Klasifikasi *Text Mining Review* Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan *Algoritma Naive Bayes*

Elly Indrayuni
Program Studi Sistem Informasi Akuntansi Kampus Kota Bogor
Fakultas Teknologi Informasi, Universitas Bina Sarana Informatika
Jl. Merdeka No. 168 Bogor
Email : elly.eiy@bsi.ac.id

ABSTRAKSI

Saat ini produk kosmetik sudah menjadi kebutuhan utama kaum wanita yang merupakan target utama dari industri kosmetik. Banyak website yang menyediakan informasi tentang produk kosmetik dengan memberikan banyak informasi berupa gambar dan *review* pengguna. Membaca semua review yang ada pada sebuah website tentu sangat memakan waktu. Oleh karena itu, analisa sentimen merupakan salah satu solusi mengatasi masalah untuk mengelompokan opini atau *review* menjadi opini positif atau negatif secara otomatis. *Naive Bayes* memiliki kelebihan yaitu sederhana, cepat dan memiliki akurasi yang tinggi. N-gram dianggap dapat mengurangi selisih antara klasifikasi kelas positif dan negatif sehingga dapat meningkatkan rata-rata akurasi akhir suatu algoritma. Nilai akurasi yang dihasilkan akan menjadi tolak ukur untuk mencari model pengujian terbaik untuk kasus klasifikasi sentimen. Evaluasi dilakukan menggunakan 10 *fold cross validation*. Pengukuran akurasi diukur dengan *confusion matrix* dan kurva ROC. Hasil penelitian menunjukkan penerapan *generate n-gram* pada tahap *preprocessing* mempengaruhi nilai akurasi dan nilai AUC yang dihasilkan. Nilai akurasi terbaik yang dihasilkan pada penelitian ini yaitu 90.50% dengan nilai AUC sebesar 0.715 pada penerapan *generate n-gram* = 2.

Kata Kunci: ***review, naive bayes, n-gram***

ABSTRACT

Nowadays cosmetic products have become the main needs of women who are the main target of the cosmetics industry. Many websites provide information about cosmetic products by providing a lot of information in the form of images and user reviews. Reading all the reviews on a website is certainly very time consuming. Therefore, sentiment analysis is one solution to overcome the problem to classify opinions or reviews into positive or negative opinions automatically. Naive Bayes has advantages that are simple, fast and have high accuracy. N-gram is considered to reduce the difference between the classification of positive and negative classes so as to increase the average final accuracy of an algorithm. The resulting accuracy value will be a benchmark for finding the best testing model for sentiment classification cases. Evaluation is done using 10 fold cross validation. Accuracy measurements were measured by confusion matrix and ROC curve. The results showed that the application of generate n-gram at the preprocessing stage affected the value of accuracy and the AUC value produced. The best accuracy value produced in this study is 90.50% with the AUC value of 0.715 on the application of generate n-gram = 2.

Keyword: review, naive bayes, n-gram

1. PENDAHULUAN

Saat ini produk kosmetik sudah menjadi kebutuhan utama kaum wanita yang merupakan target utama dari industri kosmetik. Semakin banyak merk kosmetik yang bermunculan baik dari dalam negeri maupun luar negeri membuat konsumen memiliki banyak pilihan dan pertimbangan dalam menggunakan suatu produk kosmetik tertentu, seperti kecocokan formula dengan jenis kulit, harga produk, daya tahan kosmetik sampai dengan hasil *make up* yang ingin dihasilkan. Untuk menentukan suatu produk kosmetik itu bagus atau tidak biasanya

konsumen mencari informasi dengan membaca review tentang produk kosmetik tersebut.

Banyak website yang menyediakan informasi tentang produk kosmetik dengan memberikan banyak informasi berupa gambar dan *review* pengguna. Membaca semua review yang ada pada sebuah website tentu sangat memakan waktu, karena terlalu banyak opini yang ada dari berbagai sumber website yang berbeda. Oleh karena itu, analisa sentimen merupakan salah satu solusi mengatasi masalah untuk mengelompokan opini atau *review* menjadi opini positif atau negatif secara otomatis.

