

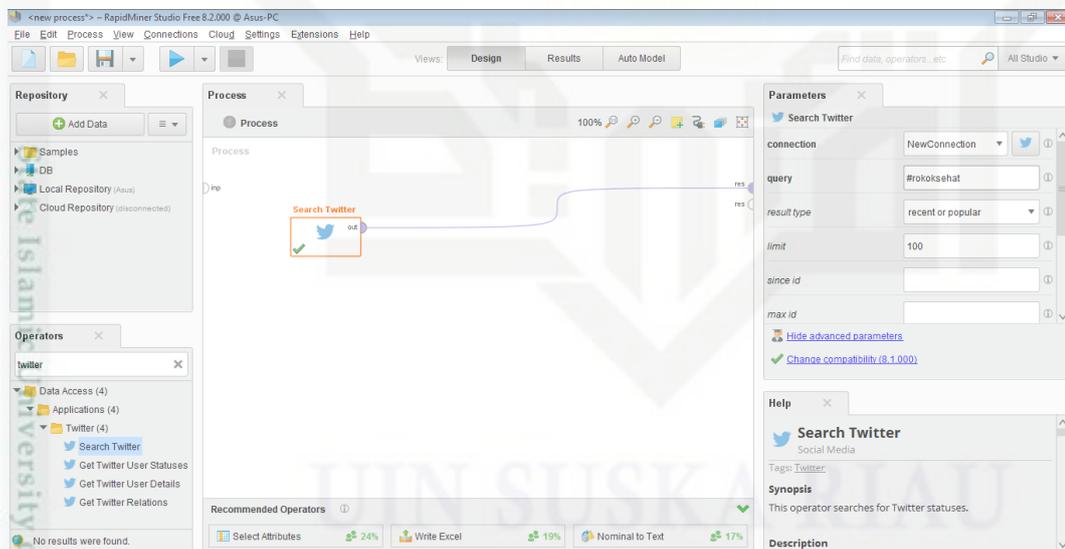
BAB IV

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* :



