TESIS

KOMPARASI DAN INTEGRASI ALGORITMA KLASIFIKASI MACHINE LEARNING DAN *FEATURE SELECTION* UNTUK ANALISIS SENTIMEN REVIEW FILM

Oleh:

VINITA CHANDANI

P31.2011.01066

PROGRAM PASCASARJANA

MAGISTER TEKNIK INFORMATIKA

UNIVERSITAS DIAN NUSWANTORO

SEMARANG

2014

UNIVERSITAS DIAN NUSWANTORO

PENGESAHAN STATUS TESIS

JUDUL : KOMPARASI DAN INTEGRASI ALGORITMA KLASIFIKASI MACHINE LEARNING DAN *FEATURE SELECTION* UNTUK ANALISIS SENTIMEN REVIEW FILM

SAYA : VINITA CHANDANI

mengijinkan Tesis Magister Komputer ini disimpan di Perpustakaan Universitas Dian Nuswantoro dengan syarat-syarat kegunaan sebagai berikut:

1. Tesis adalah hak milik Universitas Dian Nuswantoro
2. Perpustakaan Universitas Dian Nuswantoro dibenarkan membuat salinan untuk tujuan referensi saja
3. Perpustakaan juga dibenarkan membuat salinan Tesis ini sebagai bahan pertukaran antar institusi pendidikan tinggi
4. Berikan tanda √ sesuai dengan kategori Tesis
   - □ Sangat Rahasia
   - □ Rahasia
   - □ Biasa

Disahkan oleh,

.....	.....
Vinita Chandani	Purwanto, Ph.D
Alamat Tetap:	
JL. K. H. Nakhrawi No. 10 Tegal	
Tanggal: 12 Agustus 2014	Tanggal: 12 Agustus 2014

UNIVERSITAS DIAN NUSWANTORO

PERNYATAAN PENULIS

JUDUL TESIS : KOMPARASI DAN INTEGRASI ALGORITMA KLASIFIKASI MACHINE LEARNING DAN *FEATURE SELECTION* UNTUK ANALISIS SENTIMEN REVIEW FILM

PENYUSUN : VINITA CHANDANI

NPM : P31.2011.01066

"Saya menyatakan dan bertanggung jawab dengan sebenarnya bahwa Tesis ini adalah hasil karya saya sendiri kecuali cuplikan dan ringkasan yang masing-masing telah saya jelaskan sumbernya. Jika pada waktu selanjutnya ada pihak lain yang mengklai bahwa Tesis ini sebagai karyanya, yang disertai dengan bukti-bukti yang cukup, maka saya bersedia untuk dibatalkan gelar Magister Komputer saya beserta segala hak dan kewajiban yang melekat pada gelar terserbut."

Semarang, 12 Agustus 2014

VINITA CHANDANI

Penulis

UNIVERSITAS DIAN NUSWANTORO

PERSETUJUAN TESIS

JUDUL : KOMPARASI DAN INTEGRASI ALGORITMA KLASIFIKASI MACHINE LEARNING DAN *FEATURE SELECTION* UNTUK ANALISIS SENTIMEN REVIEW FILM

NAMA : VINITA CHANDANI

NPM : P31.2011.01066

Tesis ini telah diperiksa dan disetujui,

Semarang, 12 Agustus 2014

Purwanto, Ph.D	Romi Satria Wahono, M. Eng
Pembimbing Utama	Pembimbing Pembantu

UNIVERSITAS DIAN NUSWANTORO

PENGESAHAN TESIS

JUDUL : KOMPARASI DAN INTEGRASI ALGORITMA KLASIFIKASI MACHINE LEARNING DAN *FEATURE SELECTION* UNTUK ANALISIS SENTIMEN REVIEW FILM

NAMA : VINITA CHANDANI

NPM : P31.2011.01066

Tesis ini telah diujikan dan dipertahankan dihadapan Dewan Penguji pada Sidang Tesis tanggal 22 Juli 2014. Menurut pandangan kami, Tesis ini memadai dari segi kualitasi maupun kuantitas untuk tujuan penganugerahan gelar Magister Komputer (M. Kom.)

