Volume 7 Nomor 1 Januari 2022

INFORMASI INTERAKTIF

JURNAL INFORMATIKA DAN TEKNOLOGI INFORMASI

PROGRAM STUDI INFORMATIKA – FAKULTAS TEKNIK -UNIVERSITAS JANABADRA

INFORMASI INTERAKTIF	Vol. 7	No. 1	Hal. 1 - 78	Yogyakarta Januari 2022	ISSN 2527-5240

INFORMASI INTERAKTIF

ISSN 2527-5240

DEWAN EDITORIAL

Alamat Redaksi : Program Studi Informatika Fakultas Teknik
Universitas Janabadra
Jl. Tentara Rakyat Mataram No. 55-57
Yogyakarta 55231
Telp./Fax : (0274) 543676
E-mail: informasi.interaktif@janabadra.ac.id
Website : http://e-journal.janabadra.ac.id/

Frekuensi Terbit : 3 kali setahun

JURNAL INFORMASI INTERAKTIF merupakan media komunikasi hasil penelitian, studi kasus, dan ulasan ilmiah bagi ilmuwan dan praktisi dibidang Informatika. Diterbitkan oleh Program Studi Informatika Fakultas Teknik Universitas Janabadra di Yogyakarta, tiga kali setahun pada bulan Januari, Mei dan September.

INFORMASI INTERAKTIF ISSN 2527-5240

DAFTAR ISI

ISSN 2527-5240

PENGANTAR REDAKSI

Puji syukur kami panjatkan kehadirat Allah Tuhan Yang Maha Kuasa atas terbitnya JURNAL INFORMASI INTERAKTIF Volume 7, Nomor 1, Edisi Januari 2022. Pada edisi kali ini memuat 10 (sepuluh) tulisan hasil penelitian dalam bidang informatika.

Harapan kami semoga naskah yang tersaji dalam JURNAL INFORMASI INTERAKTIF edisi Januari tahun 2022 dapat menambah pengetahuan dan wawasan di bidangnya masing-masing dan bagi penulis, jurnal ini diharapkan menjadi salah satu wadah untuk berbagi hasil-hasil penelitian yang telah dilakukan kepada seluruh akademisi maupun masyarakat pada umumnya.

Redaksi

IMPLEMENTASI SMOTE UNTUK MENGATASI IMBALANCED DATA PADA SENTIMEN ANALISIS SENTIMEN HOTEL DI NUSA TENGGARA BARAT DENGAN MENGGUNAKAN ALGORITMA SVM

Erry Maricha Oki Nur Haryanto[1], Adhien Kenya Anima Estetikha[2], Rahmad Arif Setiawan[3]

[123] Program Studi MTI, Universitas Amikom Yogyakarta
Jl. Ring Road Utara, Ngringin, Condongcatur, Depok, Sleman, Yogyakarta

Email : [1]errymaricha@janabadra.ac.id

ABSTRACT

The development of a digital platform that connects all tourism stakeholders in Indonesia has been widely applied, especially for lodging services. Dozens of inns with various facilities offered. The development of the world of machine learning has many researchers regarding sentiment analysis that can be associated with the phenomenon of the increasing tourism industry. Many tourists tend to be confused about finding a hotel or inn that suits what they want. One of them is by reading from the reviews of previous visitors. However, sometimes the many reviews create confusion for tourists. Sentiment analysis is an evaluation to determine a person's sentiments, emotions, expressions, and attitudes and usually uses a dataset in machine learning. This research is an analysis of the Support Vector Machine (SVM) algorithm: Sequential Minimal Optimization (SMO) with Synthetic Minority Over-Sampling Technique (SMOTE) for data classification given Sentiment Analysis dataset from reviews of hotel visitors in West Nusa Tenggara from the traveloka site and the collection process it uses scrapy. By applying the imbalance dataset handling method, it is hoped that a classification model with the SVM algorithm will be more accurate and able to handle biases in the classification results. The results of this study using the SVM algorithm without applying the Synthetic Minority Over-Sampling Technique (SMOTE) get an accuracy of 87.62% and the results using the SVM SMOTE algorithm get an accuracy of 87.99%

