TUGAS AKHIR – KS184822

KLASIFIKASI SENTIMEN TERHADAP *REVIEW* LAYANAN HOTEL BINTANG TIGA DI SURABAYA PADA SITUS *TRAVELOKA* MENGGUNAKAN *NAÏVE BAYES CLASSIFIER* (NBC) DAN REGRESI LOGISTIK BINER

SILVIA ASTRI RAHMANINGRUM
NRP 062117 4500 0011

Dosen Pembimbing
Pratnya Paramitha Oktaviana, S.Si., M.Si.

PROGRAM STUDI SARJANA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA, KOMPUTASI, DAN SAINS DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2019

TUGAS AKHIR – KS184822

KLASIFIKASI SENTIMEN TERHADAP *REVIEW* LAYANAN HOTEL BINTANG TIGA DI SURABAYA PADA SITUS *TRAVELOKA* MENGGUNAKAN *NAÏVE BAYES CLASSIFIER* (NBC) DAN REGRESI LOGISTIK BINER

SILVIA ASTRI RAHMANINGRUM
NRP 062117 4500 0011

Dosen Pembimbing
Pratnya Paramitha Oktaviana, S.Si., M.Si.

PROGRAM STUDI SARJANA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA, KOMPUTASI, DAN SAINS DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2019

FINAL PROJECT – KS184822

SENTIMENT CLASSIFICATION OF THREE STAR HOTEL SERVICE REVIEW IN SURABAYA ON TRAVELOKA SITES USING NAÏVE BAYES CLASSIFIER (NBC) AND BINARY LOGISTIC REGRESSION

SILVIA ASTRI RAHMANINGRUM
SN 062117 4500 0011

Supervisor
Pratnya Paramitha Oktaviana, S.Si., M.Si.

UNDERGRADUATE PROGRAMME
DEPARTMENT OF STATISTICS
FACULTY OF MATHEMATICS, COMPUTING, AND DATA SCIENCE
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2019

LEMBAR PENGESAHAN

KLASIFIKASI SENTIMEN TERHADAP *REVIEW* LAYANAN HOTEL BINTANG TIGA DI SURABAYA PADA SITUS *TRAVELOKA* MENGGUNAKAN *NAÏVE BAYES CLASSIFIER* (NBC) DAN REGRESI LOGISTIK BINER

TUGAS AKHIR

Diajukan untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Statistika
pada
Program Studi Sarjana Departemen Statistika
Fakultas Matematika, Komputasi, dan Sains Data
Institut Teknologi Sepuluh Nopember

Oleh :
Silvia Astri Rahmaningrum
NRP. 062117 4500 0011

Disetujui oleh Pembimbing:
Pratnya Paramitha Oktaviana, S.Si., M.Si. ( )
NIP. 1300201405001

Mengetahui,
Kepala Departemen Statistika

Dr. Suhartono
NIP. 19710929 199512 1 001

SURABAYA, JULI 2019

KLASIFIKASI SENTIMEN TERHADAP *REVIEW* LAYANAN HOTEL BINTANG TIGA DI SURABAYA PADA SITUS *TRAVELOKA* MENGGUNAKAN *NAÏVE BAYES CLASSIFIER* (NBC) DAN REGRESI LOGISTIK BINER

Nama Mahasiswa : Silvia Astri Rahmaningrum
NRP : 062117 4500 0011
Departemen : Statistika
Dosen Pembimbing : Pratnya Paramitha Oktaviana, S.Si., M.Si

Abstrak

Pemanfaatan situs media sosial seperti Traveloka dapat membantu pemasaran hotel dengan menyediakan informasi berupa ulasan terkait pencarian dan pemesanan hotel secara online. Ulasan yang diberikan bisa menjadi feedback bagi hotel terkait serta bisa membantu pengunjung dalam memilih hotel yang tepat. Informasi feedback berupa ulasan merupakan data teks penting sehingga diperlukan suatu metode untuk mengklasifikasikannya. Sumber data didapatkan dari proses web scraping yang bertujuan untuk mendapatkan data secara online pada halaman website dengan mengumpulkan review pengunjung Favehotel dan Hotel Gunawangsa yang bersumber dari situs Traveloka. Penelitian ini menggunakan Naïve Bayes Classifier (NBC) dan Regresi Logistik Biner yang mana labelling sentimen dilakukan berdasarkan lexicon based. Visualisasi word cloud menunjukkan bahwa kata kunci yang terbanyak yang mengarah pada kedua hotel dengan sentimen positif terbesar yaitu kata 'bersih' dan 'nyaman'. Perbandingan metode antara NBC dan Regresi Logistik Biner untuk Favehotel dan Hotel Gunawangsa didapatkan keputusan bahwa metode Regresi Logistik Biner dengan SMOTE lebih baik jika dibandingkan dengan NBC yang mana nilai AUC pada data testing untuk Favehotel sebesar 0,84 dan Hotel Gunawangsa sebesar 0,82.