Semarang, 22 Juli 2014

Dewan Penguji

Ika Novita Dewi, S.Kom, M.C.S Anggota	M. Arief Soeleman, M.Kom Anggota
Dr. Abdul Syukur Direktur Pascasarjana	Dr. Abdul Syukur Ketua Penguji

ABSTRACT

Sentiment analysis is the process aiming to determine whether the polarity of a textual corpus (document, sentence, paragraph etc.) tends towards the positive, negative or neutral. People's opinion has become one of the extremely important sources for various services in social networks. Reviews are a major source of information and products can reduce uncertainty and help consumers infer product quality. Classification algorithms such as Naïve Bayes (NB), Support Vector Machine (SVM), and Artificial Neural Network (ANN) were proposed by many researchers to be used in sentiment analysis of movie reviews. Sentiment analysis has a problem on the number attributes used in dataset. Feature selection can be used to reduce the attributes that are less relevant to the dataset. Some feature selection algorithm such as information gain, chi-square, forward selection and backward elimination used in this experiments. The results of the comparison algorithms showed that, SVM obtain the best result with 81.10% accuracy and AUC 0904. The results of the comparison feature selection, information gain to get the best result with 84.57% average accuracy and average AUC is 0.899. The result of the integration of the best classification algorithm and the feature selection algorithm are 81.50% of accuracy and 0.929 of AUC. These results increase when compared to the experimental results using the SVM without feature selection. The testing results of the best feature selection algorithm for each classification algorithm is information gain get the best results for the NB, SVM and ANN. However there are some factors that could be developed to further research, namely: analysis of sentiment for a review of the film refers to an only specific kind of film. Statement on one genre of movie, may be different with other genre film. Like, sad movie review on types of drama, including a positive review, but on a kind of sad comedy review including a negative review. Sentiment analysis on research has not been paying attention to semantics, i.e. the meaning of words and sentences. Further research may use the meaning of words and sentences to determine the sentiment of a document in order to get better results.

Keywords: sentiment analysis, classification, feature selection, support vector machine, artificial neural network, naïve Bayes, information gain, chi-square, forward selection, backward elimination.

ABSTRAK

Analisis sentimen merupakan proses yang bertujuan untuk menentukan isi dari dataset yang berbentuk teks bersifat positif, negatif atau netral. Pendapat khalayak umum menjadi sumber yang penting dalam pengambilan keputusan seseorang akan suatu produk. Algoritma klasifikasi seperti Naïve Bayes (NB), Support Vector Machine (SVM), dan Artificial Neural Network (ANN) diusulkan oleh banyak peneliti untuk digunakan pada analisis sentimen review film. Namun, klasifikasi sentimen teks mempunyai masalah pada banyaknya atribut yang digunakan pada sebuah dataset. *Feature selection* dapat digunakan untuk mengurangi atribut yang kurang relevan pada dataset. Beberapa algoritma *feature selection* yang digunakan adalah information gain, chi square, forward selection dan backward elimination. Hasil komparasi algoritma, SVM mendapatkan hasil yang terbaik dengan *accuracy* 81.10% dan AUC 0.904. Hasil dari komparasi *feature selection*, information gain mendapatkan hasil yang paling baik dengan *average accuracy* 84.57% dan *average* AUC 0.899. Hasil integrasi algoritma klasifikasi terbaik dan algoritma *feature selection* terbaik menghasilkan *accuracy* 81.50% dan AUC 0.929. Hasil ini mengalami kenaikan jika dibandingkan hasil eksperimen yang menggunakan SVM tanpa *feature selection*. Hasil dari pengujian algoritma *feature selection* terbaik untuk setiap algoritma klasifikasi adalah information gain mendapatkan hasil terbaik untuk digunakan pada algoritma NB, SVM dan ANN. Proses pengurangan atribut yang kurang relevan menggunakan algoritma *feature selection* terbukti dapat meningkatkan *accuracy* dibanding dengan menggunakan metode klasifikasi saja. Namun ada beberapa faktor yang dapat dikembangkan untuk penelitian selanjutnya, yaitu: Analisis sentimen untuk review film mengacu pada jenis film tertentu saja. Pernyataan pada salah satu jenis film, mungkin akan berbeda maknanya dengan jenis film yang lain. Seperti, review film sedih pada jenis film drama termasuk review positif, tetapi pada jenis film komedi review sedih termasuk review yang negatif. Analisis sentimen pada penelitian ini belum memperhatikan semantik, yaitu makna kata dan kalimat. Penelitian selanjutnya dapat menggunakan makna kata dan kalimat untuk menentukan sentimen suatu dokumen agar didapat hasil yang lebih baik.