Penelitian tentang klasifikasi sentimen terhadap *review* film telah dilakukan oleh Dhande dan Patnaik (2014) dengan menggunakan algoritma *Naive Bayes, Neural Network*, dan *Naive Bayes Neural Classifier*. Dari hasil penelitian akhir yang diuji menggunakan ketiga algoritma tersebut menyebutkan bahwa *Naive Bayes* menghasilkan akurasi yang lebih tinggi dibandingkan *Neural Network*.

Naive Bayes merupakan klasifikasi paling sederhana dan paling umum digunakan. *Naive Bayes* menghitung probabilitas kelas berdasarkan distribusi kata-kata yang ada dalam dokumen (Medhat, Hassan, & Korashy, 2014). *Naive Bayes* memiliki beberapa keunggulan seperti sederhana, cepat dan akurasi yang tinggi. Banyak peneliti telah melakukan klasifikasi sentimen dengan menggunakan *Naive Bayes.* Namun klasifikasi ini memiliki keterbatasan utama yang tidak mungkin selalu memenuhi asumsi independensi antara atribut. Dan ini mempengaruhi tingkat akurasi klasifikasi (Dhande & Patnaik, 2014).

Penelitian lain yang pernah dilakukan Kang, Yoo, dan Han (2012) adalah analisa sentimen pada *review* restoran menggunakan algoritma *Naive Bayes* dengan fitur unigrams dan bigrams untuk meningkatkan akurasi *Naive Bayes*. Ketika mengklasifikasikan *review* dokumen dengan algoritma *Naive Bayes* terdapat perbedaan hingga 10% antara akurasi klasifikasi positif dengan negatif sehingga mengakibatkan penurunan rata-rata akurasi akhir. Pada penelitian ini dengan menerapkan metode senti leksikon yaitu fitur unigrams dan bigrams, menunjukan bahwa selisih akurasi antara *class* positif dan negatif sekitar 3,6% dibandingkan dengan penggunaan *Naive Bayes* saja. Penerapan *fitur n-gram* dianggap dapat mengurangi selisih antara klasifikasi kelas positif dan negatif sehingga dapat meningkatkan rata-rata akurasi akhir suatu algoritma. Metodologi n-gram banyak digunakan untuk memprediksi kata berikutnya yang diberikan kata-kata sebelumnya dalam pemodelan bahasa statistik.

Pada penelitian ini penulis menggunakan algoritma *Naive Bayes* untuk mengklasifikasikan teks analisa sentimen teks berbahasa Indonesia dengan penerapan *generate n-gram* pada tahap preprocessing untuk mencari nilai akurasi terbaik.

Hipotesis dari penelitian ini adalah:

1. *Naive Bayes* mampu menghasilkan akurasi tinggi untuk analisa sentimen sebagai solusi untuk mengklasifikasikan review pengguna kedalam kategori opini positif atau negatif.
2. Penambahan fitur *generate n-gram* yang akan diterapkan pada tahap *preprocessing* mampu meningkatkan nilai akurasi untuk permasalahan klasifikasi sentimen *review* produk kosmetik.

Tujuan dari penelitian ini adalah untuk mencari model algoritma terbaik berdasarkan tingkat akurasi tertinggi yang dihasilkan dalam mengklasifikasikan analisa sentimen *review* produk kosmetik menggunakan algoritma *Naive Bayes.*

