ANALISIS SENTIMEN REVIEW TRANSPORTASI MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE BERBASIS CHI SQUARE

Mokhammad Rifqi Tsani¹, Anggun Prima Gilang Rupaka², Langgeng Asmoro³, Brasie Pradana S B A⁴

e-mail: rifqi@pktj.ac.id; gilang@pktj.ac.id; langgeng@pktj.ac.id; brasie@pktj.ac.id

D4 Teknologi Rekayasa Otomotif, Politeknik Keselamatan Transportasi Jalan

Jl. Semeru No. 3 Tegal-Jawa Tengah, tlp: 0283 351061; fax: 0283 358965.

Abstrak

Analisis sentimen merupakan teknik yang bertujuan menentukan opini dari masyarakat bersifat positif atau negatif. Internet merupakan bagian terpenting dalam kehidupan sehari-hari. Banyak situs yang menyediakan berbagai macam review tentang suatu produk atau jasa yang menggambarkan pendapat pengguna. Salah satu contohnya adalah situs internet Yelp. Yelp adalah sebuah situs yang berisi berbagai macam review, seperti transportasi, media massa, restoran, hotel, makanan, elektronik dan lain sebagainya. Transportasi memiliki peran yang sangat penting dalam menunjang pertumbuhan ekonomi masyarakat dan merupakan urat nadi dalam pembangunan ekonomi suatu negara. Oleh karena itu keberhasilan pembangunan dibidang ekonomi harus ditunjang dengan pengembangan sistem transportasi yang baik, sesuai dengan kebutuhan dan perkembangan zaman. Pengguna transportasi bisa memberikan pendapat mereka mengenai kualitas dari transportasi yang sudah tersedia secara online. Jika membaca review tersebut secara keseluruhan bisa memakan waktu dan sebaliknya jika hanya sedikit review yang dibaca, hasil evaluasi akan bias. Algoritma klasifikasi sentimen review seperti support vector machine (SVM) adalah algoritma yang sering digunakan untuk klasifikasi analisis sentimen review. Namun klasifikasi sentimen review mempunyai kelemahan pada banyaknya fitur atau atribut dataset yang digunakan, metode seleksi fitur chi square dapat digunakan untuk mengurangi fitur atau atribut yang tidak relevan.

Kata kunci : *analisis sentimen, review transportasi, klasifikasi, chi Square.*

1. Pendahuluan

Analisis Sentimen (Sentiment Analysis) merupakan teknik untuk mengenali opini positif maupun negatif [1]. Transportasi adalah perpindahan manusia atau barang dari satu tempat ke tempat lainnya dengan menggunakan sebuah kendaraan yang digerakkan oleh manusia atau mesin. Transportasi digunakan untuk memudahkan manusia dalam melakukan aktivitas sehari-hari. Dengan demikian maka transportasi peran yang sangat penting dalam menunjang pertumbuhan ekonomi masyarakat dan merupakan urat nadi dalam pembangunan ekonomi suatu negara. Menurut [2] Transportasi merupakan penunjang kehidupan manusia dan sangat erat hubungannya dengan kemacetan terutama untuk daerah-daerah perkotaan besar di Indonesia, solusi dari kemacetan salah satunya adalah penggunaan angkutan umum, jika berhasil membuat pengguna kendaraan pribadi beralih ke transportasi umum akan mengurangi jumlah kendaraan di jalan yang tentunya akan mengurangi kemacetan.

Pertumbuhan jaringan sosial yang ada saat ini, membuat konsumen menggunakan konten dalam media untuk membuat keputusan yang lebih baik, analisis sentimen review diakui sebagai sumber informasi yang bermanfaat untuk pemantauan dan meningkatkan tingkat kepuasan konsumen [3]. Beberapa konsumen menuangkan opini atau pengalaman mereka melalui media sosial seperti Facebook, Twitter, atau situs media yang lainnya. Saat ini, pendapat khalayak umum telah menjadi salah satu sumber yang begitu penting dalam berbagai review produk di jejaring sosial [4]. Popularitas internet mendorong orang untuk mencari pendapat pengguna dari internet sebelum membeli produk tertentu. Pendapat orang-orang dapat mengurangi ketidakpastian terhadap suatu produk tertentu dan membantu konsumen menyimpulkan kualitas suatu produk tertentu [5]. Banyak situs yang menyediakan berbagai macam review tentang suatu produk yang dapat mencerminkan pendapat pengguna [4]. Salah satu contohnya adalah situs internet Yelp. Yelp adalah situs yang berisi berbagai macam review, seperti

transportasi, media massa, restoran, hotel, makanan, elektronik dan lain sebagainya. Jika membaca review tersebut secara keseluruhan bisa memakan waktu dan sebaliknya jika hanya sedikit review yang dibaca, hasil evaluasi akan bias. Klasifikasi sentimen bertujuan untuk mengatasi masalah ini dengan secara otomatis mengelompokkan review pengguna menjadi opini positif atau negatif.