Kata Kunci: analisis sentimen, klasifikasi, *feature selection*, support vector machine, artificial neural network, naïve bayes, information gain, chi square, forward selection, backward elimination.

KATA PENGANTAR

Puji syukur penulis panjatkan kehadirat Tuhan Yang Maha Esa yang telah melimpahkan rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan tesis dengan judul "Komparasi dan Integrasi Algoritma Klasifikasi Machine Learning dan *Feature Selection* untuk Analisis Sentimen Review Film". Dalam penyusunan laporan ini tidak lepas dari kendala dan hambatan, namun berkat bimbingan dan motivasi dari semua pihak yang telah membantu penulis dalam menyelesaikan penyusunan laporan ini, maka penulis tidak lupa menyampaikan ucapan terima kasih kepada:

1. Bapak Dr. Ir. Edi Noersasongko, M.Kom sebagai rektor Universitas Dian Nuswantoro
2. Bapak Dr. Abdul Syukur, MM sebagai direktur Pascasarjana Teknik Informatika Universitas Dian Nuswantoro.
3. Bapak Purwanto, Ph.D sebagai pembimbing utama yang bersedia meluangkan waktunya untuk membimbing penulis dalam menyelesaikan tesis ini.
4. Bapak Romi Satria Wahono, M.Eng sebagai pembimbing tesis yang bersedia meluangkan waktu, tenaga dan pikirannya untuk membimbing penulis dalam menyelesaikan tesis ini.,
5. Orang tua, keluarga, dan teman-teman yang telah memberikan dukungan moral motivasi dan semangat kepada penulis.
6. Seluruh staf pengahar (dosen) Program Pascasarjana Teknik Informatika Universitas Dian Nuswantoro yang telah memberikan pelajaran yang berarti bagi penulis selama menempuh studi.
7. Seluruh staf dan karyawan Program Pascasarjana Teknik Informatika Universitas Dian Nuswantoro yang telah melayani penulis dengan baik selama kuliah.
8. Teman-teman seperjuangan, MTI Regular Malam angkatan 20, khususnya Mas Adi Wibowo, Mba Ade Hikmah, Pak Suwondo, Mas Kiki, Mas Imam, Bu Juju, Pak Masluh, Koh Roy, yang tidak henti-hentinya memberikan semangat dan motivasi.

9. Teman-teman seperjuangan, *Group Intelligent System*, Mba Tyas, Mba Endah.

Serta semua pihak yang tidak dapat disebutkan satu persatu sehingga penulis dapat menyelesaikan tesis ini. Penulis menyadari bahwa penulisan tesis ini masih jauh dari sempurna, untuk itu penulis mohon kritik dan saran yang bersifat membangun demi kesempurnaan penulisan karya ilmiah yang penulis hasilkan yang akan datang. Akhir kata, semoga tesis ini dapat bermanfaat bagi penulis khususnya dan bagi para pembaca pada umumnya.

Semarang, 12 Agustus 2014

Vinita Chandani

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Analisis sentimen adalah proses yang bertujuan untuk memenentukan isi dari dataset yang berbentuk teks (dokumen, kalimat, paragraf, dll) bersifat positif, negatif atau netral [1]. Analisis sentimen merupakan bidang penelitian yang cukup popular, karena dapat memberikan keuntungan untuk berbagai aspek, mulai dari prediksi penjualan [2], politik [3], dan pengambilan keputusan para investor [4].

Saat ini, pendapat khalayak umum telah menjadi salah satu sumber yang begitu penting dalam berbagai produk di jejaring sosial [5]. Demikian juga dalam industri film [6]. Popularitas internet mendorong orang untuk mencari pendapat pengguna dari internet sebelum membeli produk atau melihat situs film [5]. Pendapat orang-orang dapat mengurangi ketidakpastian terhadap suatu produk tertentu dan membantu konsumen menyimpulkan kualitas suatu produk tertentu [7].

