Jurnal Ilmiah Teknik Informatika dan Komunikasi
Halaman Jurnal: http://journal.sinov.id/index.php/juitik/index
Halaman UTAMA Jurnal : https://journal.sinov.id/index.php

Komparasi Algoritma Feature Selection Pada Analisis Sentimen Review Film

Iwan Sinanto Ate [a], Ahlijati Nuraminah [b]
[a] Program Studi Ilmu Komputer, i.sinanto.a@students.esqbs.ac.id , STIMIK ESQ
[b] Program Studi Ilmu Komputer, ahlijati.nuraminah@esqbs.ac.id, STIMIK ESQ

ABSTRAK

Sentiment analysis is a process that aims to determine the content of a dataset in the form of text that is positive, negative or neutral. Currently, public opinion is an important source in making a person's decision about a product. Classification algorithms such as Naïve Bayes (NB), Support Vector Machine (SVM), and Artificial Neural Network (ANN) have been proposed by many researchers to be used in sentiment analysis of film reviews. However, text sentiment classification has problems with the number of attributes used in a dataset. Feature selection can be used to reduce irrelevant attributes in the dataset. Several feature selection algorithms used are information gain, chi square, forward selection and backward elimination. Algorithm comparison results, SVM gets the best results with an accuracy of 81.10% and AUC 0.904. The results of the feature selection comparison, information gain get the best results with an average accuracy of 84.57% and an average AUC of 0.899. The integration results of the best classification algorithm and the best feature selection algorithm produce an accuracy of 81.50% and an AUC of 0.929. These results have increased when compared to experimental results using SVM without feature selection. The result of testing the best feature selection algorithm for each classification algorithm is that information gain gets the best results for use in the NB, SVM and ANN algorithms.

***Keywords**: Sentiment analysis, classification, feature selection, support vector machine, artificial neural network, naïve bayes*

ABSTRAK

Analisis sentimen adalah suatu proses yang bertujuan untuk menentukan isi suatu dataset berupa teks positif, negatif, atau netral. Saat ini opini publik merupakan sumber penting dalam pengambilan keputusan seseorang terhadap suatu produk. Algoritma klasifikasi seperti Naïve Bayes (NB), Support Vector Machine (SVM), dan Artificial Neural Network (ANN) telah diusulkan oleh banyak peneliti untuk digunakan dalam analisis sentimen ulasan film. Namun, klasifikasi sentimen teks bermasalah dengan jumlah atribut yang digunakan dalam sebuah dataset. Pemilihan fitur dapat digunakan untuk mengurangi atribut yang tidak relevan dalam dataset. Beberapa algoritma seleksi fitur yang digunakan adalah information gain, chi square, forward selection dan backward eliminasi. Hasil perbandingan algoritma, SVM mendapatkan hasil terbaik dengan akurasi sebesar 81,10% dan AUC 0,904. Hasil perbandingan pemilihan fitur, information gain mendapatkan hasil terbaik dengan akurasi rata-rata 84,57% dan AUC rata-rata 0,899. Hasil integrasi algoritma klasifikasi terbaik dan algoritma pemilihan fitur terbaik menghasilkan akurasi sebesar 81,50% dan AUC sebesar 0,929. Hasil tersebut mengalami peningkatan jika dibandingkan dengan hasil eksperimen menggunakan SVM tanpa seleksi fitur. Hasil pengujian algoritma

Received Maret 10, 2022; Revised Mei 2, 2022; Accepted Juni 07, 2022

seleksi fitur terbaik untuk masing-masing algoritma klasifikasi adalah information gain mendapatkan hasil terbaik untuk digunakan pada algoritma NB, SVM dan JST.