2. TINJAUAN PUSTAKA

2.1. Analisa Sentimen

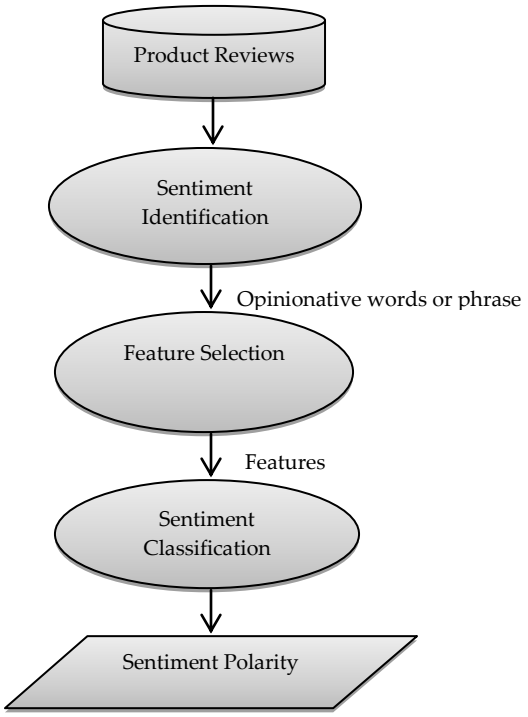
Opinion mining atau juga dikenal sebagai analisa sentimen adalah proses yang bertujuan untuk menentukan apakah polaritas kumpulan teks tulisan (dokumen, kalimat, paragraph, dll) cenderung ke arah positif, negatif, atau netral (Kontopoulos, Berberidis, Dergiades, & Bassiliades, 2013).

Menurut (Medhat et al., 2014) menyatakan bahwa "analisa sentimen adalah teknik komputasi pendapat, perasaan dan subjektivitas teks". Menurut Maynard dan Funk dalam (Medhat et al., 2014), teknik klasifikasi sentimen dapat terbagi atas:

1. *Machine Learning Approach*
   Machine learning approach menerapkan algoritma *machine learning* terkenal dengan menggunakan fitur linguistik.

2. *Lexicon Based Approach*
   Lexicon based approach bergantung pada leksikon sentimen, sebuah kumpulan istilah sentimen yang dikenal dan dikompilasi. Hal ini dibagi menjadi dua, yaitu *dictionary-based approach* dan *corpus-based approach* yang menggunakan metode statistik atau semantik untuk mencari sentimen polaritas.
3. *Hybrid Approach*
   Pada umumnya pendekatan hybrid menggabungkan kedua pendekatan leksikon sentimen sebagai peran kunci dalam sebagian besar metode.

Proses analisa sentimen untuk *review* produk sebagai berikut:



Sumber : (Medhat et al., 2014)

Gambar 1. Proses Analisa Sentimen pada Review Produk

2.2. *Pre-processing*

Pre-processing data adalah proses pembersihan dan mempersiapkan teks untuk klasifikasi (Haddi, Liu, & Shi, 2013). Teks online mengandung biasanya banyak *noise* dan bagian tidak informatif seperti tag HTML, script dan iklan. Selain itu, pada tingkat kata-kata, banyak kata-kata dalam teks tidak sesuai pada orientasi umum itu. Menjaga dimensi kata menjadi masalah besar sehingga klasifikasi lebih sulit karena setiap kata dalam teks diperlakukan sebagai satu dimensi. Berikut adalah hipotesis agar memiliki data yang benar sebelum diproses: mengurangi *noise* dalam teks dapat membantu meningkatkan kinerja classifier dan mempercepat proses klasifikasi sehingga membantu dalam analisis sentimen real time. Seluruh proses melibatkan beberapa langkah: membersihkan teks online, penghapusan ruang *spasi*, memperluas singkatan, kata dasar (*stemming)*, penghapusan kata henti *(stopword removal)*, penanganan negasi dan terakhir seleksi fitur.

2.3. N-gram

Menurut Huang et al., dalam (Gencosman, Ozmutlu, & Ozmutlu, 2014), metode n-gram adalah metode yang paling sederhana dan paling sukses dalam pemodelan bahasa. Pemodelan bahasa n -gram dapat digunakan untuk pidato atau pengenalan karakter, koreksi ejaan, pengenalan tulisan tangan, dan terjemahan mesin statistik. Meskipun metode n-gram bekerja dengan urutan kata dalam sejumlah besar teks, kesalahan ejaan tidak dapat dideteksi tanpa menimbang kata karakteristik ulasan. Kesalahan ejaan dalam query berurutan dapat dideteksi dengan karakter n-gram. Oleh karena itu, menggunakan karakter n-gram untuk memprediksi kelanjutan topik dalam *search engine* lebih logis daripada menggunakan kata n-gram *(n-gram term)*.