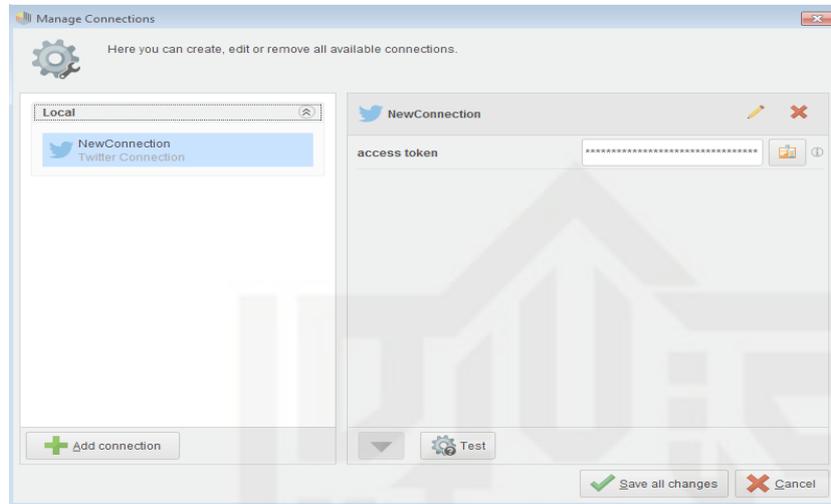
Gambar 4.1 Tampilan Awal Rapidminer

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

Hak Cipta Dilindungi Undang-Undang

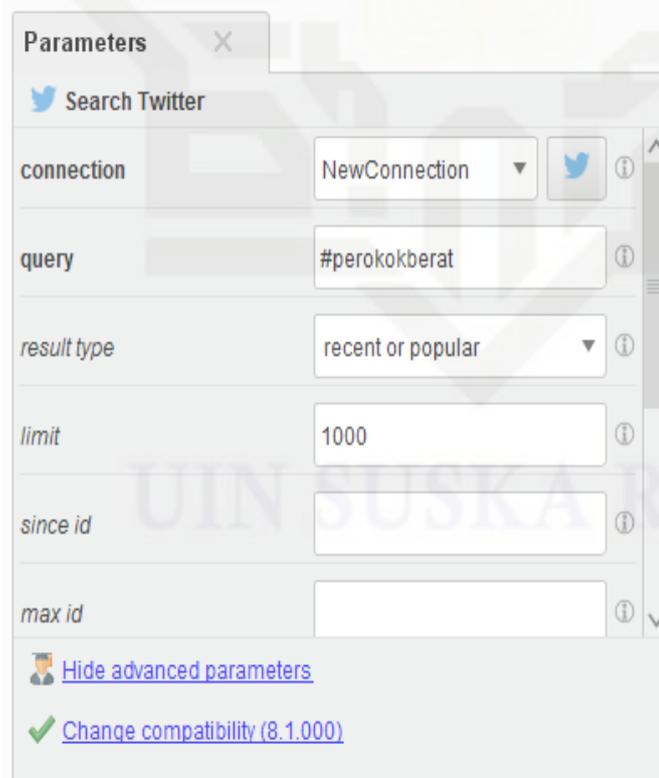
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

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



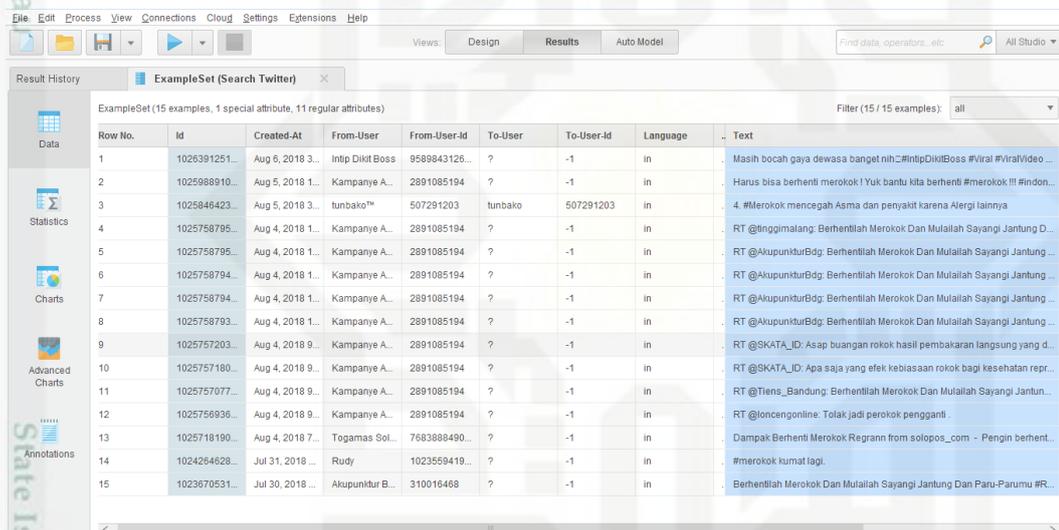
Gambar 4.2 Masukan Token Twitter API

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.



Row No.	Id	Created-At	From-User	From-User-Id	To-User	To-User-Id	Language	Text
1	1026391251...	Aug 6, 2018 3...	Intip Dikit Boss	9589843126...	?	-1	in	Masih bocah gaya dewasa banget nih? #IntipDikitBoss #Viral #ViralVideo ...
2	1025988910...	Aug 5, 2018 1...	Kampanye A...	2891085194	?	-1	in	Harus bisa berhenti merokok! Yuk bantu kita berhenti #merokok !!! #indon...
3	1025846423...	Aug 5, 2018 3...	tunbako™	507291203	tunbako	507291203	in	4 #Merokok mencegah Asma dan penyakit karena Alergi lainnya
4	1025758795...	Aug 4, 2018 1...	Kampanye A...	2891085194	?	-1	in	RT @tinggimalang: Berhentilah Merokok Dan Mulailah Sayangi Jantung D...
5	1025758795...	Aug 4, 2018 1...	Kampanye A...	2891085194	?	-1	in	RT @AkupunkturBdg: Berhentilah Merokok Dan Mulailah Sayangi Jantung ...
6	1025758794...	Aug 4, 2018 1...	Kampanye A...	2891085194	?	-1	in	RT @AkupunkturBdg: Berhentilah Merokok Dan Mulailah Sayangi Jantung ...
7	1025758794...	Aug 4, 2018 1...	Kampanye A...	2891085194	?	-1	in	RT @AkupunkturBdg: Berhentilah Merokok Dan Mulailah Sayangi Jantung ...
8	1025758793...	Aug 4, 2018 1...	Kampanye A...	2891085194	?	-1	in	RT @AkupunkturBdg: Berhentilah Merokok Dan Mulailah Sayangi Jantung ...
9	1025757203...	Aug 4, 2018 9...	Kampanye A...	2891085194	?	-1	in	RT @SKATA_ID: Asap buangan rokok hasil pembakaran langsung yang d...
10	1025757180...	Aug 4, 2018 9...	Kampanye A...	2891085194	?	-1	in	RT @SKATA_ID: Apa saja yang efek kebiasaan rokok bagi kesehatan repr...
11	1025757077...	Aug 4, 2018 9...	Kampanye A...	2891085194	?	-1	in	RT @Tiens_Bandung: Berhentilah Merokok Dan Mulailah Sayangi Jantun...
12	1025756936...	Aug 4, 2018 9...	Kampanye A...	2891085194	?	-1	in	RT @loncengonline: Tolak jadi perokok pengganti .
13	1025718190...	Aug 4, 2018 7...	Togamas Sol...	7683888490...	?	-1	in	Dampak Berhenti Merokok Regram from solopos_com - Penguin berhent...
14	1024264628...	Jul 31, 2018 ...	Rudy	1023559419...	?	-1	in	#merokok kumat lagi.
15	1023670531...	Jul 30, 2018 ...	Akupunktur B...	310016468	?	-1	in	Berhentilah Merokok Dan Mulailah Sayangi Jantung Dan Paru-Parumu #R...

