Jurnal Informatika Universitas Pamulang **ISSN: 2541-1004**
Vol. 4, No. 3, September 2019 (87-94) **e-ISSN: 2622-4615**

Analisis Sentiment *Tweets* Berbahasa Sunda Menggunakan *Naive Bayes Classifier* dengan Seleksi *Feature Chi Squared Statistic*

Yono Cahyono[1], Saprudin[2]

Teknik Informatika Universitas Pamulang
e-mail : [1]dosen00843@unpam.com, [2]dosen00845@unpam.ac.id

Submitted Date: September 3rd, 2019 Reviewed Date: September 6th, 2019
Revised Date: September 16th, 2019 Accepted Date: October 4th, 2019

Abstract

At present the development of the use of social media in Indonesia is very rapid, in Indonesia there are a variety of regional languages, one of which is the Sundanese language, where some people especially those living in West Java use Sundanese language to express comments, opinions, suggestions, criticisms and others in social media. This information can be used as valuable data for individuals or organizations in decision making. The huge amount of data makes it impossible for humans to read and analyze it manually. Sentiment analysis is the process of classifying opinions, analyzing, understanding, evaluating, emotions and attitudes towards a particular entity such as individuals, organizations, products or services, topics, events, in order to obtain information. The purpose of this research is the Naïve Bayes Classifier (NBC) classification algorithm and Feature Chi Squared Statistics selection method can be used in Sundanese-language tweets sentiment analysis on Twitter social media into positive, negative and neutral categories. Chi Square Statistic feature test results can reduce irrelevant features in the Naïve Bayes Classifier classification process on Sundanese-language tweets with an accuracy of 78.48%.

Keywords: Sentiment Analysis, Sundanese, Twitter, Naïve Bayes Classifier (NBC), *Chi Squared Statistic*

Abstrak

Saat ini perkembangan penggunaan media sosial di Indonesia sangat pesat, di Indonesia terdapat berbagai bahasa daerah, salah satunya adalah bahasa Sunda, di mana sebagian orang terutama yang tinggal di Jawa Barat menggunakan bahasa Sunda untuk menyampaikan komentar, opini, saran, kritik dan lain-lain di media sosial. Informasi ini dapat digunakan sebagai data berharga bagi individu atau organisasi dalam pengambilan keputusan. Jumlah data yang sangat besar membuat manusia tidak dapat membaca dan menganalisisnya secara manual. Analisis sentimen adalah proses mengklasifikasikan pendapat, menganalisis, memahami, mengevaluasi, emosi dan sikap terhadap entitas tertentu seperti individu, organisasi, produk atau layanan, topik, peristiwa, untuk mendapatkan informasi. Tujuan dari penelitian ini adalah algoritma klasifikasi *Naïve Bayes Classifier* (NBC) dan metode pemilihan Fitur *Chi Squared Statistics* dapat digunakan dalam analisis sentimen *tweets* berbahasa Sunda di media sosial *Twitter* ke dalam kategori positif, negatif dan netral. Hasil uji fitur *Chi Square Statistic* dapat mengurangi fitur yang tidak relevan dalam proses klasifikasi *Naïve Bayes Classifier* pada *tweets* berbahasa Sunda dengan akurasi 78,48%.

Kata kunci: Analisis Sentiment, Bahasa Sunda, *Twitter, Naïve Bayes Classifier* (NBC), *Chi Squared Statistic*

1. Pendahuluan

Saat ini pengguna media sosial *Twitter* sangat banyak. Indonesia terdapat berbagai pulau dengan beragam bahasa, pengguna media sosial *Twitter* di Indonesia tidak hanya penggunaan bahasa Indonesia untuk komunikasi, tapi juga penggunaan bahasa daerah seperti di daerah Jawa Barat menggunakan bahasa sunda. Perkembangan media *online* seperti media sosial *Twitter* dengan informasi yang tidak terbatas, menyebabkan kebutuhan untuk menggali informasi yang ada didalamnya. Pada *Twitter* terdapat istilah *tweet*

yang merupakan sebuah pesan atau status yang dibuat oleh penggunanya. Sebuah *tweet* dapat mengekspresikan sebuah perasaan atau keadaan dari pengguna *Twiter*. *Tweet* dapat mengandung sebuah opini dari penggunanya terhadap kejadian yang dialaminya. Opini tersebut dapat dimanfaatkan sebagai penilaian baik bagi perorangan atau bagi perusahaan atau instansi.

Twitter banyak digunakan oleh instansi pemerintahan atau perusahaan sebagai media komunikasi dengan konsumen. Namun menentukan dan memilah apakah suatu *tweet* berbahasa sunda dapat mengandung sebuah opini positif, negatif atau netral tentu tidak mudah jika *tweet* yang diteliti jumlahnya sangat banyak. Maka dengan permasalahan tersebut dapat di bangun sebuah sistem yang dapat melakukan analisis sentimen.