Kata Kunci: Hotel, Naïve Bayes Classifier, Regresi Logistik Biner, Traveloka

(Halaman ini sengaja dikosongkan)

SENTIMENT CLASSIFICATION OF THREE STAR HOTEL SERVICE REVIEW IN SURABAYA ON TRAVELOKA SITES USING NAÏVE BAYES CLASSIFIER (NBC) AND BINARY LOGISTIC REGRESSION

Name : **Silvia Astri Rahmaningrum**
Student Number : **062117 4500 0011**
Department : **Statistics**
Supervisor : **Pratnya Paramitha Oktaviana, S.Si., M.Si**

Abstract

Utilization of social media sites such as Traveloka can help hotel marketing by providing information as review related to search and hotel bookings online. The given review could be a feedback to the related hotel as well as to assist visitors in choosing the right hotel. Feedback information as a review is important data text so needs a method to classifiy it. Sources of data are obtained from web scraping process which that aims to obtain online data on website page by collecting visitor review of Favehotel and Gunawangsa Hotel sourced from Traveloka sites. This study using Naïve Bayes Classifier (NBC) and Binary Logistic Regression which is the process of sentiment labeling based on Lexicon dictionary. Word cloud visualization shows that the highest keywords that lead to the both of the hotel with the largest positive sentiment are 'clean' and 'comfortable'. A comparison methods between NBC and Binary Logistic Regression for Favehotel and Gunawangsa Hotel obtained a decision that Binary Logistic Regression method with SMOTE was better than NBC where AUC value in testing data for Favehotel was 0,84 and Hotel Gunawangsa was 0,82.

Keywords: ***Binary Logistic Regression, Hotel, Naïve Bayes Classifier, Traveloka***

(Halaman ini sengaja dikosongkan)

KATA PENGANTAR

Alhamdulillah, segala puji syukur bagi Allah SWT yang telah melimpahkan rahmat nikmat dan hidayah kepada makhluk-Nya serta sholawat kepada Nabi Muhammad SAW sehingga penulis dapat menyelesaikan laporan tugas akhir dengan judul **"Klasifikasi Sentimen terhadap *Review* Layanan Hotel Bintang Tiga di Surabaya pada Situs *Traveloka* Menggunakan *Naïve Bayes Classifier* (NBC) dan Regresi Logistik Biner"**. Keberhasilan dalam penyusunan tugas akhir ini tidak terlepas dari bantuan banyak pihak yang telah berperan serta dan membantu suksesnya penulisan laporan akhir ini. Pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih kepada :

1. Ibu Pratnya Paramitha Oktaviana, S.Si, M.Si. selaku dosen pembimbing yang telah membimbing, mengarahkan dan memberikan dukungan bagi penulis untuk dapat menyelesaikan Tugas Akhir ini.
2. Ibu Dr. Dra. Kartika Fithriasari, M.Si dan Ibu Irhamah, M.Si, Ph.D selaku dosen penguji yang telah memberikan saran-saran untuk kesempurnaan Tugas Akhir ini.
3. Bapak Dr. Suhartono selaku Kepala Departemen Statistika ITS yang telah menyediakan fasilitas untuk menyelesaikan Tugas Akhir.
4. Ibu Dr. Santi Wulan Purnami, S.Si., M.Si. selaku Ketua Program Studi Sarjana yang telah membimbing dan memotivasi penulis selama menjadi mahasiswa.
5. Seluruh dosen Departemen Statistika ITS yang telah memberikan wawasan ilmunya selama penulis menempuh pendidikan, beserta seluruh karyawan Departemen Statistika ITS yang telah membantu kelancaran dan kemudahan dalam pelaksanaan kegiatan perkuliahan.
6. Kedua orang tua tercinta, Ibu Sri Iswati, yang telah berjasa, menjadi motivator, serta selalu mendukung dan mendoakan keberhasilan dalam setiap langkah penulis dan (Alm.) Bapak Hari Widodo yang telah memberikan motivasi untuk terus melanjutkan pendidikan. Serta semua keluarga atas

supportnya sehingga dilancarkan dalam menyelesaikan Tugas Akhir ini.