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

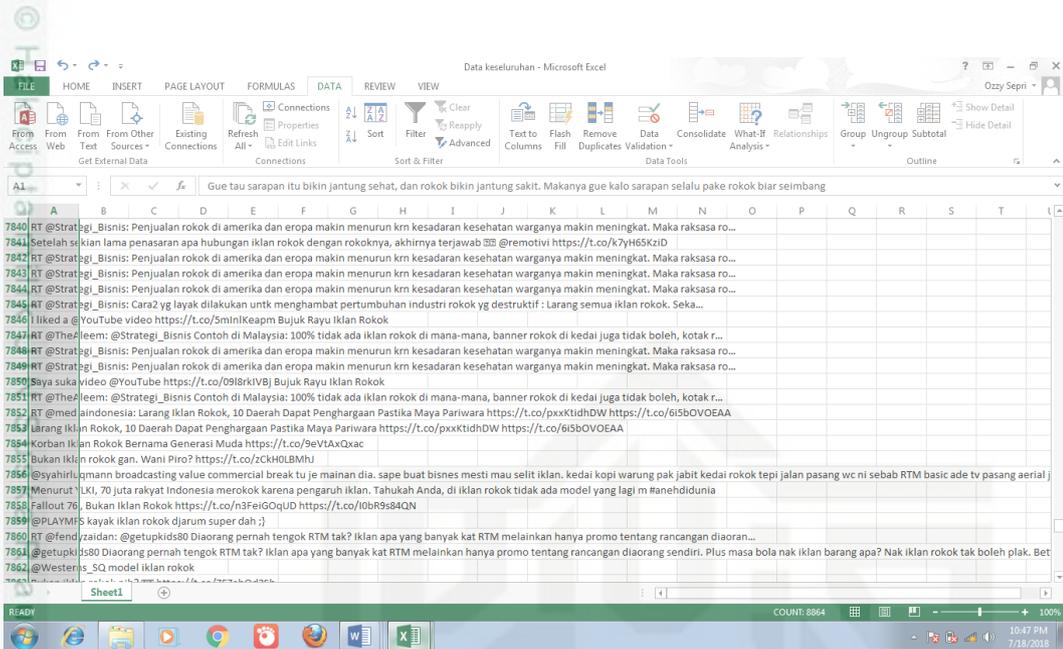
Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

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	#kawasanparokok	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