Dari beberapa teknik klasifikasi yang sering digunakan dalam proses klasifikasi data yaitu metode *Naïve Bayes* atau sering disebut dengan *Naïve Bayes Classifier* (NBC). Algoritma *Naïve Bayes* dipilih dikarenakan algoritma ini sangat cocok untuk *short* data *text*. Kelebihan dari *Naïve Bayes Classifier* adalah metode ini sederhana tetapi memiliki akurasi dan performansi yang tinggi dalam proses klasifikasi teks (Routray, 2013).

Sedangkan seleksi fitur adalah proses optimasi untuk mengurangi suatu set besar fitur dari sumber aslinya, agar diperoleh sejumlah subset fitur yang relatif kecil dan signifikan untuk meningkatkan akurasi dalam proses klasifikasi.

Untuk itu dalam penelitian ini menggunakan penggabungan metode pemilihan fitur *Chi Squared Statistic* dan *Naïve Bayes Classifier* (NBC) untuk analisis sentiment *tweets* berbahasa sunda.

2. Landasan Teori

2.1 *Sentiment Analysis*

Sentiment analysis atau disebut *opinion mining* dalam pengertian secara luas mengacu pada bidang komputasi linguistik, pengolahan bahasa alami dan *text mining*. Secara umum, tujuannya adalah untuk menentukan *attitude* dari pembicara ataupun penulis yang berhubungan dengan topik tertentu.

Pengelompokkan polaritas dari teks yang ada dalam kalimat, dokumen, atau fitur entitas. Bagaimana pendapat yang disampaikan dalam kalimat, dokumen atau fitur entitas bersifat positif, negatif atau netral merupakan tugas dasar dalam analisis sentimen (Dehaff, 2010). Lebih luas lagi *sentiment analysis* dapat mengungkapkan emosional gembira, sedih, marah dan lain-lain.

Kita juga dapat mengetahui misalnya seperti merek, produk-produk dan orang-orang dengan menentukan apakah mereka menilai positif atau negatif. Dengan ini memungkinkan bisnis dalam melacak:

a. Persepsi produk baru.
b. Deteksi *Flame* (*rants* buruk)
c. Persepsi Merek.
d. Reputasi manajemen.

Memungkinkan juga individu untuk memperoleh mengenai suatu pandangan *(review)* pada skala global (Jenkins, 2011).

Sentiment atau ekspresi fokus mengacu pada topik tertentu, pernyataan pada sesuatu topik mungkin saja akan berbeda arti/makna dengan pernyataan yang sama pada subjek yang berbeda. Sebagai contoh, misalnya pernyataan yang baik untuk mengatakan "alur film tidak terprediksi", tapi lain halnya pernyataan tidak baik jika 'tidak terprediksi' dinyatakan pada "kemudi dari kendaraan tidak terprediksi". Bahkan pada produk tertentu, kata-kata yang sama dapat menggambarkan makna sebaliknya. Oleh karena itu beberapa penelitian, terutama mengenai *review* produk, pekerjaan dimulai dengan menentukan elemen dari suatu produk yang sedang dibicarakan sebelum dimulai proses *opinion mining*.

Hal utama yang dilakukan dalam pemrosesan dokumen adalah memisahkan kumpulan karakter ke dalam kata *(token)* atau biasa disebut sebagai *tokenizing*. *Tokenizing* adalah merupakan hal yang sangat kompleks dalam sistem komputerisasi karena beberapa karakter dapat ditemukan sebagai *token delimiters*. *Delimite*r adalah karakter tab, spasi dan baris baru/*newline*, sedangkan seperti karakter "! ? ( ) < >" kadangkala dijadikan *delimiter* namun kadangkala bukan tergantung pada lingkungannya (Wulandini & Nugroho, 2009).

2.2 *Naïve Bayes Classifier*

Algoritma *Naive Bayes Classifier* merupakan algoritma yang digunakan dalam mencari nilai probabilitas tertinggi untuk mengklasifikasikan data uji pada kategori yang paling tepat (Feldman & Sanger, 2007).