7. Teman-teman S1 LJ Departemen Statistika ITS Angkatan 2017, PIONEER dan PSM ITS yang senantiasa memberikan semangat dan doa selama proses penyelesaian Tugas Akhir serta kebersamaan dan pengalaman yang dilalui selama penulis menjadi mahasiswa.
8. Semua pihak yang telah memberikan dukungan yang tidak dapat disebutkan satu persatu oleh penulis

Dengan selesainya laporan ini, penulis menyadari dalam penulisan laporan tugas akhir ini masih jauh dari kesempurnaan, untuk itu kritik dan saran sangat penulis harapkan demi perbaikan dan kesempurnaan. Semoga laporan akhir ini dapat bermanfaat bagi semua pihak.

Surabaya, Juli 2019

Penulis

DAFTAR ISI

Halaman

DAFTAR TABEL

Halaman

DAFTAR GAMBAR

Halaman

(Halaman ini sengaja dikosongkan)

DAFTAR LAMPIRAN

Halaman

(Halaman ini sengaja dikosongkan)

BAB I
PENDAHULUAN

1.1 Latar Belakang

Industri pariwisata saat ini sedang digencarkan untuk dikembangkan, termasuk di Surabaya karena perkembangan kunjungan wisatawannya terus mengalami peningkatan setiap tahun baik wisatawan nusantara (wisnus) maupun wisatawan mancanegara (wisman). Jumlah kunjungan wisatawan yang datang ke Kota Surabaya pada tahun 2017 menembus angka 24 juta orang dimana angka tersebut naik sebesar 13,72% dibandingkan tahun sebelumnya. Sebagai ibukota provinsi Jawa Timur, Kota Surabaya dinobatkan menjadi kota terbaik versi *Yokatta Wonderful Indonesia Tourism Awards* 2018 karena memiliki performa terbaik khusus untuk pengembangan pariwisata. Ada empat indikator dalam penilaian ini dimana salah satunya mengenai kinerja usaha pariwisata yang meliputi pengembangan akomodasi termasuk makan dan minum (Tempo.co, 2018). Salah satu sarana akomodasi utama bagi wisatawan untuk menentukan tujuan wisatanya adalah adanya hotel. Menurut Badan Pusat Statistik pada Kota Surabaya Dalam Angka 2018 menyatakan bahwa jumlah hotel yang ada di Kota Surabaya sebesar 233 pada tahun 2017, hal ini terjadi peningkatan daripada tahun sebelumnya sebanyak 189 (BPS, 2018).

Saat ini, pertumbuhan pesat media sosial (misalnya : ulasan, diskusi forum, blog, komentar dan postingan pada situs jejaring sosial) di *web* membuat seseorang akan menggunakan media untuk pengambilan keputusan. Sehingga jika seseorang ingin menikmati produk layanan jasa maka seseorang tersebut tidak lagi terbatas untuk meminta pendapat teman dan keluarga karena ada banyak ulasan pengguna dan diskusi di forum publik di *web* tentang produk tersebut ditambah lagi dengan penggunaan internet saat ini merupakan salah satu kebutuhan pokok bagi semua orang di seluruh dunia. Tidak bisa dipungkiri bahwa dengan perkembangan yang begitu pesat memberikan dampak besar bagi aspek kehidupan

diantaranya kemudahan akses informasi yang dapat diakses 24 jam. Pertumbuhan pengguna internet di Indonesia menurut hasil survei yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) menunjukkan bahwa terjadi peningkatan tiap tahun. Pada tahun 2016 jumlah pengguna sebesar 132,7 juta jiwa kemudian di tahun 2017 pertumbuhannya sebesar 143,26 juta jiwa dari total populasi penduduk Indonesia sebanyak 262 juta orang (APJII, 2018).