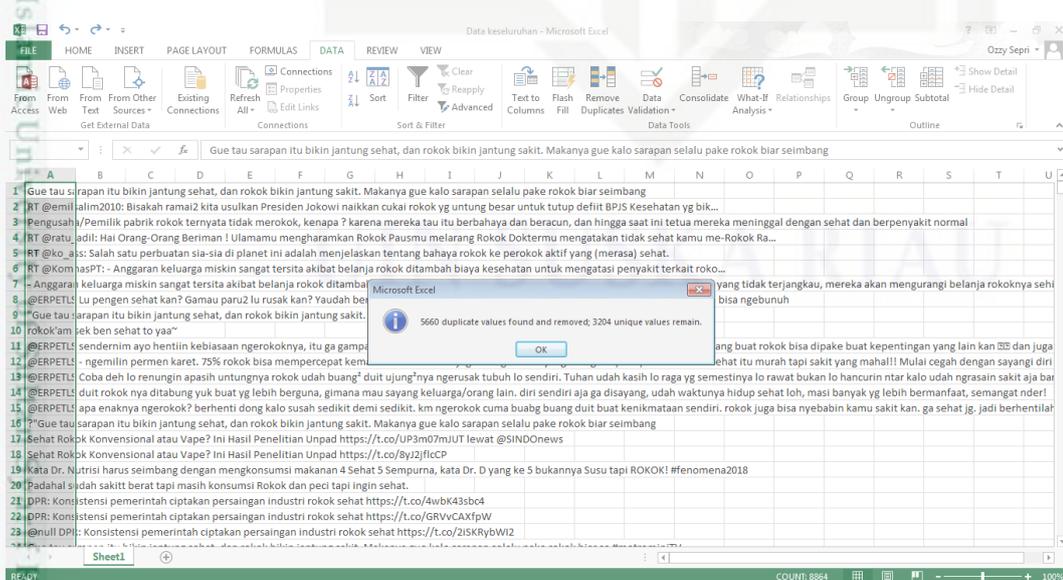
Jumlah data keseluruhan yang didapat pada API Twitter sebanyak 8.864 data *tweet*. Data komentar *tweet* yang telah didapat di export lalu disimpan dalam bentuk format .CSV. Berikut ini adalah contoh data *tweet* yang telah disimpan dalam bentuk format .CSV :

- Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
 2. Dilarang mengemukakan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.



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 microsoft excel. Dengan cara blok seluruh data *tweet* yang akan dilakukan redundancy. Kemudian pilih pada remove duplicates pada Microsoft excel. Dari proses redundancy tadi telah dibuang data yang sama sebanyak 5.660 sehingga didapat hasil *tweet* menjadi 3.204 dari *tweet* awal 8.864. Berikut adalah gambar redundancy data *tweet*

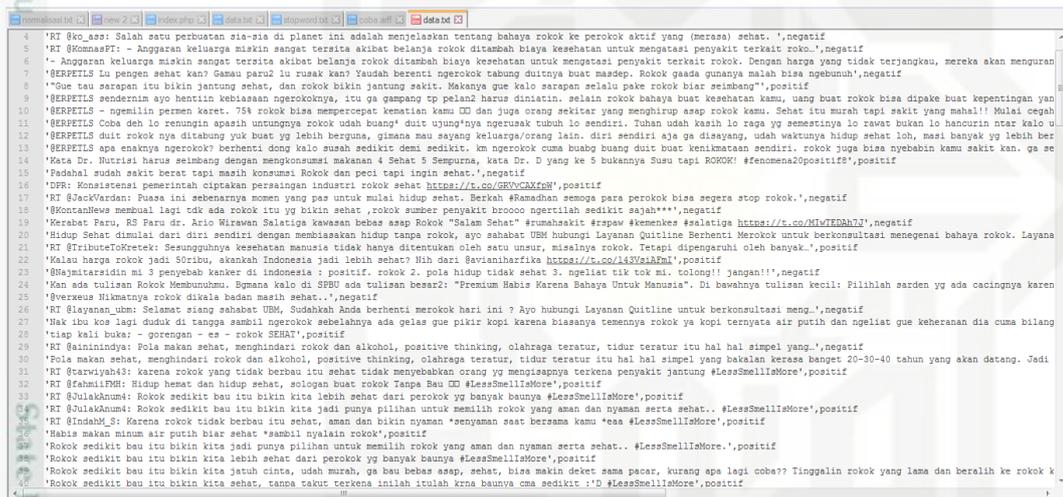


Gambar 4.6 hasil pembuangan data yang sama

Data *tweet* yang telah di redundancy dilakukan pelabelan, 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:



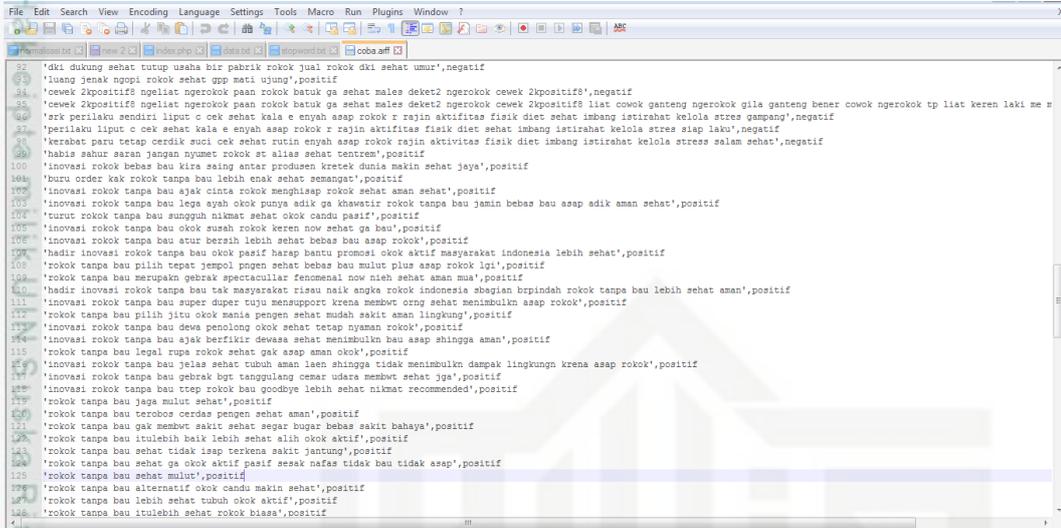
```

4 'RT @ko_ass: Salah satu perubahan sia-sia di planet ini adalah menjelaskan tentang bahaya rokok ke perokok aktif yang (merasa) sehat. ',negatif
5 'RT @komasPT: - Anggaran keluarga miskin sangat terancam akibat belanja rokok ditambah biaya kesehatan untuk mengatasi penyakit terkait rokok.',negatif
6 '- Anggaran keluarga miskin sangat terancam akibat belanja rokok ditambah biaya kesehatan untuk mengatasi penyakit terkait rokok. Dengan harga yang tidak terjangkau, mereka akan menguras
7 'BERPETS Lu pengen sehat kan? Gausa paru2 lu rusak kan? Yaudah berenti ngerokok tabung duitnya buat masdep. Rokok gaada gunanya malah bisa ngebunuh',negatif
8 '**Que tau serapan itu bikin jantung sehat, dan rokok bikin jantung sakit. Makanya gue kalo sarapan selalu pake rokok biar seimbang'',positif
9 'BERPETS sandermin ayo hentuin kebiasaan ngerokoknya, itu ga gampang tp pelan2 harus diniatin. selain rokok bahaya buat kesehatan kamu, uang buat rokok bisa dipake buat kepentingan yan
10 'BERPETS - ngemilin permen karet. 75k rokok bisa mempercep Kemiatan kamu UD dan juga orang sekitar yang menghirup asap rokok kamu. Sehat itu murah tapi sakit yang mahal!! Hala! cegah
11 'BERPETS Coba deh lo renungin apasih urungnya rokok udah buang' duit yg'ny'nya ngerusak tubuh lo sendiri. Tahun udah kasih lo raga yg semestinya lo rawat bukan lo hancurin otar kalo ui
12 'BERPETS duit rokok nya ditabung yuk buat yg lebih berguna, gimana mau sayang keluarga/orang lain. diri sendiri aja ga disayang, udah waktunya hidup sehat loh, masi banyak yg lebih ber
13 'BERPETS apa anaknya ngerokok? berhenti dong kalo susah sedikit demi sedikit. km ngerokok cuma buabg duit buat kenikmatan sendiri. rokok juga bisa nyebabin kamu sakit kan. ga se
14 *Hata Dr. Nutrisi harus seimbang dengan mengkonsumsi makanan 4 Sehat 5 Sempurna, kata Dr. D yang ke 5 bukannya Susu tapi ROKOK! #fenomena2positif0',positif
15 'Psdahal sudah sakit berat tapi masih konsumsi Rokok dan peci tapi ingin sehat.',negatif
16 'DPR: Konsistensi pemerintah ciptakan persaingan industri rokok sehat https://t.co/8WVCMXf9n',positif
17 'RT @JasKardian: Puasa ini sebenarnya momen yang pas untuk mulai hidup sehat. Berkah #Ramadhan semoga para perokok bisa segera stop rokok.',negatif
18 'Kontanews membal lagi tdk ada rokok itu yg bikin sehat ,rokok sumber penyakit broooo ngertilah sedikit sajah'',negatif
19 'Herabot Paru, RS Paru dr. Ario Mirewan Salatiqa Kawasan bebas asap Rokok "Salam Sehat" #rumahsehat #rspaw #kemenkes #salatiqa https://t.co/4WU7EhBdLJ',negatif
20 'hidup Sehat dimulai dari diri sendiri dengan membiasakan hidup tanpa rokok, ayo sahabat URM hubungi Layanan Quitline Berhenti Merokok untuk berkonsultasi mengenai bahaya rokok. Layana