Keywords: *bias, imbalance dataset, SVM, SMOTE.*

1. PENDAHULUAN

Ulasan pada situs penyedia tiket hotel seperti Traveloka sangat mempengaruhi keputusan calon pengunjung baru, Penelitian yang berjudul sentiment analysis on user satisfaction level of mobile data services using Support Vector Machine (SVM) algorithm bahwa perlu dilakukan klasifikasi pada ulasan dari pengunjung hotel sebelumnya guna mengetahui kepuasan pengunjung selama menginap di hotel tersebut[1]. Selanjutnya penelitian berjudul Comparing SVM dan Naïve Bayes Classifiers for Text Categorization with Wikitology as Knowledge Enrichmentdimana penelitian yang membandingkan antara algoritma naïve bayes dengan SVM dengan hasil naïve bayes lebih baik daripada SVM, dengan akurasi Bayes 28.78% sedangkan SVM 6.36%[2]. Salah satu parameter yang dapat digunakan untuk mengetahui tingkat efektivitas review pengunjung adalah dengan rating review. Rating adalah suatu nilai 1 sampai 10 dari ulasan yang diberikan oleh pengunjung sebelumnya. Rating biasanya diukur dengan pengalaman selama menginap di masa lampau.

Dataset yang digunakan pada penelitian ini diperoleh hasil *scaping* data menggunakan *tool* scrapy. Dataset yang digunakan data historis rating dan ulasan pengalaman menginap. Pada pengumpulan dataset sering terjadi situasi di mana jumlah label data cenderung berat sebelah. Situasi ini disebut masalah *imbalance class*, berdampak menurunnya kinerja algoritma klasifikasi. Untuk mengatasi permasalahan imbalance class, salah satu metode yang digunakan adalah sampling[3].

Dengan mengacu pada penelitian yang terdahulu, peneliti bermaksud untuk melakukan analisis sentimen terhadap ulasan atau review hotel yang ada pada Traveloka. Analisis klasifikasi untuk mencari ulasan positif dan negatif. Pada ekstraksi fitur menggunakan beberapa model dengan kombinasi parameter tf, tf-idf serta stopward pada saat preprosesing dan juga untuk menguji keefektifan metode yang diusulkan, dilakukan dua skenario: pertama, algoritma SVM langsung tanpa memperdulikan imbalance data. Kemudian skenario kedua, Implementasi metode over-sampling SMOTE

digunakan untuk menyeimbangkan dataset dengan menambah jumlah data yang minoritas menjadi dataset yang seimbang.

2. TINJAUAN PUSTAKA

2.1 Data Mining

Data mining adalah topik yang melibatkan pembelajaran dalam arti praktis, non-teoretis. Kami tertarik pada teknik untuk menemukan dan menggambarkan pola struktural dalam data, sebagai alat untuk membantu menjelaskan data itu dan membuat prediksi darinya[4]. Menurut Aswini text mining merupakan keilmuan yang mendalami pada dokumen text. Teks mining adalah teknik penambangan data penting yang mencakup teknik paling sukses untuk mengekstrak pola yang efektif. Dalam bidang text mining, teknik pattern mining digunakan untuk menemukan pola teks, seperti frequent item set, closed frequent item set, co-occurring terms. [5].

2.2 Analisis Sentimen

Suatu metode yang digunakan untuk memahami dan mengolah data tekstual untuk memperoleh informasi negatif, positif dan netral dalam suatu kumpulan dokumen. Analisis sentimen sendiri memiliki level, yaitu level sentimen netral, emosi negatif dan positif per baris, level dokumen, analisis sentimen seluruh dokumen netral atau negatif atau positif dan level aspek, yaitu emosional. analisis dengan menerapkan kelompok ke tingkat ini di mana semua atribut dengan kesamaan dikumpulkan menjadi satu, lalu akhirnya tingkat pengguna, begitulah cara kami menggunakan data analitik untuk berinteraksi dengan lingkungan sosial[6].