Ada beberapa penelitian yang sudah dilakukan dalam hal pengklasifikasian sentimen terhadap review yang tersedia, diantaranya adalah penelitian [6], dimana pengklasifikasian sentimen pada review di internet menggunakan algoritma Support Vector Machine (SVM) dan Naïve Bayes didapatkan hasil yang terbaik adalah SVM. SVM mempunyai kelebihan yaitu bisa diterapkan untuk data yang berdimensi tinggi, tetapi SVM sulit untuk digunakan untuk data dengan jumlah yang besar [7]. Pada umumnya, atribut atau fitur dari klasifikasi sentimen teks sangat besar, dan jika semua atribut tersebut digunakan, maka akan mengurangi kinerja dari klasifikasi sentimen dan SVM. Sehingga akurasi analisis sentimen review dengan SVM [9] menjadi rendah. Algoritma klasifikasi sentimen review seperti SVM [9] diusulkan oleh [6] untuk analisis sentimen review [10].

Seleksi fitur termasuk bagian penting dalam mengoptimalkan sebuah kinerja dari klasifikasi sentimen review [11]. Seleksi Fitur dapat dilakukan untuk pengurangan ruang fitur yang besar, misalnya dengan mengeliminasi atribut atau fitur yang kurang relevan [12]. Penggunaan algoritma seleksi fitur yang tepat dapat meningkatkan akurasi [13]. Terdapat dua kategori seleksi fitur yaitu seleksi fitur supervised dan unsupervised. Seleksi fitur yang termasuk ke dalam kategori supervised adalah chi square, information gain dan mutual information, sedangkan seleksi fitur yang termasuk ke dalam kategori unsupervised adalah term strength, term contribution, entropy-based ranking dan document frequency [14]. Perbedaan antara keduanya adalah keberadaan informasi awal tentang kategori dari suatu dokumen [15]. Dalam pengklasifikasian, chi square adalah salah satu kategori supervised seleksi fitur yang mampu menghilangkan banyak fitur tanpa

mengurangi tingkat akurasi [16] dan chi square juga banyak digunakan untuk menghasilkan kinerja yang bagus dalam keakuratan. Algoritma seleksi fitur chi square dapat mereduksi data yang besar [17], sehingga proses yang dibutuhkan menjadi cepat. Terdapat beberapa algoritma seleksi fitur seperti document frequency, information gain, chi square, term strength dan mutual information. Pada penelitian yang dilakukan [18] menunjukkan bahwa chi square termasuk algoritma seleksi fitur efisien dan lebih baik dari metode frequency-based. Untuk itu chi square tepat digunakan sebagai algoritma seleksi fitur pada analisis sentimen review dengan SVM.

2. Metode Penelitian

Penelitian ini adalah penelitian eksperimen analisis sentimen review dengan dataset review transportasi. Penelitian ini bertujuan untuk menguji model klasifikasi yang terbaik untuk klasifikasi analisis sentimen review. Pada gambar 3.1 menggambarkan model desain penelitian dari tahap pengumpulan data, pengolahan awal, metode yang diusulkan, eksperimen dan pengujian, evaluasi dan validasi.

Gambar 1 Model Desain Penelitian

Pengumpulan Data

Dataset review transportasi yang dikumpulkan dari review yang ada di <http://www.yelp.com/nyc>, data diambil dari tahun 2010 sampai 2019. Review yang digunakan hanya 200 dokumen review transportasi yang terdiri dari 100 review positif dan 100 review negatif. Data tersebut masih berupa sekumpulan teks yang terpisah dalam bentuk dokumen. Data review positif disatukan dalam satu folder dan diberi nama positif, sedangkan data review

negatif disatukan dalam satu folder dan diberi nama negatif.