Banyak kemudahan dan manfaat yang diperoleh dengan adanya internet termasuk diantaranya perusahaan jasa yang bergerak di bidang pariwisata seperti hotel. Seiring dengan berkembanganya media *online*, maka para pelaku industri perhotelan bisa memanfaatkan media tersebut untuk melakukan promosi dan pemasaran. Salah satu situs yang menyediakan layanan *online* pariwisata adalah situs *Traveloka*. *Traveloka* merupakan situs yang menyediakan layanan pencarian dan pemesanan tiket pesawat dan hotel secara *online* yang paling populer di Asia Tenggara, selain itu situs ini juga menawarkan *review* atau ulasan dari jutaan wisatawan khususnya dalam mencari hotel terbaik secara cepat, aman dengan harga termurah tanpa tambahan biaya apapun sebagai opsi dan fitur perencanaan akomodasi di tempat tujuan (Traveloka, 2019).

Keberadaan hotel sangat berperan penting sebagai sarana akomodasi tempat menginap sementara bagi para wisatawan. Industri pariwisata akan mengalami kesulitan dalam perkembangannya tanpa adanya akomodasi penginapan. Pertumbuhan jumlah hotel berbintang menandakan bahwa industri pariwisata juga mengalami perkembangan, namun di sisi lain pertumbuhan hotel akan meningkatkan tingkat persaingan usaha sehingga berbagai upaya promosi dan pemasaran terus dilakukan untuk menarik perhatian pengunjung guna mempertahankan tingkat hunian dengan banyaknya hotel lainnya. Maka dari itu, kepuasan pelanggan harus diperhatikan untuk menjaga stabilitas kedatangan wisnus maupun wisman hingga menjadikan mereka pelanggan tetap bagi suatu hotel tersebut. Sejalan dengan

perkembangan media *online*, kritik dan saran dari pengunjung bisa dilihat oleh banyak orang salah satunya lewat situs *Traveloka.*

Review atau ulasan yang telah ditulis oleh para pengunjung secara *online* di situs *Traveloka* dapat bermanfaat bagi pihak hotel mengenai layanan yang telah diberikan untuk mengetahui kepuasan pelanggan. Selain itu, membantu calon pengunjung sebagai referensi untuk memilih hotel sesuai keinginan. Hotel Gunawangsa MERR dan Favehotel Rungkut Surabaya adalah hotel bintang tiga yang memanfaatkan aplikasi *Traveloka* sebagai pemasarannya. *Review* dari pengunjung ini merupakan data teks penting yang bisa digunakan sebagai bahan evaluasi terhadap kinerja Hotel Gunawangsa MERR dan Favehotel Rungkut Surabaya sehingga kedepannya bisa mempertahankan pelayanan yang sudah baik dan meningkatkan pelayanan yang masih kurang. Berdasarkan *review* yang telah diberikan oleh pengunjung maka diperlukan suatu metode untuk mengklasifikasikan apakah *review* tersebut cenderung positif atau negatif. Metode yang tepat digunakan untuk pengklasifikasian data teks adalah *text mining.*

Text mining bertujuan untuk mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen. Terdapat beberapa metode klasifikasi teks yang digunakan diantaranya *Support Vector Machine* (SVM), *Naïve Bayes Classifier* (NBC), *K-Nearest Neighbor*, *Classification and Regression Trees* (CART). Berdasarkan beberapa metode yang digunakan maka pada penelitian ini menggunakan metode *Naïve Bayes Classifier* (NBC). Metode NBC digunakan karena memiliki kelebihan yaitu hanya membutuhkan sedikit data *training* untuk memperkirakan parameter yang diperlukan untuk klasifikasi dan memiliki akurasi yang tinggi (Dey, Chakraborty, Biswas, Bose, & Tiwari, 2016). Sedangkan metode klasik yang digunakan untuk membandingkan ketepatan klasifikasi dengan NBC yaitu Regresi Logistik Biner. Regresi Logistik Biner digunakan untuk menjelaskan hubungan antara variabel prediktor yang berskala interval atau kategorik dengan variabel respon berupa data biner dimana (Y=1) dianggap

sebagai kejadian sukses, sedangkan (Y=0) dianggap sebagai kejadian gagal. Dalam penelitian ini 1 sebagai pendapat yang positif dan 0 sebagai pendapat negatif.