Banyak situs yang menyediakan review tentang suatu produk yang dapat mencerminkan pendapat pengguna [5]. Salah satu contohnya adalah situs *Internet Movie Database* (IMDb). IMDb adalah situs yang berhubungan dengan film dan produksi film. Informasi yang diberikan IMDb sangat lengkap. Siapa saja aktor/aktris yang main di film itu, sinopsis singkat dari film, link untuk trailer film, tanggal rilis untuk beberapa negara dan review dari user-user yang lain. Ketika seseorang ingin membeli atau menonton suatu film, komentar-komentar orang lain dan peringkat film biasanya mempengaruhi perilaku pembelian mereka.

Algoritma klasifikasi sentimen seperti naïve bayes (NB) [8], artificial neural network (ANN) [9] [10], support vector machine (SVM) [9] [11] diusulkan oleh banyak peneliti [12] untuk analisis sentimen *review restaurant* [8], dokumen [9] [11], dan teks [10]. ANN mempunyai kelebihan dalam hal kemampuan untuk generalisasi, yang bergantung pada seberapa baik ANN meminimalkan resiko empiris namun ANN mempunyai kelemahan dimana menggunakan data pelatihan cukup besar [13]. SVM

mempunyai kelebihan yaitu bisa diterapkan untuk data yang berdimensi tinggi, tetapi SVM sulit untuk digunakan untuk data dengan jumlah yang besar [14]. NB mempunyai kelebihan mudah diimplementasikan, performance NB lebih baik. Pengklasifikasian pada NB didasarkan pada probabilitas bersyarat dari fitur salah satu kelas setelah fitur seleksi menggunakan algoritma yang ada [15].

Beberapa peneliti telah melakukan komparasi menggunakan beberapa algoritma pada beberapa dataset. Penelitian yang dilakukan oleh B. Pang et al [16] membandingkan algoritma NB, maximum entropy dan SVM. Didapatkan hasil yang terbaik adalah SVM. Rodrigo Moraes et al [9] membandingkan antara ANN, SVM dan NB. Didapatkan hasil yang terbaik adalah ANN. Ziqiong Zhang et al [17] membandingkan antara SVM dan NB dan NB merupakan hasil yang terbaik. Songbo Tan et al [11] 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 [18] menggunakan dataset review film. Rodrigo Moraes et al [9] menggunakan dataset review film, *Global Positioning System* (GPS), buku dan kamera. Ziqiong Zhang [17] et al menggunakan dataset *review restaurant*, dan Songbo Tan [19] et al menggunakan dataset dokumen berbahasa cina.

Salah satu masalah pada klasifikasi sentimen teks adalah banyaknya atribut yang digunakan pada sebuah dataset [20]. Pada umumnya, atribut dari klasifikasi sentimen teks sangat besar, dan jika semua atribut tersebut digunakan, maka akan mengurangi kinerja dari *classifier* [21]. Atribut yang banyak membuat *accuracy* menjadi rendah. Untuk mendapatkan *accuracy* yang lebih baik, atribut yang ada harus dipilih dengan algoritma yang tepat [22].

Feature selection merupakan bagian penting untuk mengoptimalkan kinerja dari *classifier* [20]. *Feature selection* dapat didasarkan pada pengurangan ruang fitur yang besar, misalnya dengan mengeliminasi atribut yang kurang relevan [12]. Penggunaan algoritma *feature selection* yang tepat dapat meningkatkan *accuracy* [22] [23]. Algoritma *feature selection* dapat dibedakan menjadi dua tipe, yaitu *filter* dan *wrapper* [24]. Contoh dari tipe *filter* adalah *information gain* (IG), *chi-square*, dan *log*

likelihood ratio. Contoh dari tipe *wrapper* adalah *forward selection* dan *backward elimination* [25]. Hasil *precision* dari tipe *wrapper* lebih tinggi daripada tipe *filter*, tetapi hasil ini tercapai dengan tingkat kompleksitas yang besar. Masalah kompleksitas yang tinggi juga dapat menimbulkan masalah [12].

Yang dan Perdersen [26] membandingkan lima algoritma *feature selection* pada klasifikasi dokumen. Lima algoritma tersebut adalah *document frequency*, IG, *chi-square*, *term strength* dan *mutual information*. Hasil penelitian mereka menunjukkan bahwa IG dan *chi-square* paling efisien. Forman [23] membandingkan 12 algoritma *feature selection* pada 229 klasifikasi teks menjadi dua kategori. Hasil penelitian menunjukkan IG dan *chi-square* mendapatkan hasil yang lebih baik dibandingkan metode *Bi-Normal Separation* yang diusulkan peneliti. Tan dan Zang [11] menggunakan algoritma *feature selection* untuk analisis sentimen dokumen berbahasa Cina. Hasil yang didapat IG mendapatkan yang paling baik.