Kata Kunci: Analisis sentimen, klasifikasi, pemilihan fitur, support vector machine, jaringan syaraf tiruan, naïve bayes

1. PENDAHULUAN

Analisis sentimen adalah proses yang bertujuan untuk memenentukan isi dari dataset yang berbentuk teks (dokumen, kalimat, paragraf, dll) bersifat positif, negatif atau netral (Kontopoulos, Berberidis, Dergiades, & Bassiliades, 2013). Analisis sentimen merupakan bidang penelitian yang cukup popular, karena dapat memberikan keuntungan untuk berbagai aspek, mulai dari prediksi penjualan (Yang Liu, Huang, An, & Yu, 2007), politik (Park, Ko, Kim, Liu, & Song, 2011), dan pengambilan keputusan para investor (Dergiades, 2012). Saat ini, pendapat khalayak umum telah menjadi salah satu sumber yang begitu penting dalam berbagai produk di jejaring sosial (C.-L. Liu, Hsaio, Lee, Lu, & Jou, 2012). Demikian juga dalam industri film (Tsou & Ma, 2011). Popularitas internet mendorong orang untuk mencari pendapat pengguna dari internet sebelum membeli produk atau melihat situs film (C.- L. Liu et al., 2012). Pendapat orang-orang dapat mengurangi ketidakpastian terhadap suatu produk tertentu dan membantu konsumen menyimpulkan kualitas suatu produk tertentu (Koh, Hu, & Clemons, 2010).

Beberapa peneliti telah melakukan komparasi menggunakan beberapa algoritma pada beberapa dataset. Penelitian yang dilakukan oleh B. Pang et al (Pang, Lee, Rd, & Jose, 2002) membandingkan algoritma NB, maximum entropy dan SVM. Didapatkan hasil yang terbaik adalah SVM. Rodrigo Moraes et al (Moraes et al., 2013) membandingkan antara ANN, SVM dan NB. Didapatkan hasil yang terbaik adalah ANN. Ziqiong Zhang et al (Z. Zhang, Ye, Zhang, & Li, 2011) membandingkan antara SVM dan NB dan NB merupakan hasil yang terbaik. Songbo Tan et al (S Tan & Zhang, 2008) membandingkan NB, centroid classifier, k-nearest neighbor (KNN), winnow classifier dan SVM merupakan hasil yang terbaik. Dataset yang digunakan dalam penelitian di atas berbeda-beda. Penelitian yang dilakukan oleh B. Pang et all (Pang & Lee, 2002) menggunakan dataset review film. Rodrigo Moraes et al (Moraes et al., 2013) menggunakan dataset review film, Global Positioning System (GPS), buku dan kamera. Ziqiong Zhang (Z. Zhang et al., 2011) et al

menggunakan dataset review restaurant, dan Songbo Tan (Songbo Tan & Wang, 2011) et al menggunakan dataset dokumen berbahasa Cina.

Salah satu masalah pada klasifikasi sentimen teks adalah banyaknya atribut yang digunakan pada sebuah dataset (Wang, Li, Song, Wei, & Li, 2011). Pada umumnya, atribut dari klasifikasi sentimen teks sangat besar, dan jika semua atribut tersebut digunakan, maka akan mengurangi kinerja dari classifier (Wang, Li, Zhao, & Zhang, 2013). Atribut yang banyak membuat accuracy menjadi rendah. Untuk mendapatkan accuracy yang lebih baik, atribut yang ada harus dipilih dengan algoritma yang tepat (Xu, Peng, & Cheng, 2012).

2. PENELITIAN TERKAIT

Salah satu masalah pada klasifikasi sentiment teks adalah data yang berdimensi tinggi sehingga menyebabkan banyaknya atribut yang kurang relevan. Jika semua atribut tersebut digunakan, maka akan mengurangi kinerja dari sebuah classifier (Wang et al., 2013). Atribut yang banyak membuat accuracy menjadi rendah. Untuk mendapatkan accuracy yang lebih baik, atribut yang ada harus dipilih dengan algoritma yang tepat (Xu et al., 2012). Feature selection merupakan bagian penting untuk mengoptimalkan kinerja dari classifier (Wang et al., 2011). Feature selection dapat digunakan untuk mengeliminasi atribut yang kurang relevan (Koncz & Paralic, 2011).