21 'RT @ributeTofretek: Seungguhnya kesehatan manusia tidak hanya ditentukan oleh satu unsur, misalnya rokok. Tetapi dipengaruhi oleh banyak.',positif
22 'Kalaupun harga rokok jadi 50ribu, akankah Indonesia jadi lebih sehat? Mh dari @avianharfika https://t.co/143Vas4mI',positif
23 'Beyntaridin ml 3 penyebab kanker di Indonesia : positif. rokok 2. pola hidup tidak sehat 3. ngeliat tik tok ml. tolong!! jangan!!',negatif
24 'Hn ada tulisan Rokok Membunuhmu. Emana kalo di SFU ada tulisan besar2: "Premiun Besar Karena Bahaya Untuk Manusia". Di bewalnya tulisan kecil: Pilihlah garden yg ada cacingnya karen
25 'Overkses Nikmatnya rokok dikala badan masih sehat.',negatif
26 'RT @layanan_umi: Selamat siang sahabat URM, Sudahkah Anda berhenti merokok hari ini ? Ayo hubungi Layanan Quitline untuk berkonsultasi meng.',negatif
27 'Hak ibu kos lagi duduk di tangga sambil ngerokok sebelahnya ada gelas kopi pikir kopi karena biasanya temennya rokok ya kopi ternyata air putih dan ngeliat gue keheranan dia cuma bilang
28 'sip kali bukas - gorengan - es - rokok SEHAT',positif
29 'RT @ainindiyah: Pola makan sehat, hindari rokok dan alkohol, positive thinking, olahraga teratur, tidur teratur itu hal hal simpel yang bakal kerasa banget 20-30-40 tahun yang akan datang. Jadi
30 'Pola makan sehat, hindari rokok dan alkohol, positive thinking, olahraga teratur, tidur teratur itu hal hal simpel yang bakal kerasa banget 20-30-40 tahun yang akan datang. Jadi
31 'RT @arvivyah43: karena rokok yang tidak berbau itu sehat tidak menyebabkan orang yg mengisapnya terkena penyakit jantung #LessSmellMore',positif
32 'RT @famliRM: Hidup hemat dan hidup sehat, sologan buat rokok Tanpa Bau UD #LessSmellMore',positif
33 'RT @JalakKunni: Rokok sedikit bau itu bikin kita lebih sehat dari perokok yg banyak buanya #LessSmellMore',positif
34 'RT @Kindahh_S: Karena rokok tidak berbau itu sehat, aman dan bikin nyaman *nyaman saat berzama kamu *eas #LessSmellMore',positif
35 'Habis makan minum air putih biar sehat *sambil nyalin rokok',positif
36 'Rokok sedikit bau itu bikin kita jadi punya pilihan untuk memilih rokok yang aman dan nyaman serta sehat.. #LessSmellMore',positif
37 'Rokok sedikit bau itu bikin kita lebih sehat dari perokok yg banyak buanya #LessSmellMore',positif
38 'Rokok sedikit bau itu bikin kita lebih sehat dari perokok yg banyak buanya #LessSmellMore',positif
39 'Rokok sedikit bau itu bikin kita lebih sehat dari perokok yg banyak buanya #LessSmellMore',positif
40 'Rokok sedikit bau itu bikin kita lebih sehat, udah murah, ga bau bebas asap, sehat, bisa makin dekat sama pacar, kurang apa lagi coba?? Tinggalin rokok yang lama dan beralih ke rokok k
41 'Rokok sedikit bau itu bikin kita sehat, tanpa takut terkena inlah itulah krna bauanya sma sedikit :D #LessSmellMore',positif

