KOMPARASI ALGORITMA NAIVE BAYES DAN SUPPORT VECTOR MACHINE UNTUK ANALISA SENTIMEN REVIEW FILM

Elly Indrayuni

Manajemen Informatika
AMIK BSI Pontianak
http://www.bsi.ac.id
elly.eiy@bsi.ac.id

Ciptaan disebarluaskan di bawah Lisensi Creative Commons Atribusi-NonKomersial 4.0 Internasional.

Abstract— *Film is a subject of interest by a large number of people among the social networking community who have significant differences in their opinions or sentiments. Sentiment analysis or opinion mining is one solution to overcome the problem to classify opinions or reviews into positive or negative opinions automatically. The technique used in this study is Naive Bayes and Support Vector Machines (SVM). Naive Bayes has advantages that are simple, fast and have high accuracy. Whereas SVM is able to identify a separate hyperplane that maximizes the margin between two different classes. The results of the sentiment classification in this study consisted of two class labels, namely positive and negative. The value of accuracy produced 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 accuracy value for the Naive Bayes algorithm was 84.50%. While the accuracy value of the Support Vector Machine (SVM) algorithm is greater than Naive Bayes which is equal to 90.00%.*

Intisari— Film merupakan subjek yang diminati oleh sejumlah besar orang diantara komunitas jaringan sosial yang memiliki perbedaan signifikan dalam pendapat atau sentimen mereka. Analisa sentimen atau *opinion mining* merupakan salah satu solusi mengatasi masalah untuk mengelompokan opini atau *review* menjadi opini positif atau negatif secara otomatis. Teknik yang digunakan dalam penelitian ini adalah *Naive Bayes* dan *Support Vector Machines* (SVM). *Naive Bayes* memiliki kelebihan yaitu sederhana, cepat dan memiliki akurasi yang tinggi. Sedangkan SVM mampu mengidentifikasi *hyperplane* terpisah yang memaksimalkan margin antara dua kelas yang berbeda. Hasil klasifikasi sentimen pada penelitian ini terdiri dari dua label *class*, yaitu positif dan negatif. 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 nilai akurasi untuk algoritma *Naive* Bayes sebesar 84.50%. Sedangkan nilai akurasi algoritma *Support Vector Machine* (SVM) lebih besar dari *Naive Bayes* yaitu sebesar 90.00%.

Kata Kunci*: Analisa Sentimen, Review, Naive Bayes, SVM*

PENDAHULUAN

Dalam beberapa tahun terakhir ini, perkembangan *smartphone* dan aplikasi yang sangat pesat memungkinkan para penggunanya untuk mengomentari berbagai *platform* melalui layanan internet *mobile,* media sosial, dan lain-lain (Zhang, Hua, Wang, Qian, & Zheng, 2014). Sistem blogging mikro (seperti Twitter) digunakan oleh setiap individu yang berbeda untuk menunjukkan sentimen (opini) mereka tentang berbagai subjek, sehingga Twitter merupakan sumber informasi yang berguna dari sentimen individu (Samad, Basari, Hussin, Pramudya, & Zeniarja, 2013).

Film merupakan subjek yang diminati oleh sejumlah besar orang diantara komunitas jaringan sosial yang memiliki perbedaan signifikan dalam pendapat atau sentimen mereka. Opini penambangan ulasan film diukur lebih menantang daripada penambangan opini dari kategori ulasan lainnya, seperti ulasan produk (Samad et al., 2013). Analisa sentimen atau *opinion mining* adalah studi komputasi mengenai pendapat, perilaku dan emosi seseorang terhadap entitas. Entitas tersebut dapat menggambarkan individu, kejadian atau topik (Medhat, Hassan, & Korashy, 2014). Oleh karena itu, analisa sentimen atau

opinion mining merupakan salah satu solusi mengatasi masalah untuk mengelompokan opini atau *review* menjadi opini positif atau negatif secara otomatis.