Dari semua hasil penelitian yang sudah dilakukan belum ditemukan model yang paling tepat untuk analisis sentimen. Maka dari itu penulis akan melakukan komparasi terhadap beberapa algoritma klasifikasi (NB, SVM dan ANN), komparasi terhadap beberapa algoritma *feature selection* (IG, *chi-square, forward selection, backward elimination*) dan melakukan integrasi dari hasil komparasi algoritma klasifikasi dan algoritma *feature selection* yang terbaik pada dataset review film.

1.2 Identifikasi Masalah

Berdasarkan uraian pada latar belakang, dirumuskan suatu permasalahan yaitu dataset pada analisis sentimen review film mempunyai dimensi yang tinggi, banyak atribut yang kurang relevan sehingga membuat tingkat klasifikasi menjadi rendah.

1.3 Rumusan Masalah

Berdasarkan latar belakang penelitian dan identifikasi masalah di atas, maka rumusan masalah pada penelitian ini, yaitu:

1. Algoritma klasifikasi apa yang paling akurat untuk menangani masalah atribut yang kurang relevan pada analisis sentimen review film?
2. Algoritma *feature selection* apa yang performanya paling baik untuk menangani masalah atribut yang kurang relevan pada analisis sentimen review film?
3. Seberapa meningkat *accuracy* pada analisis sentimen review film bila hasil yang terbaik dari algoritma klasifikasi dan *feature selection* diintegrasikan?
4. Dari hasil komparasi algoritma *feature* selection terbaik, algoritma *feature selection* apa yang tepat untuk setiap algoritma klasifikasi?

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk melakukan komparasi dan integrasi algoritma klasifikasi machine learning dan *feature selection* untuk analisis sentimen review film. Tujuan penelitian secara spesifik dari penelitian ini yaitu:

1. Mengetahui algoritma klasifikasi apa yang paling akurat untuk untuk menangani masalah atribut yang kurang relevan pada analisis sentimen review film.
2. Mengetahui algoritma *feature selection* apa yang performanya paling baik untuk menangani masalah atribut yang kurang relevan pada analisis sentimen review film.
3. Mengetahui seberapa meningkatnya *accuracy* algoritma klasifikasi terbaik dan algoritma *feature selection* terbaik jika diintegrasikan pada analisis sentimen review film.
4. Mengetahui algoritma *feature selection* apa yang paling tepat jika digunakan untuk setiap algoritma klasifikasi dari hasil komparasi algoritma *feature selection* terbaik.

1.5 Batasan Masalah

Algoritma klasifikasi machine learning sangat banyak. Oleh karena itu, penulis membatasi algoritma yang digunakan hanya NB, SVM dan ANN dengan menggunakan dataset IMDb review film.

1.6 Korelasi Masalah – Pertanyaan – Tujuan Penelitian

Berdasarkan masalah penelitian (RP), pertanyaan penelitian (RQ) dan tujuan penelitian (RO) yang telah dijelaskan di atas, maka hubungan antara ketiganya dapat dilihat pada Tabel 1.1 berikut:

Tabel 1.1 Korelasi Masalah – Pertanyaan – Tujuan penelitian

Research Problems (RP)		**Research Questions (RQ)**		**Research Objectives (RO)**	
RP 1	Dataset pada analisis sentimen review film berdimensi tinggi, mempunyai atribut yang kurang relevan sehingga membuat tingkat *accuracy* klasifikasi menjadi rendah	RQ 1	Algoritma klasifikasi apa yang paling akurat untuk menangani masalah atribut yang kurang relevan pada analisis sentimen review film?	RO 1	Untuk mengidentifikasi algoritma klasifikasi yang paling akurat untuk menangani atribut yang kurang relevan apabila diterapkan pada analisis sentimen review film.
		RQ 2	Algoritma *feature selection* apa yang performanya paling baik untuk menangani masalah atribut yang kurang relevan pada analisis sentimen review film?	RO 2	Untuk mengidentifikasi algoritma *feature selection* yang memiliki performa paling baik apabila diterapkan untuk menangani atribut yang kurang relevan pada analisis sentimen review film.