Pengolahan Awal

Kemudahan proses komputerisasi secara otomatis tergantung pada struktur data yang baik. Pola teks yang sembarang perlu proses perubahan dalam bentuk pola teks yang terstruktur sehingga mempercepat proses penggalian informasi pada analisis sentimen. Hal ini yang disebut sebagai proses preprocessing. Ada beberapa tahapan pada preprocessing adalah sebagai berikut :

1. Tokenization

Dalam proses tokenization ini, semua kata yang ada di dalam tiap dokumen dikumpulkan dan dihilangkan tanda bacanya, serta dihilangkan jika terdapat simbol atau apapun yang bukan huruf.

2. StopwordsRemovel

Dalam proses ini, kata-kata yang tidak relevan akan dihapus, seperti kata the, of, for, with yang merupakan kata-kata yang tidak mempunyai makna tersendiri jika dipisahkan dengan kata yang lain dan tidak terkait dengan dengan kata sifat yang berhubungan dengan sentimen.

3. Stemming

Dalam proses ini kata-kata akan dikelompokkan ke dalam beberapa kelompok yang memiliki kata dasar yang sama, seperti drug, drugged, dan drugs di mana kata dasar dari semuanya adalah kata drug.

Metode yang Diusulkan

Penelitian ini menggunakan algoritma SVM dengan menggunakan seleksi fitur chi square dan bootstrapping untuk ketidakseimbangan kelas.

Ekperimen dan Pengujian

Eksperimen dilakukan dengan melakukan uji coba klasifikasi pada dataset sentimen *review* transportasi. Pengujian pertama dilakukan pada klasifikasi dengan menggunakan metode SVM, pengujian kedua SVM dengan fitur seleksi chi square, menggunakan dataset sentimen *review*.

Evaluasi dan Validasi Penelitian

Pada penelitian klasifikasi untuk evaluasi hasil pengujian dilakukan dengan menghitung nilai dari accuracy, precision dan recall [9]. Accuracy adalah tingkat dari dokumen yang benar diidentifikasi, sedangkan recall adalah tingkat keberhasilan sistem dalam menemukan kembali informasi dan precision adalah perbandingan jumlah data yang sesuai dengan data yang diminta. Persamaan dari ketiganya bisa digambarkan sebagai berikut :

Tabel 1 Confusion Matrix

		Predicted Class	
		Class=Yes	Class=No
Actual Class	Class=Yes	TP(true positif)	FN(false negatif)
	Class=No	FP(false positif)	TN(true negatif)

3. Hasil dan Pembahasan

Dataset *review* Transportasi diambil dari tahun 2015 sampai 2019 di situs review Yelp. *Review* yang digunakan 200 dokumen *review* transportasi yang terdiri dari 100 *review* positif dan 100 *review* negatif dengan 1975 atribut. Data tersebut masih berupa sekumpulan teks yang terpisah dalam bentuk dokumen. Data *review* positif disatukan dalam satu folder dan diberi nama positif, sedangkan data *review* negatif disatukan dalam satu folder dan diberi nama negatif. Pada penelitian sentiment analisis *review* transportasi ini untuk eksperimen pengolahan data menggunakan tools RapidMiner 7.2, proses atau tahapan analisisnya dapat dilihat pada gambar 4.1 proses analisis dengan algoritma SVM dan Chi Square, gambar 4.2 proses preprocessing dan gambar 4.3 proses validation analisis dengan SVM dan Chi Square.

Gambar 2 Proses SVM dan Chi Square

Gambar 3 Proses Pre Processing

Gambar 4 Proses Validation

Pada penelitian klasifikasi untuk evaluasi hasil pengujian dilakukan dengan menghitung nilai dari accuracy, precision dan recall. Eksperimen ini menggunakan data sebanyak 200 dokumen review transportasi. Pada tabel 4.1 menggambarkan confusion matrik dari eksperimen yang dilakukan menggunakan pengujian 7folds menggunakan algoritma SVM dan Chi Square.

Tabel 2 Confusion Matrik Algoritma SVM dan Chi Square

accuracy: 88.29% +/- 4.45% (mikro: 88.29%)			
	True Negatif	True Positif	Class Precision
Prediksi Negatif	115 (TN)	7 (FP)	94.26%
Prediksi Positif	32 (FN)	179 (TP)	84.83%
Class Recall	78.23%	96.24%	

a. Nilai Accuracy

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

$$\text{Accuracy} = \frac{179 + 115}{179 + 115 + 7 + 32}$$

$$\text{Accuracy} = \frac{294}{333}$$

$$\text{Accuracy} = 0.8829 = 88.29 \%$$

b. Nilai Recall

$$\text{Recall} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{179}{179 + 7}$$