Naive Bayes merupakan klasifikasi paling sederhana dan paling umum digunakan. *Naive Bayes* menghitung probabilitas kelas berdasarkan distribusi kata-kata yang ada dalam dokumen (Medhat et al., 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 tentang klasifikasi sentimen terhadap *review* film dengan menggunakan algoritma *Naive Bayes, Neural Network*, dan *Naive Bayes Neural Classifier* (Dhande & Patnaik, 2014). Dari hasil penelitian akhir yang diuji menggunakan ketiga algoritma tersebut menyebutkan bahwa *Naive Bayes* menghasilkan akurasi yang lebih tinggi dibandingkan *Neural Network*. Dan algoritma *Naive Bayes Neural Classifier* yang merupakan penggabungan antara metode *Naive Bayes* dan *Neural Network* menghasilkan akurasi yang paling tinggi diantara kedua algoritma tersebut.

Support Vector Machines (SVM) telah menjadi metode klasifikasi dan regresi yang populer untuk masalah linear dan nonlinear. Keistimewaan dari *Support Vector Machines* berasal dari kemampuan untuk menerapkan pemisahan linear pada input data non linear berdimensi tinggi, dan ini diperoleh dengan menggunakan fungsi kernel yang diperlukan. Efektivitas *Support Vector Machines* sangat dipengaruhi oleh jenis fungsi kernel yang dipilih dan diterapkan berdasarkan karakteristik data (Haddi, Liu, & Shi, 2013). Banyak peneliti telah melaporkan bahwa *Support Vector Machines* metode yang paling akurat untuk teks klasifikasi (Moraes, Valiati, & Gavião Neto, 2013).

Penelitian klasifikasi sentimen yang telah dilakukan adalah komparasi algoritma *Support Vector Machines* dan *Artificial Neural Networks* untuk klasifikasi sentimen level dokumen (Moraes et al., 2013). Hasil penelitian ini menunjukan bahwa *Artificial Neural Networks* memperoleh hasil yang lebih unggul atau setidaknya sebanding dengan *Support Vector Machines.* Penelitian ini juga memberitahukan beberapa keterbatasan dari kedua model yang jarang dibahas dalam sentimen klasifikasi teks.

Pada penelitian ini penulis menggunakan algoritma *Naive Bayes* dan *Support Vector Machines (SVM)* untuk mengklasifikasikan teks analisa sentimen pada *review* film untuk mencari nilai akurasi terbaik dengan membandingkan hasil akurasi seluruh model yang diterapkan.

BAHAN DAN METODE

Metode yang dilakukan penulis pada penelitian ini adalah metode penelitian eksperimen, dengan tahapan sebagai berikut:

1. Pengumpulan data

   Data review film diambil dari situs www.cinemablend.com. Pengumpulan data ini dilakukan dengan cara melakukan *filter* untuk data *review* yang berisi opini positif dan opini negatif. Penulis menggunakan 200 data review film yang terdiri dari 100 review untuk opini positif dan 100 review untuk 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. Tahap *preprocessing* yang digunakan penulis untuk membantu menghasilkan nilai akurasi terbaik, antara lain:

   a. *Tokenization*

      Pada proses *tokenize* ini, semua tanda baca, simbol, atau apapun yang bukan huruf dihilangkan sehingga menjadi sekumpulan kata secara utuh.

   b. *Filter Stopword*

      Pada tahap ini terjadi penghapusan kata-kata yang tidak relevan, seperti *the, for, of* , dan sebagainya sehingga dihasilkan sekumpulan teks yang memiliki arti dan berkaitan dengan klasifikasi sentimen.
3. Metode Yang Diusulkan

   Metode yang diusulkan pada penelitian ini adalah penggunaan algoritma *Naive Bayes* dan *Support Vector Machine.*
4. Eksperimen dan Pengujian Metode