2.3 SVM

Metode pengelompokan teknik SVM bergantung pada peningkatan tepi di antara kesempatan dan partisi hyperplane. SVM bukan pengklasifikasi lurus yang dapat mengisolasi kelas secara langsung[7]

2.4 Imbalance Class

Imbalance Class adalah kondisi distriusi yang tidak seimang antara kelas-kelas dari suatu kumpulan data di mana satu kelas memiliki jumlah data yang sangat besar diandingkan dengan kelas-kelas lainnya[8]. Perbedaan jumlah data antar kelas yang besar dapat menyebakan model klasifikasi sering gagal memprediksi dan terjadi bias.

2.5 Synthetic Minority Over-sampling Technique (SMOTE)

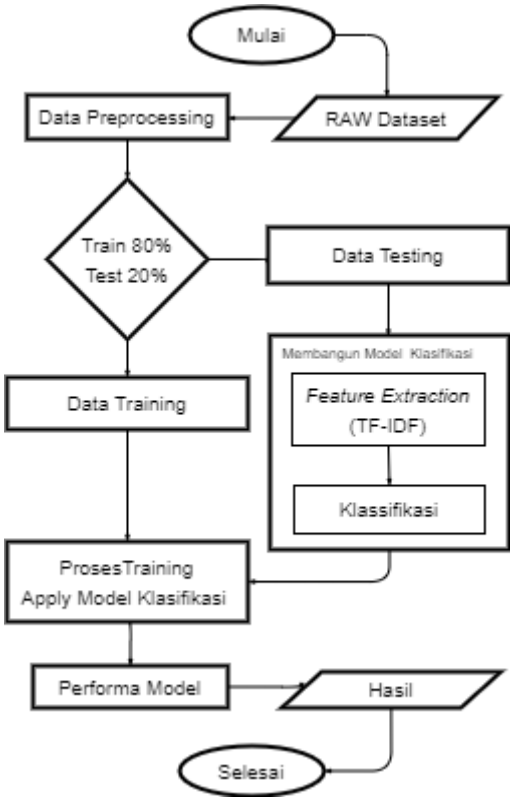
SMOTE adalah metode over-sampling dimana data pada kelas minoritas diperbanyak dengan menggunakan data sintetik yang berasal dari replikasi data pada kelas minoritas. Over-sampling pada SMOTE mengambil instance dari kelas minoritas lalu mencari k-nearest neighbor dari setiap instance, kemudian menghasilkan instance sintetik daripada mereplikasi instance kelas minoritas; oleh karena itu, dapat menghindari masalah overfitting yang berlebihan[8].

Algoritma yang bekerja pada SMOTE pertama akan mengambil nilai selisih antara vektor dari fitur pada kelas minoritas dan nilai nearest neighbor dari kelas minoritas lalu mengalikan nilai tersebut dengan angka acak antara 0 sampai 1. Selanjutnya, hasil kalkulasi tersebut ditambahkan dengan vektor fiturnya sehingga didapatkan hasil nilai vektor yang baru [9]

3. METODE PENELITIAN

Pembuatan model klasifikasi menggunakan algoritma SVM. Klasifikasi respon pengguna hotel di Nusa Tenggara Barat pada penelitian ini menggunakan data rating dari pengguna aplikasi traveloka, dengan jumlah data sebanyak 6.642 ulasan. Pada data ulasan diambil dari tahun 2012 –2021. Dari 15 hotel yang dipilih berdasarkan ulasan terbanyak, jika terdapat data ulasan yang kosong akan dihapus.

Proses pengambilan data ulasan menggunakan tool scrapy. Sehingga mendapatkan atribut data rating ulasan dan isi ulasan.



Gambar 1. Metodologi Penelitian

Metode penelitian yang dilakukan adalah jenis metode penelitian eksperimen, dengan dua skenario utama, yaitu:

A. SVM Imbalance data

Pada sekenario klasifikasi SVM Imbalance data ini. Setelah proses preprosessing dan pada data training tidak dilakukan proses balance data antara positif dan negatif untuk langsung di uji untuk membangun model klasifikasi SVM dan setelah model terbentuk dikalkulasikan dengan data testing.