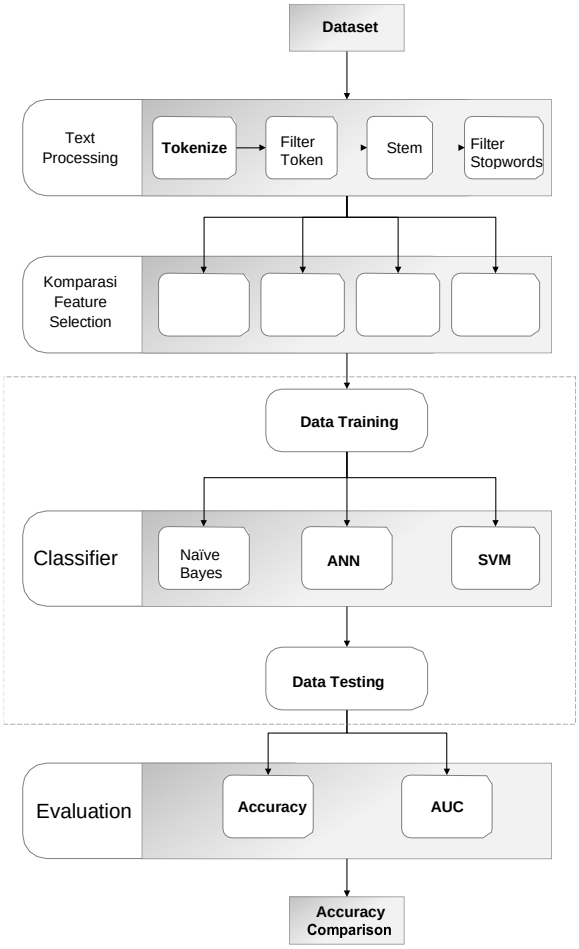
Beberapa peneliti telah mengkomparasi beberapa algoritma klasifikasi dan algoritma feature selection untuk mendapatkan hasil yang terbaik. Penelitian yang dilakukan oleh Peter Koncz dan Jan Paralic (Koncz & Paralic, 2011) menggunakan SVM untuk algoritma klasifikasinya dan algoritma feature selection n-grams+ document frequency dibandingkan dengan Information Gain (IG). Hasil yang diadapatkan IG lebih baik daripada algoritma yang diusulkan.

Rodrigo Moraes, Joao Francisco Valiati, Wilson P (Moraes et al., 2013) mengkomparasi algoritma klasifikasi SVM, Naïve Bayes (NB) dan Artificial Neural Network (ANN). Feature selection yang digunakan adalah expert knowledge, minimum frequency, IG, chi-square. Hasil yang tebaik untuk algoritma klasifikasi adalah ANN dan untuk feature selection terbaik adalah IG.

Zhu Jian, Xu Chen dan Wang Han Shi (Zhu et al., 2010) mengkomparasi algoritma klasifikasi individual model (imodel) berbasis ANN dibandingkan dengan hidden markov model dan SVM. Feature selection yang digunakan adalah odd ratio. Hasil algoritma klasifikasi yang terbaik adalah i-model based on ANN. Songbo Tan dan Jin Zhang (S Tan & Zhang, 2008) mengkomparasi lima algoritma klasifikasi (centroid classifier, K-nearest neighbor, winnow classifier, NB dan SVM), empat algoritma feature selection (Mutual Information, IG, chi-square dan Document Frequency). Hasil eksperimen menunjukan bahwa IG mendapatkan hasil yang terbaik untuk feature selection dan algoritma SVM mendapatkan hasil yang terbaik untuk klasifikasi sentimen.

3. METODOLOGI PENELITIAN

Peneliti mengusulkan untuk mengkomparasi tiga algoritma klasfiikasi (SMV, NB dan ANN) dan mengkomparasi empat algoritma feature selection (IG, Chi Square, Forward Selection dan Backward Elimination). Gambar 1 menunjukan komparasi algoritma klasifikasi dan feature selection yang diusulkan. Sebelum dilakukan komparasi, dataset dilakukan text processing terlebih dahulu. Text processing bertujuan untuk mempersiapkan dokumen teks yang tidak terstruktur menjadi data terstruktur yang siap digunakan untuk proses selanjutnya. Tahapan text processing meliputi: 1. Tokenize merupakan proses untuk memisah-misahkan kata. Potongan kata tersebut disebut dengan token atau term (Manning, Raghavan, & Schutze, n.d.). 2. Filter Token merupakan proses mengambil kata-kata penting dari hasil token (Langgeni, Baizal, & W, 2010). 3. Stem yaitu proses pengubahan bentuk kata menjadi kata dasar. Metode pengubahan bentuk kata menjadi kata dasar ini menyesuaikan struktur bahasa yang digunakan dalam proses stemming (Langgeni et al., 2010). 4. Filter stopwords adalah proses menghilangkan kata-kata yang sering muncul namun tidak memiliki pengaruh apapun dalam ekstraksi sentimen suatu review. Kata yang termasuk seperti kata penunjuk, kata tanya (Langgeni et al., 2010).