Pada algoritma *Naïve Bayes Classifier* setiap dokumen direpresentasikan dengan pasangan atribut "*x1, x2, x3,...xn*" di mana *x1* adalah kata pertama, *x2* adalah kata kedua dan seterusnya. Sedangkan *V* merupakan himpunan pada kategori *tweet*. Pada saat proses klasifikasi algoritma akan

mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan *(V_{MAP})*, di mana persamaannya adalah sebagai berikut:

$$V_{MAP} = argmax_{VjeV} \frac{P(X_1,X_2,X_3,\ldots\ldots X_n|V_j)P(V_j)}{P(X_1,X_2,X_3,\ldots\ldots X_n)} \quad (2.1)$$

Untuk *P(x1, x2, x3,...xn)* nilainya konstan untuk semua kategori *(Vj)* sehingga persamaannya dapat ditulis sebagai berikut:

$$V_{MAP} = \underset{VjeV}{argmax}\, P(x_1,x_2,x_3,\ldots x_n|V_j)P(V_j) \quad (2.2)$$

Persamaan di atas dapat disederhanakan lagi menjadi sebagai berikut:

$$V_{MAP} = \underset{VjeV}{argmax} \prod_{i-1}^{n} P(x_i|V_j)P(V_j) \quad (2.3)$$

Keterangan :

Vj = Kategori *tweet j =1, 2, 3,n*. Di mana dalam penelitian ini *j1* = kategori *tweet* sentimen negatif, *j2* = kategori *tweet* sentimen positif, dan *j3* = kategori *tweet* sentiment netral

P(xi|Vj)=Probabilitas *xi* pada kategori *Vj*

P(Vj) = Probabilitas dari *Vj*

Untuk *P(Vj)* dan *P(xi|Vj)* dihitung pada saat pelatihan di mana persamaannya adalah sebagai berikut:

$$P(V_j) = \frac{|docs\ j|}{|contoh|} \quad (2.4)$$

$$P(X_i|V_j) = \frac{nk+1}{n+|kosakata|} \quad (2.5)$$

Keterangan:

|*docs* j| = merupakan jumlah dokumen pada setiap kategori *j*

|*contoh*| = merupakan jumlah dokumen dari semua kategori

nk = merupakan jumlah frekuensi kemunculan setiap kata

n = merupakan jumlah frekuensi kemunculan kata dari setiap kategori

|*kosakata*| = merupakan jumlah semua kata dari semua kategori

3. Metode Penelitian

3.1 Pengumpulan Data (*Dataset*)

Data yang digunakan untuk proses *sentiment analysis* ini didapatkan dengan cara *crawl* (mengumpulkan) data dari media sosial *Twitter*. Media sosial *Twitter* dipilih mengingat media sosial *Twitter* saat ini merupakan salah satu media sosial yang populer dan banyak digunakan dalam mengungkapkan pendapat atau opini mengenai sesuatu hal.

3.2 *Pre-Processing*

Pre-processing (pemrosesan awal dokumen) merupakan tahapan awal yang berfungsi dalam mentransformasikan dokumen ke dalam bentuk representasi yang lain. Tujuan dari tahap ini adalah untuk mempermudah untuk proses pencarian *query* ke dalam dokumen, mempercepat dalam pemrosesan terhadap dokumen, dan mempermudah dalam proses mengurutkan dokumen-dokumen yang diambil (*retrieved*). Tahapan proses yang dilakukan dalam *pre-processing* adalah *casefolding, tokenize* dan *stopword removal* (Berry & Kogan, 2010).

a. Case folding

Pada tahapan ini dilakukan untuk pengubah huruf dalam dokumen menjadi huruf kecil. Yang diterima hanya huruf 'a' sampai dengan 'z'. Sedangkan karakter selain huruf dianggap sebagai *delimiter* dan dihilangkan.

b. Tokenizing

Tahapan ini dilakukan setelah *input* data uji melewati tahap *Case Folding*. Di mana *Tokenizing* merupakan proses pemotongan *string input* berdasarkan tiap kata yang menyusunya, serta membedakan karakter-karakter tertentu yang dapat diperlakukan sebagai pemisah kata atau bukan.

c. *Stopwords removal*

Selanjutnya untuk tahap *filter stopwords (dictionary)* adalah akan menghilangkan kata-kata pada daftar *stopwords* yang tidak memiliki arti.

3.3 *Seleksi fitur Chi Square Statistic*

Pada penelitian ini menggunakan seleksi fitur *Chi Square Statistic*. Dengan cara menghitung nilai seleksi fitur *Chi Square Statistic* dengan persamaan sebagai berikut (Yang & Pedersen, 1997):

$$X^2(t,c) = \frac{N\,x\,(AD-CB)^2}{(A+C)x(B+D)x(A+B)x(C+D)} \quad (3.1)$$

Keterangan:

A : Banyaknya dokumen dalam kategori *c* yang mengandung *term t*

B : Banyaknya dokumen yang bukan kategori *c* tetapi mengandung *term t*

C : Banyaknya dokumen dalam katgori *c* tetapi tidak mengandung *term t*

D : Banyaknya dokumen yang bukan kategori *c* dan tidak mengandung *term t*
N : Total keseluruhan dokumen

Seleksi fitur *Chi Square Statistic* digunakan untuk melakukan kesesuaian pegamatan (*goodness of fit*) dari kategori dengan *terms*. Uji *Chi Square Statistic* dalam statistika diterapkan untuk menguji independensi dari dua peristiwa. Sedangkan dalam seleksi fitur berdasarkan teori statistika, dua peristiwa tersebut di antaranya adalah kemunculan fitur dan kemunculan kategori.