```

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 :

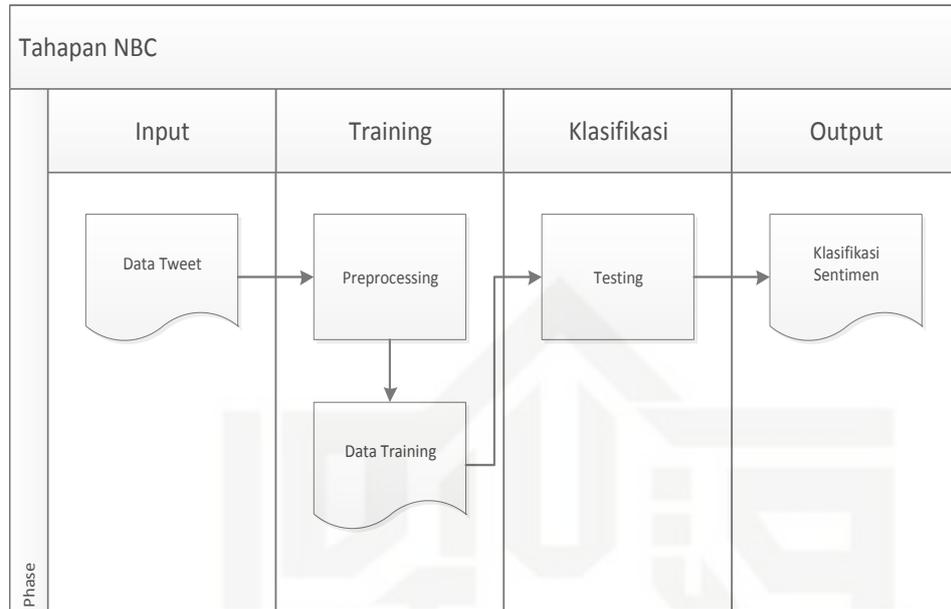


Gambar 4.8 Format data tweet *.arff

Pada gambar terlihat bahwa ada @relation, @attribute dan @data. @relation menunjukkan 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 :

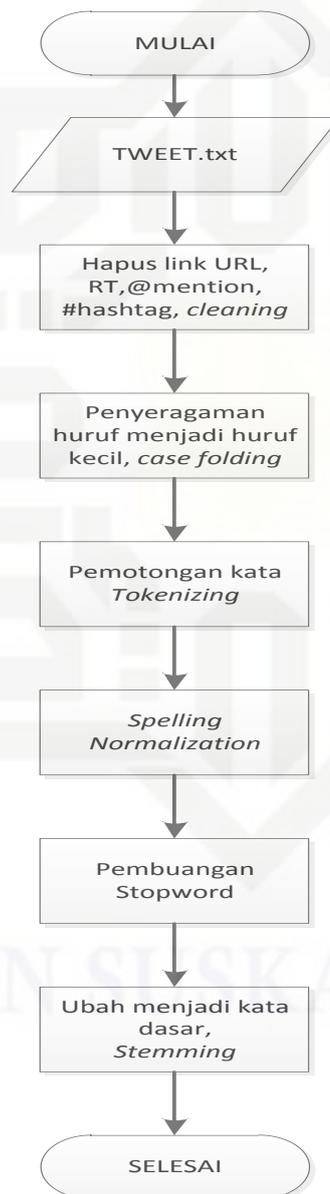
1. **Input**
Bagian input dari penelitian ini adalah seluruh data *tweet* yang berjumlah 1000 dan sudah dilabel manual serta dibagi menjadi 90% dari 1000 data sebagai data latih dan 10% dari 1000 data sebagai data uji. Adapun data *tweet* tersebut harus sesuai dengan format masukan Weka.
2. **Training**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

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

4.3 Text Preprocessing

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 Preprocessing

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

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

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

<i>Tweet</i> (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*

<i>Tweet</i> (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 kenyamanan selain darimu	Positif
7	kasian nanti petani pabrik tembakau bangkrut	Positif
8	Rokok itu punya ciri khas tersendiri	Positif
9	Turunkan harga rokok	Negatif