$$\text{Recall} = \frac{179}{186}$$

$$\text{Recall} = 0.9624 = 96.24 \%$$

c. Nilai Precision

$$\text{Precision} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{179}{179 + 32}$$

$$\text{Precision} = \frac{179}{211}$$

$$\text{Precision} = 0.8483 = 84.83\%$$

4. Kesimpulan

Dari Hasil penelitian analisis sentiment review Transportasi diperoleh metode Chi Square untuk seleksi fitur dan algoritma SVM menghasilkan akurasi yang lebih baik dalam review transportasi. Pada analisis sentimen review Transportasi menggunakan SVM dengan seleksi fitur Chi Square mendapatkan hasil akurasi sebesar 88.29%. Dengan demikian dapat disimpulkan bahwa penelitian klasifikasi dengan algoritma SVM dan Chi Square pada analisis sentimen review Transportasi dapat juga digunakan untuk menilai dan mengelompokkan sebuah review tentang transportasi yang nantinya dapat digunakan untuk membuat sebuah kebijakan terkait transportasi.

5. Daftar Pustaka

- [1] Nasukawa, Tetsuya, and Jeonghee Yi. "Sentiment analysis: Capturing favorability using natural language processing." *Proceedings of the 2nd international conference on Knowledge capture*. ACM, 2003.
- [2] Depari, Eduard, MacAndrews, and Colin. "Peranan Komunikasi Massa Dalam Pembangunan , Suatu. Kumpulan Karangan, Gadjah Mada University Press, Yogyakarta, 1978,," p. 1978, 1978.
- [3] Kang, Daekook, and Yongtae Park. "Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach." *Expert Systems with Applications* 41.4 (2014): 1041-1050.
- [4] Liu, Chien-Liang, et al. "Movie rating and review summarization in mobile environment." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.3 (2012): 397-407.
- [5] N. Sian, N. Hu, and E. K. Clemons, "Electronic Commerce Research and Applications Do online reviews reflect a product ' s true perceived quality? An investigation of online movie reviews across cultures," *Electron. Commer. Res. Appl.*, vol. 9, no. 5, pp. 374–385, 2010.
- [6] Monarizqa, Nurvirta, and Ir Lukito Edi Nugroho. Penerapan Analisis Sentimen Pada Twitter Berbahasa Indonesia Sebagai Pemberi Rating. Diss. Universitas Gadjah Mada, 2014.
- [7] Nugroho, A. S., Witarto, A. B., & Handoko, D. (2003). "Support Vector Machine Teori dan Aplikasinya dalam Bioinformatika. IlmuKomputer.Com.,," p. 2003, 2003.
- [8] Naufal, Abdul Razak, Romi Satria Wahono, and Abdul Syukur. "Penerapan Bootstrapping untuk Ketidakseimbangan Kelas dan Weighted Information Gain untuk Feature Selection pada Algoritma Support Vector Machine untuk Prediksi Loyalitas Pelanggan." *Journal of Intelligent Systems* 1.2 (2015): 98-108.
- [9] Rakhman, Arif, and Mokhammad Rifqi Tsani. "ANALISIS SENTIMEN REVIEW MEDIA MASSA MENGGUNAKAN METODE C4. 5 BERBASIS FORWARD SELECTION." *Smart Comp: Jurnalnya Orang Pintar Komputer* 8.2 (2019): 78-82.
- [10] L. Ladha and T. Deepa, "Feature Selection Methods And Algorithms," *International Journal on Computer Science and Engineering (IJCSSE)*, vol. 3, no. 5, pp. 1787–1797, 2011.
- [11] S. Wang, D. Li, X. Song, Y. Wei, and H. Li, "A feature selection method based on improved fisher ' s discriminant ratio for text sentiment classification," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8696–8702, 2011.
- [12] Koncz, Peter, and Jan Paralic. "An approach to feature selection for sentiment analysis." *Intelligent Engineering Systems (INES)*, 2011 15th IEEE International Conference on. IEEE, 2011.
- [13] T. Xu, Q. Peng, and Y. Cheng, "Identifying the semantic orientation of terms using S-HAL for sentiment analysis," *Knowledge-Based Syst.*, vol. 35, pp. 279–289, 2012.
- [14] Liu, Tao, et al. "An evaluation on feature selection for text clustering." *Icml*. Vol. 3. 2003.
- [15] Supriyanto, Catur, and Affandy . "Kombinasi Teknik Chi Square Dan Singular Value Decomposition Untuk Reduksi Fitur Pada Pengelompokan Dokumen." *Semantik* 1.1 (2011).