3.4 *Tahap Cross Validation*

Dalam tahap *cross-validation*, setiap *record* akan digunakan beberapa kali yaitu untuk data *training* dan untuk data *testing*. Untuk dapat mengilustrasikan pada metode ini, anggaplah data akan dipartisi ke dalam dua *subset*. Yang pertama dipilih satu *subset* tersebut untuk *training* dan satu lagi untuk *testing*. Kemudian akan dilakukan pertukaran fungsi dari *subset* sehingga *subset* yang sebelumnya sebagai *training set* akan menjadi *testing set* demikian sebaliknya.

10-fold cross-validation akan melakukan mengulang pengujian sebanyak 10 kali dan hasil pengukuran adalah nilai rata-rata dari 10 kali pengujian. Metode ini merupakan evaluasi standar yaitu *stratified 10-fold cross-validation* karena menunjukkan bahwa *10-fold cross-validation* adalah merupakan pilihan terbaik untuk mendapatkan hasil validasi yang lebih akurat. Keuntungan dari metode ini adalah menghindari *overlapping* pada data *testing*. *Test set* bersifat *mutually exclusive* dan secara efektif mencakup keseluruhan *data set*. Namun kekurangan dari pendekatan ini ialah banyaknya proses komputasi untuk melakuan pengulangan prosedur sebanyak *N* kali. (Gorunescu, 2011).

3.5 *Tahap Klasifikasi*

Pada tahap klasifikasi menggunakan algoritma *Naive Bayes Classifier* yang merupakan proses klasifikasi dengan metode probabilitas, yaitu memprediksi peluang pada masa yang akan datang, berdasarkan pengalaman pada masa lalu sehingga dikenal sebagai *teorema Bayes*.

3.6 *Evaluasi*

Evaluasi performansi dilakukan untuk menguji hasil dari proses klasifikasi dengan cara mengukur nilai performansi dari sistem yang telah dibuat. Parameter pengujian yang digunakan untuk evaluasi adalah diperoleh dari tabel *Confussion Matrix* untuk perhitungan tingkat akurasinya,

Pada Tabel 1, untuk *True* Positif (TP) merupakan tupel positif di *dataset* yang diklasifikasikan positif, sedangkan *False* Positif (FP) adalah tupel positif di *dataset* yang diklasifikasikan negatif atau netral. Untuk *True* Negatif (TN) merupakan tupel negatif di *dataset* yang diklasifikasikan negatif, sedangkan *False* Negatif (FN) merupakan jumlah tupel negatif di *dataset* yang diklasfikasikan positif atau netral. Untuk *True* Netral (TNet) merupakan tupel netral di *dataset* yang diklasifikasikan netral, sedangkan *False* Netral (FNet) adalah tupel netral di *dataset* yang diklasifikasikan positif atau negatif.

Tabel 1 *Confussion Matrix*

	true **Negatif**	***true*** **Positif**	***true*** **Netral**
pred. **Negatif**	TN	FP	FNet
pred. **Positif**	FN	TP	FNet
pred. **Netral**	FN	FP	TNet

Selanjutnya untuk menghitung *accuracy*, *Recall Positives*, *Recall Negatives*, *Recall Neutral*, *Positives Predicted Value (PPV)*, *Negatives Predicted Value (NPV)*, *Neutral Predicted Value (NetPV)*.

$$Accuracy = \frac{TP+TN+TNet}{TP+TN+FP+FN+FNet+TNet} \quad (3.2)$$

$$Recall\ Positives = \frac{TP}{TP+FN+FNet} \quad (3.3)$$

$$Recall\ Negatives = \frac{TN}{TN+FP+FNet} \quad (3.4)$$

$$Recall\ Neutral = \frac{TNet}{TNet+FP+FN} \quad (3.5)$$

$$PPV = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives + Number\ of\ False\ Positives} \quad (3.6)$$

$$NPV = \frac{Number\ of\ True\ Negatives}{Number\ of\ True\ Negatives + Number\ of\ False\ Negatives} \quad (3.7)$$

$$NetPV = \frac{Number\ of\ True\ Neutral}{Number\ of\ True\ Neutral + Number\ of\ False\ Neutral} \quad (3.8)$$

Recall Positives (perolehan positif) adalah jumlah kasus dengan perolehan positif, *Recall Negatives* (perolehan negatif) adalah jumlah kasus dengan perolehan negaitif, dan *Recall Neutral* (perolehan netral) adalah jumlah kasus dengan perolehan netral. , *NPV* (nilai prediktif negatif) adalah jumlah kasus dengan hasil diagnosa negatif, *PPV* (nilai prediktif positif) adalah jumlah kasus dengan hasil diagnosa positif dan *NetPV* (nilai prediktif netral) adalah jumlah kasus dengan hasil diagnosa netral.