Penelitian sebelumnya oleh Kurniasari (2018) mengenai data ulasan pengunjung The Phoenix Hotel Yogyakarta yang didapatkan dengan cara *web scraping* pada situs *TripAdvisor* menggunakan metode SVM menunjukkan bahwa akurasi pada ulasan berbahasa Inggris sebesar 96,07% dan untuk ulasan berbahasa Indonesia sebesar 84,77%. Wardhani dkk. (2018) yang melakukan perbandingan antara *Naïve Bayes* dan SVM mengenai analisis sentimen artikel berita di koordinator menteri bidang maritim yang menghasilkan bahwa akurasi NBC sebesar 89,50% memberikan nilai akurasi yang lebih tinggi daripada SVM sebesar 87,50%. Selain itu, penelitian mengenai opini pengunjung hotel bintang lima oleh Putri dan Alamsyah (2017) yang berjudul *Opinion Mining of TripAdvisor Review Towards Five-Star Hotels in Bandung City* menggunakan metode NBC menunjukkan bahwa *accuracy* yang didapatkan sebesar 90%, *precision* 92% dan 89,67% *recall*. Kemudian penelitian yang dilakukan oleh Dey, Chakraborty, Biswas, Bose, & Tiwari (2016) mengenai *Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier* didapatkan bahwa metode terbaik untuk ulasan film adalah NBC karena memberikan akurasi di atas 80%. Penelitian mengenai regresi logistik telah dilakukan oleh Yuniarto (2009) yakni mengenai klasifikasi angkatan kerja propinsi Bengkulu menggunakan metode CART dan regresi logistik yang memberikan kesimpulan bahwa ketepatan klasifikasi regresi logistik sebesar 96,71% lebih tinggi daripada metode CART sebesar 82,23%.

Penelitian yang akan dilakukan pada penelitian ini adalah mengimplementasikan teknik *web scraping* untuk mengumpulkan data *review* pengunjung terhadap layanan hotel bintang tiga yaitu Hotel Gunawangsa MERR dan Favehotel Rungkut Surabaya berbasis *Text Mining* pada situs *Traveloka* untuk kemudian dilakukan analisis data eksploratif pada tingkat hunian kedua hotel

tersebut serta mengklasifikasikannya menjadi sentimen positif dan negatif menggunakan *Naïve Bayes Classifier* (NBC) dan Regresi Logistik Biner. Setelah diperoleh hasil klasifikasi kemudian membandingkan performa klasifikasi antara kedua metode dan melakukan visualisasi dengan *Word cloud*. Variabel yang digunakan pada penelitian ini terdiri dari variabel independen yaitu kata dasar dari *review* yang telah dilakukan *preprocessing* data dan variabel dependen yaitu klasifikasi sentimen *review* positif dan negatif. Melalui penelitian ini diharapkan dapat memberikan saran kepada Hotel Gunawangsa MERR dan Favehotel Rungkut Surabaya dalam hal evaluasi mengenai pelayanan yang harus ditingkatkan dan dipertahankan untuk mengoptimalkan kepuasan pelanggan.

1.2 Rumusan Masalah

Review dari pengunjung terhadap layanan hotel merupakan hal yang tidak dapat dikesampingkan sehingga untuk mempercepat pengklasifikasian pendapat diperlukan metode statistika yang tepat yaitu *Naïve Bayes Classifier* dan metode yang digunakan sebagai pembanding adalah Regresi Logistik Biner yang mana dari kedua metode tersebut akan dipilih satu metode terbaik yang kemudian akan digunakan untuk mengklasifikasikan pendapat masyarakat mengenai layanan Hotel Gunawangsa MERR dan Favehotel Rungkut Surabaya. Selain itu perlu juga diketahui *review* yang sering diutarakan oleh pengunjung terkait kedua hotel tersebut.

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut.

1. Mendapatkan hasil klasifikasi sentimen positif dan negatif dengan metode NBC dan Regresi Logistik Biner.
2. Mengetahui kata yang sering muncul dengan menggunakan visualisasi *Word Cloud.*
3. Memperoleh perbandingan performa klasifikasi antara metode NBC dan Regresi Logistik Biner.