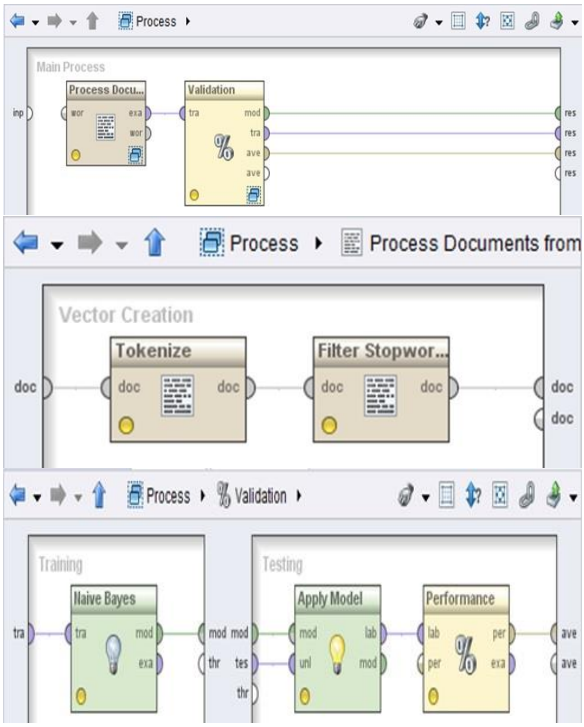
   Dalam melakukan pengujian metode yang diusulkan pada eksperimen ini, digunakan *software* sebagai alat bantu untuk menghitung tingkat akurasi yaitu Rapidminer.
5. Evaluasi dan Validasi Hasil

   Evaluasi dilakukan untuk mengetahui hasil akurasi dari eksperimen yang telah dilakukan. Setelah nilai akurasi didapatkan maka dilakukan proses validasi untuk mendapatkan nilai akurasi terbaik menggunakan *confusion matrix* dan *ROC Curve*. Kurva ROC akan digunakan untuk mengukur AUC (Area Under Curve). Kurva ROC membagi hasil positif dalam sumbu y dan hasil negative dalam sumbu x (Aulianita, 2016).

HASIL DAN PEMBAHASAN

A. Hasil eksperimen dan Pengujian Metode Algoritma *Naive Bayes*

Pengklasifikasian teks menggunakan *Naive Bayes* melalui proses yang cukup sederhana.



Sumber: (Indrayuni, 2018)

Gambar 1. Desain Model Algoritma *Naive Bayes*

Pada klasifikasi sentimen ini digunakan beberapa kata yang menjadi atribut sebagai penentuan data *review* film tersebut termasuk kategori opini positif atau opini negatif antara lain seperti *good, entertaining* dan *informative* untuk mewakili opini positif. Sedangkan atribut yang mewakili opini negatif adalah *bad, bored,* dan *disappointed.*

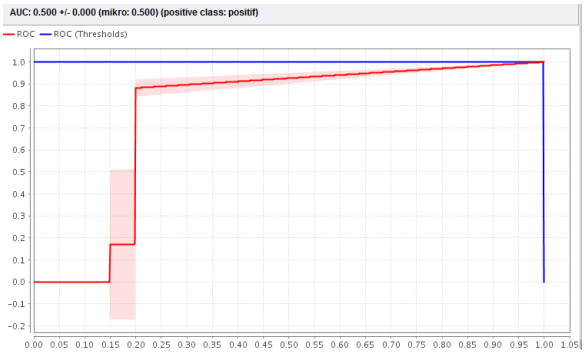
Hasil akurasi pengklasifikasian teks opini dengan menggunakan algoritma *Naive Bayes* dapat dilihat pada tabel berikut ini:

Tabel 1. Eksperimen Penentuan Nilai *Training Cycles Naive Bayes*

	Naive Bayes (NB)
Accuracy	84.50%
AUC	0.500

Sumber : (Indrayuni, 2018)

Berdasarkan hasil eksperimen yang telah dilakukan, akurasi yang dihasilkan sebesar 84.50% dengan nilai AUC sebesar 0.500. Nilai AUC tersebut termasuk *Poor Classification*. Berikut tampilan kurva ROC:



Sumber: (Indrayuni, 2018)

Gambar 2. Kurva ROC Algoritma *Naive Bayes*

Hasil pengolahan 200 data training menggunakan algoritma *Naive Bayes* pada tabel *confusion matrix* dapat dilihat pada tabel berikut ini.

Tabel 2. Model *Confusion Matrix* untuk *Algoritma Naive Bayes*

Accuracy : 84.50%			
	True positif	*True negative*	*Class Precission*
Prediksi positif	88	19	82.24%
Prediksi negative	12	81	87.10%
Class Recall	88.00%	81.00%	

Sumber: (Indrayuni, 2018)

Berdasarkan tabel *confusion matrix* menunjukkan bahwa jumlah *true positive* (tp) adalah 88 opini, *false negative* (fn) sebanyak 12 opini. Berikutnya 81 opini untuk *true negative* (tn) dan 19 opini untuk *false positif* (fp). Nilai *accuracy, sensitivity, specificity, ppv* dan *npv* hasilnya dapat dilihat pada Tabel 3.