4. Hasil Penelitian

4.1 *Implementasi Menggunakan RapidMiner*

Penelitian ini dilakukan dengan menggunakan *tool rapidminer* versi 5.3. Penelitian dilakukan dengan kombinasi metode atau melakukan penentuan jumlah fitur pada proses seleksi fitur. Harapan dari proses seleksi fitur ini adalah semakin sedikit jumlah fitur yang digunakan, semakin rendah waktu komputasi dan semakin tinggi akurasi yang dicapai.

Tahapan-tahapan dalam penelitian ini dapat diuraikan sebagai berikut:

a. Persiapan *Dataset*

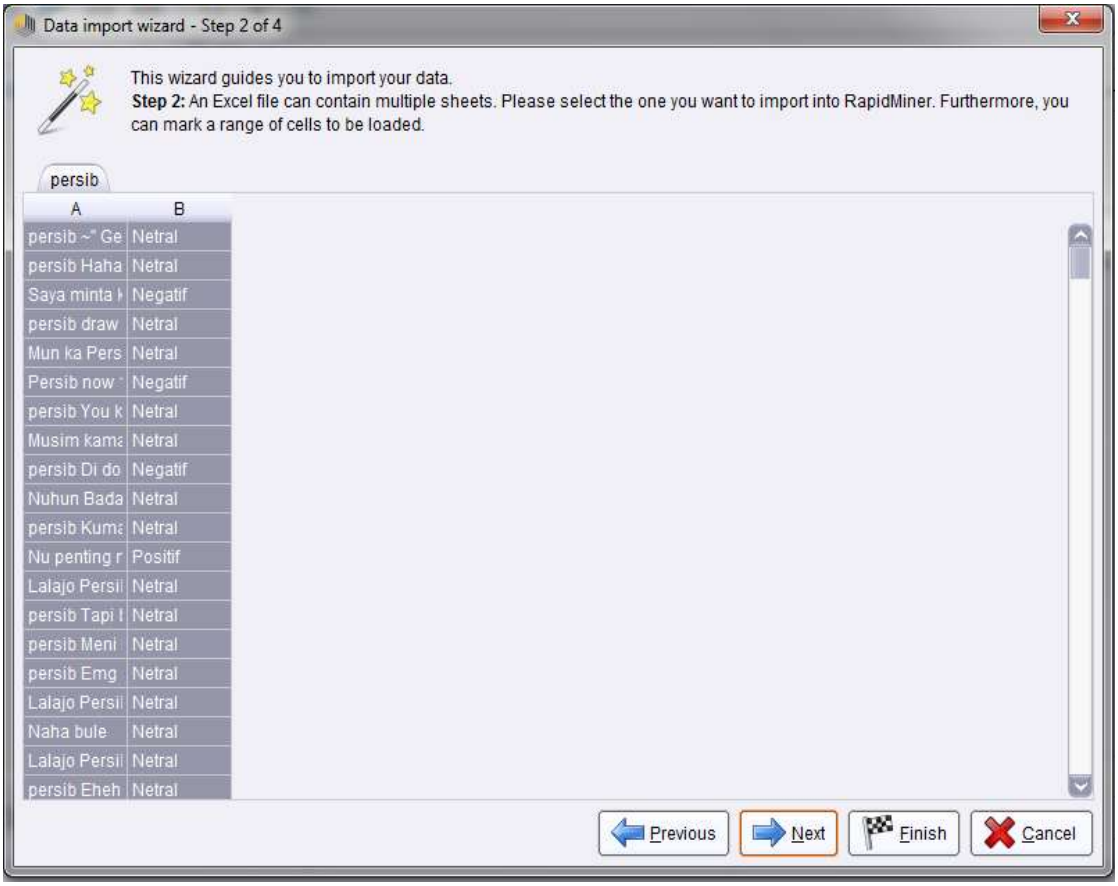
Tahap persiapan *Dataset* ini data yang digunakan sebanyak 316 *tweets*, yang disimpan dalam *file excel*, di mana pada setiap *record* berisi sebuah kalimat yang mengandung bahasa sunda, namun dari data *tweets* yang diperoleh terdapat campuran bahasa sunda dan Indonesia. Berikut contoh *tweets* untuk *dataset* dalam penelitian ini ditunjukkan pada Tabel 2.

b. Proses Klasifikasi Pada *RapidMiner*

Proses klasifikasi diawali dengan penentuan *dataset* yang disimpan di dalam *file excel*. Ditujukan pada Gambar 1 Penentuan *Dataset*.