1.4 Manfaat Penelitian

Manfaat yang diperoleh dari penelitian ini adalah membantu pihak hotel yang bersangkutan dalam memahami *review* pengunjung baik positif maupun negatif mengenai layanan Hotel Gunawangsa MERR dan Favehotel Rungkut Surabaya. Serta kedepannya diharapkan bisa menjadi bahan evaluasi mengenai pelayanan yang harus ditingkatkan dan dipertahankan untuk mengoptimalkan kepuasan pelanggan.

1.5 Batasan Masalah

Batasan masalah yang digunakan pada penelitian ini sebagai berikut.

1. Penelitian menggunakan studi kasus Hotel Bintang Tiga di Surabaya yaitu Hotel Gunawangsa MERR dan Favehotel Rungkut Surabaya.
2. Data yang digunakan berupa *review* atau ulasan berbahasa Indonesia dari situs *Traveloka.*
3. Penggunaan kamus *lexicon* untuk tahap *labelling* sentimen positif dan negatif.

BAB II
TINJAUAN PUSTAKA

2.1 *Text Mining*

Menurut Feldman and Sanger (2007), *text mining* dapat didefinisikan sebagai sebuah proses menggali informasi dimana seorang *user* berinteraksi dengan sekumpulan dokumen menggunakan *tools analysis* serta mampu mengeskstraksi informasi yang berguna dari sumber data melalui identifikasi dan eksplorasi pola yang menarik. *Text mining* memberikan solusi dari permasalahan seperti pemrosesan, pengorganisasian atau pengelompokan dan menganalisa *unstructured text* dalam jumlah besar. *Text Mining* merupakan penerapan konsep dari teknik *data mining* untuk mencari pola dalam teks dan bertujuan untuk mencari informasi yang berguna dari sekumpulan dokumen berupa data teks yang memiliki format tidak terstruktur atau minimal semi terstruktur (Korhonen dkk., 2012). Adapun tugas khusus dari *text mining* yaitu pengkategorisasian teks (*text categorization*) dan pengelompokan teks (*text clustering*) (Kang, Ahn, & Lee, 2018).

Text categorization atau *text classification* dapat dianggap sebagai proses untuk membentuk golongan (kelas-kelas) dari dokumen berdasarkan pada kelas kelompok yang sudah diketahui sebelumnya (*supervised*). Sedangkan *text clustering* berhubungan dengan menemukan sebuah struktur kelompok yang belum kelihatan (*unsupervised*). Penerapan *text mining* pada *website* yang berisi pendapat, komentar, *review, feedback* dan kritik merupakan salah satu hal yang sangat penting, karena apabila dikelola dengan baik maka dapat memberikan keuntungan berupa informasi yang bermanfaat untuk membantu individu atau organisasi di dalam pengambilan suatu keputusan (Li & Liu, 2010).

2.2 *Sentiment Analysis*

Menurut Bing Liu dalam buku *Sentiment Analysis and Opinion Mining, sentiment analysis* atau biasa disebut *opinion mining* merupakan salah satu cabang penelitian *Text Mining* yang

bertujuan menganalisis pendapat, sentimen, evaluasi, sikap, penilaian dan emosi seseorang melalui suatu topik, produk, layanan, organisasi, individu dan atribut lainnya. *Sentiment analysis* dapat digunakan dalam berbagai kemungkinan domain dari produk konsumen, jasa kesehatan, jasa keuangan, peristiwa sosial hingga pemilihan politik. Kecenderungan penelitian tentang analisis sentimen berfokus pada pendapat yang menyatakan atau menyiratkan suatu sentimen positif dan negatif. Suatu pendapat dapat mewakili hampir semua aktifitas manusia, karena pendapat dapat mempengaruhi terhadap perilaku seseorang. Setiap kita perlu membuat keputusan maka kita ingin tahu pendapat orang lain tersebut. Dalam dunia nyata, bisnis dan organisasi selalu ingin melihat opini publik tentang suatu produk atau jasa (Liu, 2012). Pendekatan untuk analisis sentimen dapat dilakukan dengan *supervised learning,* yaitu mempelajari model klasifikasi berdasarkan serangkaian data berlabel. Pelabelan data *review* dalam penelitian ini menerapkan pendekatan berbasis kamus (*lexicon-based aproach*) yaitu pemberian label berdasarkan kata sentimen yang mana kata-kata tersebut mengekspresikan sentimen positif atau negatif. Untuk memperoleh label sentimen, *lexicon* bekerja dengan mengikuti aturan seperti pada persamaan (2.1) (Mohammad, Kiritchenko, & Zhu, 2013).