2.4. *Naive Bayes Classifier*

Menurut Tseng et al. dalam (Dhande & Patnaik, 2014), *Naive Bayes Classifier* adalah model sederhana untuk klasifikasi. Model ini bekerja dengan baik untuk klasifikasi teks. Model ini merupakan bentuk sederhana dari *Bayesian Network*, dimana semua atribut independen diberi nilai kelas variabel. *Naive Bayes Classifier* memiliki beberapa keunggulan seperti sederhana, cepat dan akurasi yang tinggi.

Menurut (Dhande & Patnaik, 2014) untuk memperkirakan prior probabilitas dapat ditunjukkan dengan persamaan:

$$\gamma(\alpha) = \frac{Nc}{N}$$

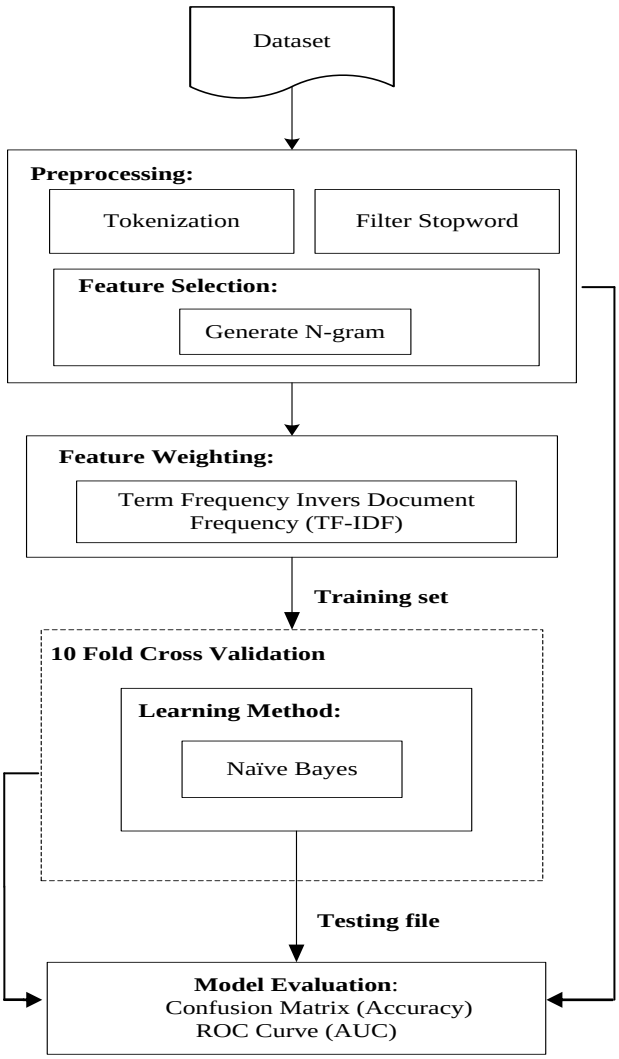
Keterangan:

Nc : jumlah dokumen kelas α
N : total jumlah dokumen keseluruhan

3. METODOLOGI

Metode penelitian yang penulis lakukan adalah metode penelitian eksperimen, dengan tahapan sebagai berikut:

1. Pengumpulan Data
   Penulis menggunakan data review produk kosmetik berbahasa Indonesia yang didapat dari situs https://femaledaily.com/ yang terdiri dari 100 review positif dan 100 review negatif. Pengumpulan data ini dilakukan dengan cara memfilter secara manual untuk data *review* yang berisi opini positif dan opini negatif.
2. Pengolahan Awal Data
   Pada tahap pengolahan awal data untuk klasifikasi teks atau sentiment digunakan tahap *preprocessing* agar teks yang *noise* atau bersifat tag HTML, symbol ataupun tanda baca dapat dihilangkan. Ada beberapa tahap *preprocessing* yang digunakan penulis dalam penelitian ini, antara lain:
   a. *Tokenization*
      Proses mengumpulkan semua kata yang muncul dan menghilangkan semua tanda baca, simbol, atau apapun yang bukan huruf sehingga menjadi sekumpulan kata secara utuh.
   b. *Filter Stopword*
      Proses penghapusan kata-kata yang tidak relevan sehingga dihasilkan sekumpulan teks yang memiliki arti dan berkaitan dengan klasifikasi sentimen.
   c. *Generate N-gram*
      Pada proses ini digunakan generate N-gram yaitu Character N-gram sebesar 2, 3, dan 4. Dari hasil proses penerapan generate N-gram ini akan dicari nilai akurasi tertinggi.
3. Model Yang Diusulkan
   Metode yang diusulkan pada penelitian ini adalah metode *Naive Bayes* dengan penambahan fitur *generate n-gram* pada tahap *preprocessing*.
4. Eksperimen dan Pengujian Metode
   Software yang digunakan sebagai alat bantu untuk menghitung tingkat akurasi dalam eksperimen dan pengujian metode dalam penelitian ini adalah Rapidminer.
5. Evaluasi dan Validasi Hasil
   Setelah didapatkan hasil akurasi dari beberapa model pengujian yang telah dilakukan, diperlukan proses evaluasi dan validasi untuk mendapatkan nilai akurasi terbaik. Teknik validasi yang digunakan menggunakan *cross validation*, sedangkan *confusion matrix* digunakan untuk menghitung akurasi dan nilai *training cycle*. Sedangkan untuk melihat kualitas hasil olahan data dapat diukur melalui nilai *Area Under Curve* (AUC) pada *ROC Curve.*



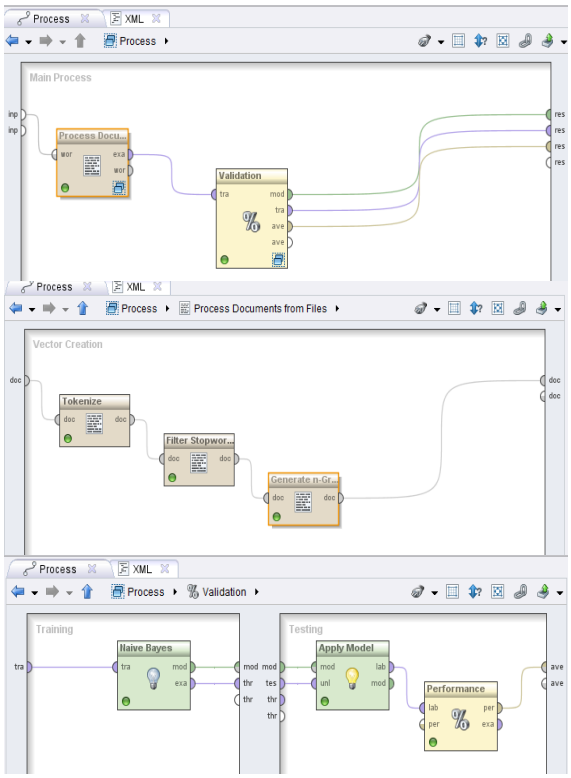
Sumber : (Indrayuni, 2019)