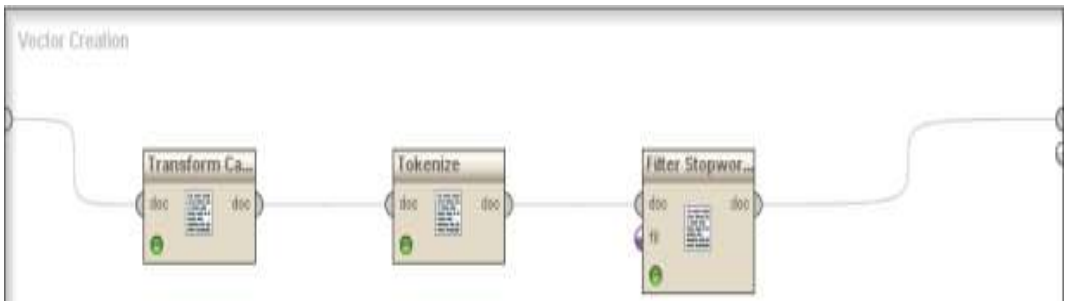
Tabel 2 *Tweets* Berbahasa Sunda

persib Sakitu kituna butut, lini tukang lapuk weuh harepan	Negatif
13 x main 3 x menang sisanya draw jeung eleh. 2019 kemunduran persib pisan	Negatif
Tetep masih dukung persib, Erek susah erek seneng tetep persib	Positif
persib Teu ngaruh oge	Netral
Jauh ti dulur Jauh ti lembur Aink tetep PERSIB	Positif
kamari abi hente lalajo persib	Netral
Da urang mah orang Sunda nya wajib we ngadukung Persib mah	Positif



Gambar 1 Penentuan *Dataset*

Untuk *text preprocessing* ditunjukkan pada Gambar 2 *Text Preprocessing* yaitu dengan tahapan menggunakan *transform case*, *tokenize*, dan *filter stopwords (dictionary)*.



Gambar 2 *Text Preprocessing*

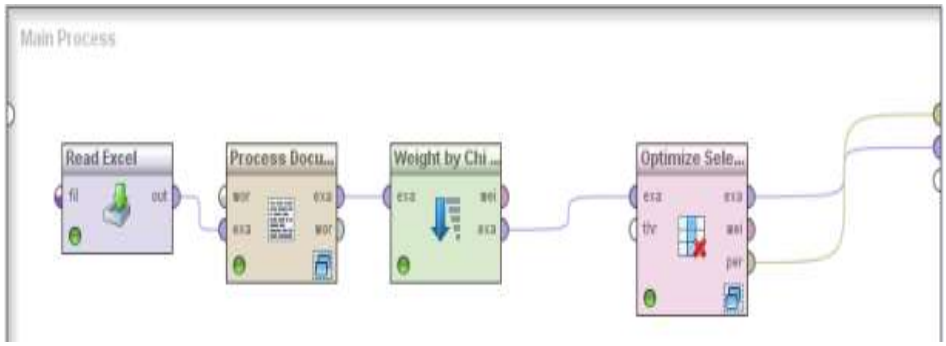
Tabel 3 Kata-kata S*topwords* Bahasa Indonesia dan Bahasa Sunda

ada	berapa	anu	ayeuna
adalah	berapakah	anyar	badag
adanya	berapalah	apa	bade
adapun	berapapun	aranjeun	bagean
agak	berarti	arek	baheula
agaknya	berawal	asa	bakal
agar	berbagai	atanapi	baruk

Tahap *transform case* dilakukan untuk merubah huruf dalam dokumen menjadi huruf kecil dan menghapus karakter simbol. Setelah melalui tahap *transform case* selanjutnya untuk tahap *tokenize* yaitu memotong setiap kata dalam teks. Selanjutnya untuk tahap *filter stopwords (dictionary)* adalah akan menghilangkan atau menghapus kata-kata yang tidak memiliki arti, di mana kata-kata tersebut terdapat pada daftar *stopwords* bahasa Indonesia dan bahasa sunda terdiri dari 999 kata-kata. Contoh kata-kata *stopwords* bahasa Indonesia dan bahasa sunda ditunjukkan pada Tabel 3.

c. Seleksi fitur menggunakan *Chi Square Statistic*

Pada tahap berikutnya seleksi fitur *Chi Square Statistic* dilakukan *Optimize Selection* dengan menggunakan *forward selection*. Tahap penyeleksian fitur ini menggunakan *Chi Square Statistic* di mana proses ini adalah untuk memilih kata atau *term* apa saja yang dapat dijadikan sebagai wakil penting untuk kumpulan dokumen yang akan dianalisis. Untuk proses klasifikasi dengan seleksi fitur *Chi Square Statistic* ditunjukkan pada Gambar 3.

