
	***Tweet* (9)**	Turunkan harga rokok	Negatif
Data Uji	***Tweet***	**belajar rokok usaha henti**	**Negatif**

4.6 Proses Pembelajaran dan Model

Pengolahan data dalam penelitian tugas akhir ini menggunakan *tools* Weka, yang terdiri dari proses pembelajaran (*training*) untuk menghasilkan model dan proses pengujian (*testing*). Pada Weka, terdapat 4 *test option*, penjelasannya adalah sebagai berikut :

1. *Use training set*

   Proses pengujian dilakukan dengan menggunakan data latih itu sendiri. Proses ini disebut juga proses pembelajaran yang bertujuan untuk mendapatkan model.

2. *Supplied test set*

   Pengujian dilakukan dengan menggunakan data yang lain atau data uji.

3. *Cross-validation*

   Cross-validation merupakan salah satu proses pengujian pada data, dimana user harus menginput nilai *fold* yang akan digunakan. Nilai *default fold cross-validation* pada Weka adalah 10. Hal ini berarti, data latih dibagi menjadi k buah *subset* (sub himpunan), dimana k adalah nilai dari *fold*. Selanjutnya, untuk tiap dari subset, akan dijadikan data uji dari hasil klasifikasi yang dihasilkan dari k-1 subset lainnya. Jadi, akan ada 10 kali tes. Dimana, setiap data akan menjadi data tes sebanyak 1 kali, dan menjadi datalatih sebanyak k-1 kali.

4. *Percentage split*

   Data akan di pisahkan sebanyak k%, dimana nilai k merupakan masukan dari *user*. Sebagai contoh k = 90%, artinya data akan dipisah sebanyak 90% sebagai

data latih dan sisanya menjadi data uji. Namun dalam penelitian ini tidak menggunakan pilihan pengujian ini karena data latih dan uji telah dipisah sebelumnya.

4.7 Klasifikasi dan evaluasi

Klasifikasi dilakukan terhadap data *training*. Data hanya melalui proses training kemudian data siap untuk diklasifikasikan dan untuk mendapatkan ketelitian digunakan data *confusion matrix* yang merupakan *output* dari klasifikasi NBC.