Gambar 2. Model Yang Diusulkan

4. HASIL DAN PEMBAHASAN

4.1. Model Eksperimen Algoritma Naive Bayes

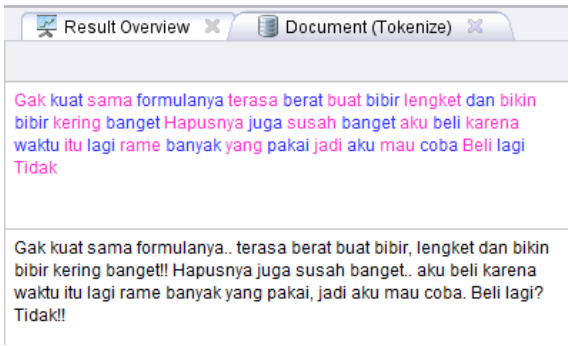
Pengklasifikasian teks menggunakan *Naive Bayes* melalui proses yang cukup sederhana. Pada penelitian ini ditambahkan fitur *generate N-gram* pada tahap *preprocessing*.



Sumber : (Indrayuni, 2019)

Gambar 3. Desain Model Algoritma Naive Bayes

Pada gambar 3 dapat dilihat penerapan proses *preprocessing* yang meliputi tokenize, filter stopword dan generate N-gram. Hasil proses tahap tokenize dengan menggunakan Rapidminer dapat dilihat pada gambar 4.

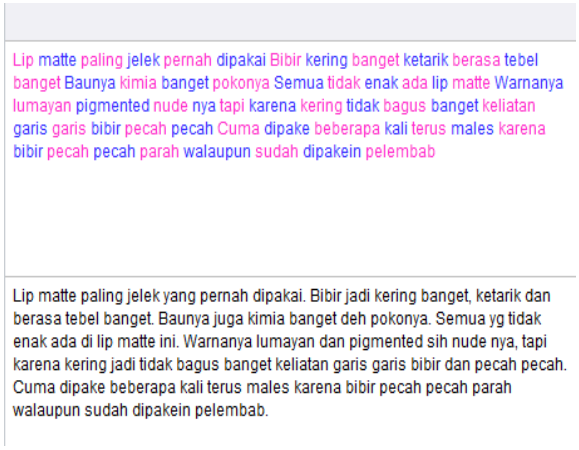


Sumber : (Indrayuni, 2019)

Gambar 4. Proses Tokenize pada Rapidminer

Pada proses *tokenize* terlihat bahwa semua semua tanda baca, simbol, atau apapun yang bukan huruf hilang sehingga teks review menjadi sekumpulan kata secara utuh.

Untuk proses *filter stopword* karena review yang diambil menggunakan teks bahasa Indonesia, maka yang digunakan pada Rapidminer dalah *filter stopword (Dictionary)*. Pada proses ini diperlukan file seperti kamus yang berisi kata-kata yang tidak relevan dalam bahasa Indonesia seperti "ini", "dan", "yang", "jadi", yang nantinya akan dihilangkan sehingga teks yang tersisa hanya teks yang memiliki arti dan berkaitan dengan klasifikasi sentimen.



Sumber : (Indrayuni, 2019)

Gambar 5. Proses Filter Stopword pada Rapidminer

Pada klasifikasi sentimen ini digunakan beberapa kata yang menjadi parameter sebagai penentuan data review produk kosmetik tersebut termasuk kategori opini positif atau opini negatif. Untuk opini positif antara lain seperti "bagus", "ringan", "awet". Sedangkan parameter yang mewakili opini negatif adalah "jelek", "berat", "lengket". Berdasarkan parameter tersebut maka setiap review dapat diklasifikasikan ke dalam opini positif atau opini negatif dengan menggunakan Rapidminer.

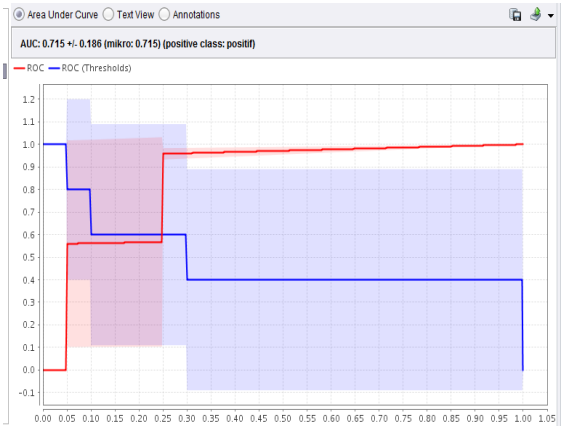
Hasil eksperimen pengklasifikasian teks opini dengan menggunakan algoritma *Naive Bayes* dengan penambahan fitur *generate N-gram* dan pengujian 10 *cross validation* dilihat pada tabel 1.