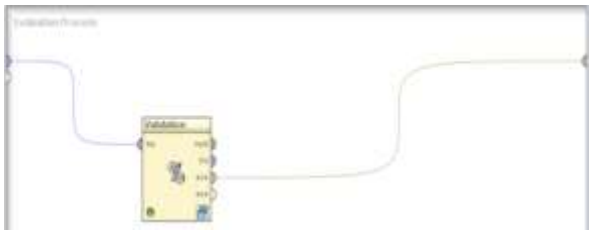


Gambar 3 Proses klasifikasi dengan seleksi fitur *Chi Square Statistic*

Read Excel dipakai untuk membaca *dataset* yang disimpan di dalam *file excel*. *Process Documents* merupakan tahap *Text Preprocessing* yaitu terdiri dari tahapan *transform case, tokenize*, dan *filter stopwords (dictionary).Weight by Chi* merupakan tahap penyeleksian fitur menggunakan *Chi Square Statistic.* Dan *Optimize Selection* dengan memilih parameter *forward selection. Forward selection* dimulai dengan tidak adanya fitur dan akan memulai menambahkan satu persatu fitur, sampai tidak ada lagi fitur yang mungkin dapat menurunkan *error* secara signifikan.

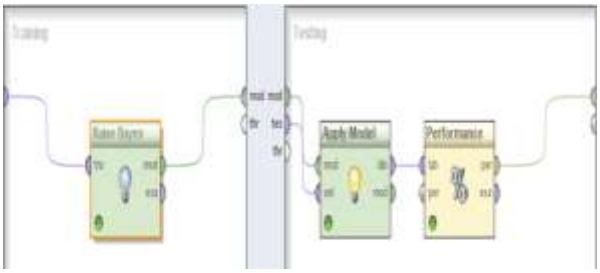
Cross-validation disini membagi data ke dalam dua *subset,* satu untuk data *training* dan satu lagi untuk data *testing.* Lalu akan dilakukan pertukaran fungsi dari dua *subset* tersebut, sehingga *subset* yang sebelumnya sebagai data *training* akan menjadi data *testing* demikian sebaliknya. *10-fold cross-validation* akan

melakukan mengulang pengujian sebanyak 10 kali dan hasil pengukuran adalah nilai rata-rata dari 10 kali pengujian. Proses *10-fold validation* dilakukan untuk meningkatkan akurasi klasifikasi. Proses ini diletakan didalam proses *optimize selection*. Didalam proses *10-fold validation* terdapat proses klasifikasi yang dilakukan oleh algoritma *naïve bayes* seperti ditunjukkan pada Gambar 4 *10-fold validation*.



Gambar 4 *10-Fold Validation*

Proses klasifikasi di sini adalah untuk menentukan sebuah kalimat terklasifikasi ke dalam kategori positif, negatif dan netral, berdasarkan nilai perhitungan probabilitas *naïve bayes* yang lebih besar. Misal jika hasil dari probabilitas kalimat untuk kategori negatif lebih besar daripada kategori positif dan netral maka kalimat tersebut termasuk ke dalam kategori negatif. Begitu juga sebaliknya dengan kategori positif dan netral.



Gambar 5 Klasifikasi *Naïve Bayes*

4.2 *Hasil Pengujian Chi Square Statistic dan Naïve Bayes.*

Tabel 4 *Confussion Matrix Chi Square Statistic* dan *Naïve Bayes*

	true **Netral**	***true*** **Negatif**	***true*** **Positif**
pred. **Netral**	119	47	11
pred. **Negatif**	6	98	1
pred. **Positif**	3	0	31

Untuk dapat mengetahui tingkat akurasi dari penggunaan seleksi fitur *Chi Square Statistic* dilakukan pengujian dengan perhitungan pada tabel *Confussion Matrix*. Ditunjukkan pada tabel 4 *Confussion Matrix Chi Square Statistic* dan *Naïve Bayes*.