B. SVM Blance data

Pada sekenario klasifikasi SVM Balance data ini, mengimplementasikan Metode Over-Sampling yang digunakan untuk menyeimbangkan data training. Pada data training yang tidak seimbang label ulasan positif dan negatif, maka akan diseimbangkan dengan teknik SMOTE supaya data antara positif dan negatif untuk membuat model klasifikasi tidak berat sebelah. Teknik SMOTE ini diimplementasikan sebelum membuat sebuah model klasifikasi. Setelah data seimbang, selanjutnya proses pembuatan modelklasifikasi SVM dan di uji dengan data testing

4. HASIL DAN PEMBAHASAN

4.1 Pengumpulan data

Tahapan pengumpulan data adalah pengumpulan data dari ulasan atau review dari website traveloka menggunakan aplikasi scrapy pada phyton dimana aplikasi ini secara otomatis dapat melakukan scrab data ulasan pada Traveloka yang kemudian disajikan dalam format *.csv.

4.2 Pelabelan data

Analisis sentimen ini menggunakan model supervised learning. Pada halaman komentar atau review Traveloka, selain ulasan, pengguna hotel juga bisa memberikan bintang atas pelayanan yang telah dinikmati. Pelabelan data negatif menggunakan tingkatan rating 1 sampai kurang dari 7 sedangkan pelabelan positif menggunakan cara pemilihan ulasan berdasarkan rating 7 sampai dengan 10. Pada penelitian ini tidak menggunakan label netral dikarenakan ulasan yang dibutuhkan hanya positif, dan negative. Berikut beberapa contoh hasil dari proses pelabelan yang dapat dilihat pada Tabel 1

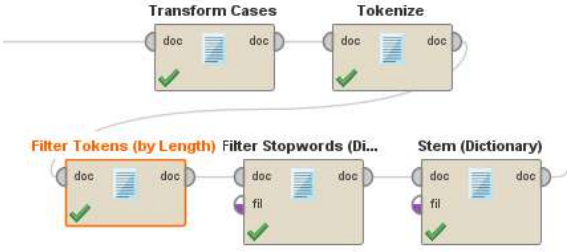
Table 1. Pelabelan data

Rating	Review	Label
7.5	Sangat cocok untuk bulan madu. kekurangannya hanya air kamar mandinya agak payau dan jauh dari penjual makanan.	positif
5.5	Kolam renang kotor, kebersihan kamar buruk masuk ke hotel banyak laron di lantai, menu sarapan terbatas, staf sangat ramah dan view hotel luar biasa	negatif
4	Sayangnya dapat kamar yang tidak sesuai ekspektasi.	negatif
6.5	Memenuhi harapan. pencahayaan yang cukup di kamar mandi dan bau karena kipas tidak bekerja	negatif
9	Romantic dinnernya bener-bener suasana yang bagus. Pelayanannya juga sangat memuaskan.	positif

Dari proses pelabelan didapatkan kelas negatif sebesar 865 data dan kelas positif sebesar 5.777 data. Perbandingan kelas positif dan negatif tidak seimbang yaitu sebesar 87% : 13%.

4.3 Data Preprosessing

Pada data preprosesing, terdapat 3 langkah yang dilalui untuk dapat menghasilkan data yang baik, diantaranya adalah :
Pertama adalah pembersihan data dari noise yang ada yaitu karakter selain huruf baik itu berupa tanda baca, emoticon, dan karakter lain yang tidak diperlukan serta tidak pentin, kedua adalah proses stopword, dan yang terakhir adalah proses tokenizing yaitu memotong dokumen menjadi beberapa bagian yang biasa disebut dengan token, pada proses ini dilakukan pemisahan kata, dan entitas lain.



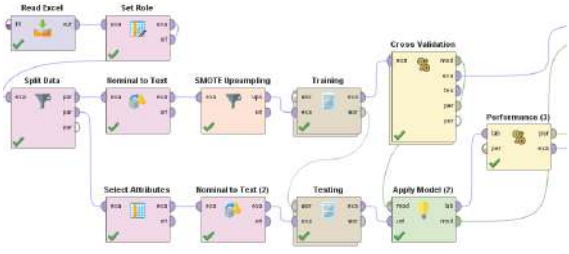
Gambar 2. Proses Teks Preprocessing

4.4 Pemrosesan algoritma SVM dan SMOTE

Data yang telah dilakukan proses preprosesing selanjutnya dilakukan pemrosesan algoritma menggunakan seleksi fitur yaitu TF-IDF atau Term Frequency Inverse Document Frequency dimana ini dilakukan untuk memberikan bobot term secara statistik. Setelah proses pembobotan selesai, dilakukan proses input algoritma SVM dan SMOTE untuk mengatasi imbalance data.