Gambar 1. Alur Penelitian

4. HASIL DAN PEMBAHASAN

Penelitian dilakukan menggunakan komputer dengan spesifikasi CPU Intel Core i5 1.6GHz, RAM 8GB, dan sistem operasi Microsoft Windows 7 Professional 64-bit. Aplkiasi yang digunakan adalah RapidMiner 5.2. Data penelitian ini menggunakan Data Movie Review Polarity Dataset V2.0 (Pang & Lee, 2002) yang diperoleh dari data movie review yang digunakan oleh Pang and Lee. Data ini dapat diambil di situs http://www.cs.cornell.edu/people/pabo/movie-review-data/. Data ini diambil dari situs IMDb. Data yang digunakan dalam penelitian terdiri dari 1000 review film, berisi 500 review positif dan 500 review negatif.

Tabel 1. Hasil Komparasi Algoritma

Algoritma	**Accuracy**	**AUC**
ANN	51.80%	0.500
SVM	81.10%	0.904
NV	74.00%	0.734

Tabel 1 merupakan rangkuman hasil komparasi algoritma klasifikasi. Akurasi algoritma ANN sebesar 51.80%, algoritma SVM 81.10% dan algoritma NV sebesar 74%.

Tabel 2. Komparasi Accuracy dan AUC Algoritma Feature Selection

	Information Gain		**Chi Square**		**Forward Selection**		**Backward Elimination**	
	Top K (K=200)		**Top K (K=100)**					
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
ANN	91.40%	0.914	79.60%	0.900	75.50%	0.781	70.20%	0.724
SVM	81.50%	0.929	80.80%	0.853	67.67%	0.698	79.25%	0.844
NV	80.80%	0.853	80.30%	0.867	79.00%	0.807	71.25%	0.689
AVERAGE	84.57%	0.899	80.23%	0.873	74.06%	0.762	73.57%	0.752

Berdasarkan Tabel 1 dan Tabel 2 didapat hasil terbaik adalah SVM dengan accuracy = 81.10% dan AUC = 0.904. Hal ini mengkonfirmasi pada penelitian yang dilakukan oleh Songbo Tan (S Tan & Zhang, 2008) dalam mengkomparasi algoritma klasifikasi, dan SVM mendapatkan nilai yang paling baik. Klasifikasi pada analisis sentimen sangat tergantung pada data yang diuji. Untuk pengujian data IMDB review film, SVM merupakan algoritma yang paling baik.

5. KESIMPULAN

Hasil dari komparasi algoritma klasifikasi antara Support Vector Machine (SVM), Naïve Bayes (NB) dan Artificial Neural Network (ANN) didapatkan SVM dengan hasil terbaik dengan nilai accuracy = 81.10% dan nilai AUC = 0.904. Hasil dari komparasi algoritma feature selection antara information gain, chi square, forward selection, backward elimination didapatkan information gain pada parameter top k dengan nilai k = 200 sebagai hasil terbaik, dengan nilai accuracy average adalah 84.57% dan nilai AUC = 0.899.

DAFTAR PUSTAKA

[1] Forman, G. "An Extensive Empirical Study of Feature Selection Metrics for Text Classification". *Journal of Machine Learning Research*, 3, 1289–1305, 2000. doi:10.1162/153244303322753670

[2] Nugroho, A. S., Witarto, A. B., & Handoko, D. "Support Vector Machine Teori dan Aplikasinya dalam Bioinformatika". IlmuKomputer.Com. 2000

[3] Syahfitra, Febrian Dhimas, et.al. "Implementation of Backpropagation Artificial Neural Network as a Forecasting System of Power Transformer Peak Load at Bumiayu Substation". *Journal of Electrical Technology UMY (JET-UMY)*, Vol. 1, No. 3, September 2017

[4] Syaputri, Astia Weni, Irwandi, Erno & Mustaki. "Naïve Bayes Algorithm for Classification of Student Major's Specialization". *Journal of Intelligent Computing and Health Informatics*, Vol. 1, No. 1, March 2020