Pada Tabel 4 diperoleh untuk jumlah *True* Netral (TNet) adalah 119, *True* Negatif (TN) adalah 98, *True* Positif (TP) adalah 31, *False* Netral (FNet) adalah 9, *False* Negatif (FN) adalah 47 dan *False* Positive (FP) adalah 12. Berdasarkan perhitungan tingkat akurasi dari data yang diperoleh pada tabel 4, menunjukkan bahwa penggunaan seleksi fitur *Chi Square Statistic* dan algoritma klasifikasi *Naïve Bayes Classifier* mendapatkan akurasi sebesar 78.48%.

$$Accuracy = \frac{TP+TN+TNet}{TP+TN+FN+FP+FNet+TNet} \quad (4.1)$$

$$= \frac{31+98+119}{31+98+47+12+9+119}$$

$$= \frac{248}{316} \times 100\ \% = 78{,}48\ \%$$

5. Kesimpulan

Dapat disimpulkan berdasarkan hasil analisis yang dilakukan pada penelitian ini adalah sebagai berikut:

a. Metode seleksi fitur *Chi Square Statistic* dan algoritma klasifikasi *Naïve Bayes Classifier* dapat digunakan dalam proses menganalisis sentiment *tweets* berbahasa sunda.
b. Hasil pengujian yang dilakukan penggunaan seleksi fitur *Chi Square Statistic* dapat mengurangi fitur-fitur yang tidak relevan pada proses klasifikasi *Naïve Bayes Classifier* dengan akurasi sebesar 78.48 %.

6. Saran

Saran-saran setelah dilakukan penelitian ini adalah sebagai berikut:

a. Penelitian selanjutnya memperbaiki pengolahan dokumen, identifikasi dokumen serta mengembangkan tahap *preprocessing* pada *dataset* berbahasa sunda seperti *emoticon*, singkatan dan lain-lain, untuk meningkatkan proses klasifikasi dokumen.
b. Menggunakan metode pemilihan fitur lain, seperti *Term Frequency x Inverse Document Frequency, Information gain, Mutual Information, Genetic Algorithm* dan lain-lain, agar dapat dibandingkan hasilnya.
c. Pada penelitian berikutnya dapat dikembangkan dengan metode algoritma klasifikasi lainnya seperti *Neural Network, Support Vector Machine, K-Nearest Neighbor* (K-NN) dan lain-lain.

Daftar Pustaka

Berry, M.W. & Kogan, J. 2010. "*Text Mining Aplication and theory"*. WILEY : United Kingdom.

Chandani, V., & Wahono, R. S. (2015). "Komparasi Algoritma Klasifikasi *Machine Learning* Dan *Feature Selection* pada Analisis Sentimen Review Film". *Journal of Intelligent Systems,1*(1), 56-60.

Dehaff, M. 2010. "*Sentiment Analysis, Hard But Worth It!"*.

Feldman, R & Sanger, J. 2007. "*The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data"*. Cambridge University Press : New York.

Ginting, H. S., Lhaksmana, K. M., & Murdiansyah, D. T. (2018). "Klasifikasi Sentimen Terhadap Bakal Calon Gubernur Jawa Barat 2018 Di *Twitte*r Menggunakan *Naive Bayes". eProceedings of Engineering*, *5*(1).

Gorunescu, F. 2011. "*Data Mining Concepts, Model and Techniques"*. Berlin: Springer.

Jenkins, M. C. 2011. "*How Sentiment Analysis works in machines"*.

Lidya, S. K., Sitompul, O. S., & Efendi, S. (2015). "Sentiment *Analysis* Pada Teks Bahasa Indonesia Menggunakan *Support Vector Machine* (SVM) Dan *K-Nearest Neighbor* (K-NN). In*Seminar Nasional Teknologi Informasi dan Komunikasi"*.

Ling, J., Kencana, I. P. E. N., & Oka, T. B. (2014). "Analisis Sentimen Menggunakan Metode *Naïve Bayes Classifier* Dengan Seleksi Fitur *Chi Square". E-Jurnal Matematika*, *3*(3), 92-99.

Putranti, N. D., & Winarko, E. (2014). "Analisis sentimen *twitter* untuk teks berbahasa Indonesia dengan *maximum* e*ntropy* dan *support vector machine". IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *8*(1), 91-100.

Routray, P., Swain, C. K. & Mishra, S.P., 2013. "A Survey on Sentiment Analysis. *International Journal of Computer Applications",* Agustus, 70(10), pp. 1-8

Saputra, N., Adji, T. B., & Permanasari, A. E. (2015). "Analisis sentimen data presiden Jokowi dengan *preprocessing* normalisasi dan *stemming* menggunakan metode *naive bayes* dan SVM". *Jurnal Dinamika Informatika*, *5*(1).

Wulandini, F. & Nugroho, A. N. 2009. "*Text Classification Using Support Vector Machine for Webmining Based Spation Temporal Analysis of the Spread of Tropical Diseases"*. International Conference on Rural Information and Communication Technology 2009.

Yang, Y., & Pedersen, J. O. 1997. "*A comparative study on feature selection in text categorization"*. ICML, (hal. 412--420).