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

$$= \frac{88 + 81}{88 + 81 + 19 + 12}$$

$$= 0.845$$

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

$$= \frac{88}{88 + 12}$$

$$= 0.88$$

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

$$= \frac{81}{81 + 19}$$

$$= 0.81$$

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

$$= \frac{88}{88 + 19}$$

$$= 0.8224$$

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

$$= \frac{81}{81 + 12}$$

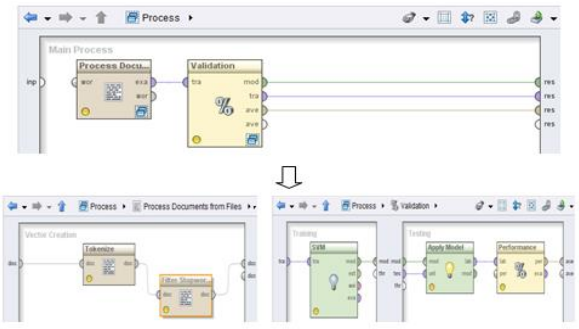
$$= 0.871$$

Tabel 3. Nilai *accuracy, sensitivity, specificity, ppv* dan *npv Algoritma Naive Bayes*

	% (dalam persen)
Accuracy	84.50
Sensitivity	88.00
Specificity	81.00
Ppv	82.24
Npv	87.10

Sumber : (Indrayuni, 2018)

B. Hasil Eksperimen dan Pengujian Metode Algoritma *Support Vector Machine*



Sumber : (Indrayuni, 2018)
Gambar 3. Desain Model Algoritma *Support Vector Machine*

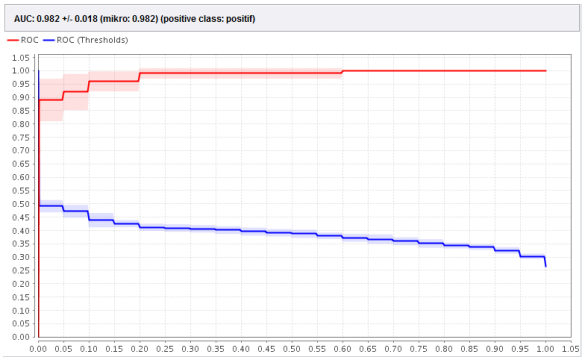
Nilai *training cycles* dalam penelitian ini ditentukan dengan cara melakukan uji coba dengan memasukkan nilai C dan epsilon pada parameter SVM. Hasil percobaan yang telah dilakukan untuk penentuan nilai *training cycles* dapat dilihat pada Tabel 4.

Tabel 4. Eksperimen Penentuan Nilai *Training Cycles* SVM

Parameter		SVM	
C	Epsilon	Accuracy	AUC
0.0	0.5	90.00%	0.980
0.5	0.5	90.00%	0.980
0.8	0.8	90.00%	0.980
1.0	0.0	90.00%	0.982
1.0	0.5	90.00%	0.980
1.0	0.8	90.00%	0.980

Sumber : (Indrayuni, 2018)

Berdasarkan eksperimen yang telah dilakukan, penentuan parameter untuk C dan epsilon dapat mempengaruhi nilai akurasi dan AUC. Hasil terbaik yang diperoleh yaitu nilai akurasi tertinggi mencapai 90.00% dengan nilai AUC sebesar 0.982 dengan penentuan nilai C= 1.0 dan epsilon= 0.0. Nilai AUC tersebut termasuk *Excellent Classification*. Berikut tampilan kurva ROC:



Sumber: (Indrayuni, 2018)
Gambar 4. Kurva ROC Algoritma *Support Vector Machine*

Berikut ini tabel *confusion matrix* hasil pengolahan data training menggunakan algoritma *Support Vector Machine.*

Tabel 5. Model *Confusion Matrix* untuk *Algoritma Support Vector Machine*

Accuracy : 90.00%			
	True positif	*True negative*	*Class Precission*
Prediksi positif	81	1	98.78%
Prediksi negative	19	99	83.90%
Class Recall	81.00%	99.00%	

Sumber : (Indrayuni, 2018)

Berdasarkan tabel *confusion matrix* hasil pengolahan data review film menggunakan algoritma *Support Vector Machine* menunjukkan bahwa jumlah *true positive* (tp) adalah 81 opini, *false negative* (fn) sebanyak 19 opini. Berikutnya 99 opini untuk *true negative* (tn) dan 1 opini untuk *false positif* (fp). Nilai *accuracy, sensitivity, specificity, ppv* dan *npv* hasilnya dapat dilihat pada Tabel 6.