Tabel 1. Eksperimen Penentuan Nilai Training Cycles Naive Bayes

Cross Validation	Generate N-gram	Accuracy (%)	AUC
5	2	90.50	0.715
	3	90.50	0.640
	4	90.50	0.500
6	2	85.01	0.596
	3	86.51	0.534
	4	86.51	0.641
7	2	87.98	0.566
	3	88.95	05.00
	4	89.46	0.547
8	2	87.50	0.553
	3	87.00	0.553
	4	88.00	0.604
9	2	88.03	0.500
	3	88.49	0.591
	4	89.50	0.595
10	2	89.50	0.500
	3	90.00	0.544
	4	90.00	0.625

Sumber : (Indrayuni, 2019)

Berdasarkan hasil eksperimen yang telah dilakukan, nilai akurasi terbaik yang dihasilkan terdapat pada penerapan *N-gram* sebesar 2 dan *5-cross validation* yaitu 90.50% dengan nilai AUC sebesar 0.715. Nilai AUC tersebut termasuk *Fair Classification*. Berikut tampilan kurva ROC:



Sumber : (Indrayuni, 2019)

Gambar 6. Kurva ROC Algoritma Naive Bayes pada Rapidminer

4.2. Hasil Pengujian Model Algoritma Naive Bayes

Berdasarkan hasil pengolahan 200 data *training* menggunakan Naive Bayes menunjukkan bahwa tingkat akurasi yang dihasilkan sebesar 90.50%. Untuk mencari nilai *accuracy, sensitivity, specificity, ppv* dan *npv* dapat digunakan nilai *true positif, false negatif, true negatif,* dan *false positif.* Untuk mengetahui nilai-nilai tersebut dapat diketahui melalui tabel *confusion matrix*

Tabel 2. Model Confusion Matrix untuk Algoritma Naive Bayes

Accuracy : 90.50%			
	True positif	True negative	Class Precision
Prediksi Positif	94	13	87.85%
Prediksi Negative	6	87	93.55%
Class Recall	94.00%	87.00%	

Sumber : (Indrayuni, 2019)

Berdasarkan tabel *confusion matrix* menunjukkan bahwa jumlah *true positive* (tp) adalah 94 opini, *false negative* (fn) sebanyak 6 opini. Berikutnya 87 opini untuk *true negative* (tn) dan 13 opini untuk *false positif* (fp). Nilai *accuracy, sensitivity, specificity, ppv* dan *npv* hasilnya dapat dilihat pada Tabel 3.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$= \frac{94 + 87}{94 + 87 + 13 + 6}$$

$$= 0.905$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$= \frac{94}{94 + 6}$$

$$= 0.94$$

$$Specificity = \frac{TN}{TN + FP}$$

$$= \frac{87}{87 + 13}$$

$$= 0.87$$

$$ppv = \frac{TP}{TP + FP}$$

$$= \frac{94}{94 + 13}$$

$$= 0.8785$$

$$npv = \frac{TN}{TN + FN}$$

$$= \frac{87}{87 + 6}$$

$$= 0.9355$$

Tabel 3. ***Nilai Accuracy, Sensitivity, Specificity, Ppv Dan Npv Algoritma Naive Bayes***

	% (dalam persen)
Accuracy	**90.50**
Sensitivity	**94.00**
Specificity	**87.00**
ppv	**87.85**
npv	**93.55**

Sumber : (Indrayuni, 2019)