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

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

<i>Tweet</i> (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 kenyamanan 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

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

Tabel 4.6 Hasil Spelling Normalization

<i>sTweet</i> t 1	<i>Tweet</i> 2	<i>Tweet</i> 3	<i>Tweet</i> 4	<i>Tweet</i> 5	<i>Tweet</i> 6	<i>Tweet</i> 7	<i>Tweet</i> 8	<i>Tweet</i> 9
demi	asap	Kataka na	batukny a	sayangi	dalam	kasihan	rokok	turunka n
kesehat anmu	rokok	tidak	seorang	diri	nikotin ku	nanti	itu	harga
berhenti lah	membu at	pada	perokok	sendiri	terdapat	petani	punya	rokok
meroko k	nafas	rokok	adalah	orang	kenyam anan	pabrik	ciri	
	ini		tanda	yang	selain	tembak au	khas	
	sesak		rusaknya	kita	dirimu	bangkru t	tersendi ri	
			bagian	cintai				
			dalam	dengan				
			tubuh	tidak				
				meroko k				

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).

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

Tabel 4.7 Hasil *Filtering*

<i>Tweet</i> 1	<i>Tweet</i> 2	<i>Tweet</i> 3	<i>Tweet</i> 4	<i>Tweet</i> 5	<i>Tweet</i> 6	<i>Tweet</i> 7	<i>Tweet</i> 8	<i>Tweet</i> 9
demi	asap	Tidak	batuknya	sayangi	nikotinku	kasihan	rokok	turunkan
kesehatannya	rokok	Rokok	perokok	diri	kenyamanan	petani	khas	harga
berhenti	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*

<i>Tweet</i> 1	<i>Tweet</i> 2	<i>Tweet</i> 3	<i>Tweet</i> 4	<i>Tweet</i> 5	<i>Tweet</i> 6	<i>Tweet</i> 7	<i>Tweet</i> 8	<i>Tweet</i> 9
demi	asap	tidak	batuk	sayang	nikotin	kasih	rokok	turun
kesehatannya	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

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

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.

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

- Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

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

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

2 Tahap Uji (*testing*)

Data uji

Komentar	Kelas
Dulu belajar merokok, Kali ini berusaha belajar berhenti merokok!!! <u>#sehattanparokok</u> <u>#stopmerokok</u> <u>#bahayamerokok</u> .	?

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 persamaan 2.3 sebagai berikut :

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

Diketahui IVI =

Count Positif = 11, count negatif = 23

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

• Probabilitas data uji kelas positif

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

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

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

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

• Probabilitas data uji kelas negatif

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

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

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

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

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 :

$$P(p/n | dt) = P(p/n) * \Pi p(w|p/n)$$

Keterangan :

$P(p/n | dt)$ = Pemilihan kelas

$P(p/n)$ = Probabilitas kelas positif/ negatif

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

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

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

$$p(\text{test}|\text{positif}) = p(\text{positif}) \times p(\text{belajar}|\text{positif}) \times p(\text{rokok}|\text{positif}) \times p(\text{usaha}|\text{positif}) \times p(\text{henti}|\text{positif})$$

$$p(\text{test}|\text{positif}) = \frac{3}{9} \times 0.0278 \times 0.0556 \times 0.0278 \times 0.0278 = 3.97$$

$$p(\text{test}|\text{negatif}) = p(\text{negatif}) \times p(\text{belajar}|\text{negatif}) \times p(\text{rokok}|\text{negatif}) \times p(\text{usaha}|\text{negatif}) \times p(\text{henti}|\text{negatif})$$

$$p(\text{test}|\text{negatif}) = \frac{6}{9} \times 0.0208 \times 0.0133 \times 0.0208 \times 0.0208 = 8.80$$

Nilai probabilitas tertinggi adalah pada kategori Positif yaitu sebesar **8.80**, sehingga komentar tersebut diklasifikasikan ke dalam kategori komentar negatif.

Tabel 4.12 Hasil penentuan kelas data uji (test)

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

	<i>Tweet (7)</i>	kasian nanti petani pabrik tembakau bangkrut	Positif
	<i>Tweet (8)</i>	Rokok itu punya ciri khas tersendiri	Positif
	<i>Tweet (9)</i>	Turunkan harga rokok	Negatif
Data Uji	<i>Tweet</i>	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 data latih 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

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

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.