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

$$= \frac{81 + 99}{81 + 99 + 1 + 19}$$

$$= 0.90$$

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

$$= \frac{81}{81 + 19}$$

$$= 0.81$$

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

$$= \frac{99}{99 + 1}$$

$$= 0.99$$

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

$$= \frac{81}{81 + 1}$$

$$= 0.9878$$

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

$$= \frac{99}{99 + 19}$$

$$= 0.839$$

Tabel 6. Nilai *accuracy, sensitivity, specificity, ppv* dan *npv Algoritma Naive Bayes*

	% (dalam persen)
Accuracy	90.00
Sensitivity	81.00
Specificity	99.00
Ppv	98.78
Npv	83.90

Sumber : (Indrayuni, 2018)

Setelah pengklasifikasian *review* film telah 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 diatas. Berdasarkan evaluasi menggunakan *confusion matrix* maupun ROC *curve* terbukti bahwa nilai akurasi yang dihasilkan oleh algoritma *Support Vector Machine* lebih besar daripada algoritma *Naive Bayes.* Untuk hasil pengujian semua algoritma secara detail dapat dilihat pada Tabel 7.

Tabel 7. Perbandingan algoritma *Naive Bayes* dan SVM

	Naive Bayes	SVM
Accuracy	84.50%	90.00%
AUC	0.500	0.982

Sumber : (Indrayuni, 2018)

KESIMPULAN

Berdasarkan hasil pengujian model menggunakan algoritma *Naive Bayes* dan *Support Vector Machines* pada eksperimen yang telah dilakukan ada beberapa hal yang dihasilkan, antara lain: algoritma *Naive Bayes* merupakan algoritma paling sederhana yang terbukti menghasilkan nilai akurasi hingga 84.50% dengan nilai AUC 0.500. Meskipun menghasilkan nilai akurasi yang tinggi namun *Naive Bayes* memiliki kekurangan karna nilai AUC-nya masih termasuk ke dalam kategori *Poor Classification*. Sedangkan untuk algoritma *Support Vector Machines* terbukti menghasilkan nilai akurasi tinggi dan lebih akurat dengan nilai akurasi hingga 90.00% dan nilai AUC *(Area Under Curve)* sebesar 0.982 yang termasuk *Excellent Classification.*

Dari uraian diatas, dapat disimpulkan bahwa *Support Vector Machines* memberikan unjuk kerja lebih baik daripada *Naive Bayes* untuk klasifikasi sentimen *review* film dengan menghasilkan nilai akurasi dan AUC yang tinggi. Algoritma *Support Vector Machines* merupakan model pengujian algoritma yang terbaik dan akurat untuk permasalahan klasifikasi sentimen *review* film.

REFERENSI

Aulianita, R. (2016). Komparasi Metode K-Nearest Neighbors dan Support Vector Machine Pada Sentiment Analysis Review Kamera, *8*(3), 71–77.

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.

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.2013.05.05.

Indrayuni, E. (2018). *Laporan Akhir Penelitian Mandiri 2018*.

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.

Moraes, R., Valiati, J. F., & Gavião Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, *40*(2), 621–633. https://doi.org/10.1016/j.eswa.2012.07.05.

Samad, A., Basari, H., Hussin, B., Pramudya, I. G., & Zeniarja, J. (2013). Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. *Procedia Engineering*, *53*, 453–462. https://doi.org/10.1016/j.proeng.2013.02.059.

Zhang, L., Hua, K., Wang, H., Qian, G., & Zheng, L. (2014). Sentiment analysis on reviews of mobile users. *Procedia Computer Science*, *34*, 458–465. https://doi.org/10.1016/j.procs.2014.07.013