4.5 Pengujian eksperimen

Pengujian dilakukan dengan membagi dataset dengan perbandingan 60% untuk data latih dan 40% untuk data uji. Skema utama pada proses rapid miner seperti pada gambar 3.



Gambar 3. Proses utama Rapid Miner

Pengujian eksperimen dilakukan dengan 2 skema yaitu tanpa dan dengan SMOTE. Proses pengujian dengan SMOTE akan memakan waktu pemrosesan yang lebih lama disbanding tanpa SMOTE.

4.6 Evaluasi dan validasi

Setelah pengujian berhasil selanjutnya hasil pengujian akan menghasilkan matrik konfusi berupa nilai True Positif, True Negatif, False Positif, serta False Negatif. Adapun nilai yang akan diuji adalah nilai akurasi, presisi dan recall. Hasil perbandingan pada pengujian tanpa SMOTE dan dengan SMOTE terdapat pada tabel 2.

Tabel 2. Hasil pengujian

SMOTE	Accuracy	Precision	Recall
-	87.62	88.20	99.00
√	87.99	91.12	95.50

5. KESIMPULAN

Pada penelitian ini diketahui bahwa SMOTE upsampling dapat digunakan untuk menangani dataset dengan kelas yang tidak seimbang (imbalance data). Hasil pengujian dengan penerapan SMOTE memiliki nilai akurasi yang lebih tinggi sebesar 87.99 dengan nilai presisi 91.12 dan nilai recall sebesar 95.00.

DAFTAR PUSTAKA

[1] R. N. Chory, M. Nasrun, and C. Setianingsih, “Sentiment analysis on user satisfaction level of mobile data services using Support Vector Machine (SVM) algorithm,” *Proc. - 2018 IEEE Int. Conf. Internet Things Intell. Syst. IOTAIS 2018*, pp. 194–200, 2019, doi: 10.1109/IOTAIS.2018.8600884.

[2] S. Hassan, M. Rafi, and M. S. Shaikh, “Comparing SVM and Naïve Bayes classifiers for text categorization with Wikitology as knowledge enrichment,” *Proc. 14th IEEE Int. Multitopic Conf. 2011, INMIC 2011*, pp. 31–34, 2011, doi: 10.1109/INMIC.2011.6151495.

[3] R. Barandela, R. M. Valdovinos, J. Salvador Sánchez, and F. J. Ferri, “The imbalanced training sample problem: under or over sampling?,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3138, pp. 806–814, 2004, doi: 10.1007/978-3-540-27868-9_88.

[4] M. A. H. Ian H. Witten, Frank Eibe, *Data Mining Practical Machine Learning Tools and Techniques*. 2008.

[5] V. Aswini and S. K. Lavanya, “Pattern discovery for text mining,” pp. 412–416, 2014, doi: 10.1109/iccpeic.2014.6915399.

[6] A. K. Fauziyyah, "Analisis Sentimen Pandemi Covid19 Pada Streaming Twitter Dengan Text Mining Python," *J. Ilm. SINUS*, vol. 18, no. 2, p. 31, 2020, doi: 10.30646/sinus.v18i2.491.

[7] Y. Al Amrani, M. Lazaar, and K. E. El Kadirp, "Random forest and support vector machine based hybrid approach to sentiment analysis," *Procedia Computer Science*, vol. 127. pp. 511–520, 2018, doi: 10.1016/j.procs.2018.01.150.

[8] and E. A. G. Haibo He, Member, IEEE, "Learning from imbalanced data," *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI*, vol. 2019-Novem, no. 9, pp. 923–930, 2019, doi: 10.1109/ICTAI.2019.00131.

[9] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," *Decis. Anal.*, vol. 2, no. 1, pp. 1–25, 2015, doi: 10.1186/s40165-014-0010-2.