4.3. Analisa Evaluasi dan Validasi Model

Setelah pengklasifikasian *review* produk kosmetik berhasil dikelompokkan menjadi kategori opini positif dan opini negatif sehingga nilai akurasinya pun telah muncul, maka tingkat akurasi dapat diuji untuk melihat kinerja dari hasil pengujian model algoritma Naive Bayes. Berdasarkan evaluasi menggunakan *confusion matrix* maupun ROC *curve* terbukti bahwa penerapan generate N-gram pada proses *preprocessing* dapat mempengaruhi nilai akurasi dan nilai AUC yang dihasilkan pada pengolahan data training menggunakan algoritma *Naive Bayes*. Untuk klasifikasi text mining review produk kosmetik dalam bahasa Indonesia ini penerapan *generate* N-*gram* dan pengujian data training menggunakan *cross validation = 5* menghasilkan nilai akurasi yang sama yaitu 90.50% namun nilai AUC yang dihasilkan berbeda. Pada penerapan generate N-gram = 2 dan , nilai akurasi yang dihasilkan 90.50 % dengan nilai AUC sebesar =0.715. Sedangkan pada penerapan generate N-gram = 3, nilai akurasi yang dihasilkan sama namun nilai AUC menurun 0.075 menjadi 0.640. Dan terakhir untuk penerapan generate N-gram = 4, nilai AUC yang dihasilkan menjadi 0.500, selisih 0.140 dengan penerapan generate N-gram = 3. Berdasarkan hasil eksperimen, maka hal ini membuktikan bahwa algoritma Naive Bayes merupakan algoritma sederhana yang dapat menghasilkan nilai akurasi yang tinggi. Namun dengan adanya fitur *generate N-gram* pada tahap *preprocessing* dan pengujian data training menggunakan *cross validation* mempengaruhi nilai akurasi dan nilai AUC yang menggambarkan kualitas pengolahan data.

5. KESIMPULAN DAN SARAN

Berdasarkan pengujian model menggunakan algoritma *Naive Bayes* pada eksperimen yang telah dilakukan terbukti bahwa algoritma *Naive Bayes* merupakan algoritma paling sederhana yang terbukti menghasilkan nilai akurasi tinggi hingga 90.50% dengan nilai AUC sebesar 0.715. Penerapan *generate N-gram* pada tahap *preprocessing* dapat meningkatkan nilai rata-rata akurasi sehingga secara keseluruhan diperoleh kesimpulan bahwa penerapan *generate N-gram* pada algoritma *Naive Bayes* merupakan salah satu model pengujian algoritma yang terbaik dan akurat untuk permasalahan klasifikasi sentimen *review* produk kosmetik untuk teks berbahasa Indonesia.

Walaupun pada penelitian ini telah dihasilkan model algoritma yang terbaik dengan nilai akurasi yang tinggi, namun ada beberapa hal yang dapat dilakukan untuk penelitian selanjutnya. Adapun sara-saran yang diusulkan antara lain perlunya penggunaan optimasi untuk selekasi fitur seperti *Genetic Algorithm* dan *Chi Square* untuk meningkatkan hasil akurasi pada sebuah algoritma.

REFERENSI

Dhande, L. L., & Patnaik, P. G. K. (2014). Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, *3*(4), 313–320.

Gencosman, B. C., Ozmutlu, H. C., & Ozmutlu, S. (2014). *Character n-gram application for automatic new topic identification*. *Information Processing and Management* (Vol. 50). Elsevier Ltd. https://doi.org/10.1016/j.ipm.2014.06.005

Haddi, E., Liu, X., & Shi, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis. *First International Conference on Information Technology and Quantitative Management*, *17*, 26–32. https://doi.org/10.1016/j.procs.2013.05.005

Indrayuni, E. (2019). *Laporan Akhir Penelitian Mandiri 2019*.

Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Expert Systems with Applications Ontology-based sentiment analysis of twitter posts. *Expert Systems With Applications*, *40*(10), 4065–4074. https://doi.org/10.1016/j.eswa.2013.01.001

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*. https://doi.org/10.1016/j.asej.2014.04.011

.