Skor = jumlah kata positif – jumlah kata negatif (2.1)

Skor merupakan suatu indikator yang digunakan untuk pelabelan sentimen dari suatu *review*. Pada penelitian ini, suatu dokumen diartikan sebagai *review* yang mana jika *review* memiliki skor > 0 akan diklasifikasikan ke dalam kelas positif. Sedangkan *review* yang memiliki skor < 0 akan diklasifikasikan ke dalam kelas negatif.

Ilustrasi pelabelan berdasarkan *lexicon* pada *review* adalah "kamarnya bersih, rapi dan nyaman, namun kekurangannya adalah sinyal wifi yang lambat". Terdapat 3 kata positif dan 1 kata negatif yang terdeteksi pada kamus *lexicon*, yaitu "bersih","rapi", dan "nyaman" sebagai kata positif, sedangkan untuk kata "lambat"

sebagai kata negatif. Berdasarkan persamaan (2.1) maka hasil perhitungannya sebagai berikut.
Skor = jumlah kata positif – jumlah kata negatif
Skor = 3 - 1
Skor = 2

Skor akhir yang didapatkan dari hasil ilustrasi perhitungan bernilai 2 yang artinya skor bernilai > 0 maka hasil klasifikasinya masuk ke dalam kelas positif.

2.3 *Web Scraping*

Web Scraping adalah proses pengambilan sebuah dokumen semi-terstruktur dari internet, umumnya berupa halaman-halaman *web* dalam bahasa *markup* seperti HTML atau XHTML, dan menganalisis dokumen tersebut untuk diambil data tertentu dari halaman tersebut untuk digunakan bagi kepentingan lain (Turland, 2010). Langkah-langkah *web scraping* menurut Josi, Suryayusra & Abdillah (2014) dalam penelitiannya adalah sebagai berikut.

1. *Create Scraping Template* : Pembuat program mempelajari dokumen HTML dari *website* yang akan diambil informasinya dari tag *HTML* yang mengapit informasi yang akan diambil.
2. *Explore Site Navigation* : Pembuat program mempelajari teknik navigasi pada *website* yang akan diambil informasinya untuk ditirukan pada aplikasi *web scraper* yang akan dibuat
3. *Automate Navigation and Extraction* : Berdasarkan informasi yang didapatkan dari langkah 1 dan 2, aplikasi *web scraper* dibuat untuk mengotomatisasi pengambilan informasi dari *website* yang ditentukan
4. *Extracted Data and Package History* : Informasi yang didapat dari langkah 3 disimpan dalam tabel atau tabel-tabel *database.*

2.4 Praproses Teks

Teks yang akan dilakukan proses *text mining*, pada umumnya memiliki beberapa karakteristik diantaranya adalah memiliki dimensi yang tinggi, terdapat *noise* pada data, dan terdapat struktur teks yang tidak baik. Sebelum menentukan fitur-fitur yang mewakili, diperlukan tahap *preprocessing* yang dilakukan secara umum dalam teks mining pada dokumen dengan tahapan sebagai berikut.

a. *Case Folding*, merupakan proses untuk mengubah semua karakter teks dalam dokumen menjadi bentuk standar (dalam hal ini huruf kecil atau *lowercase*) (Weiss, Indurkhya, Zhang, & Damerau, 2005).
b. *Cleansing*, merupakan tahapan untuk menghilangkan kata yang tidak diperlukan misalnya tanda baca (*punctuation*), *emoticons*, alamat *website*. Tahap *cleansing* diperlukan karena kata-kata tersebut dianggap *noise* yang tidak diperlukan pada proses *text mining* (Buntoro, Adji, & Purnamasari, 2014).
c. *Stemming* merupakan proses untuk mendapatkan kata dasar dengan cara menghilangkan awalan, akhiran, sisipan dan *confixes* (kombinasi dari awalan dan akhiran) (Tala, 2003).
d. *Stopwords* merupakan kosakata yang bukan termasuk kata unik atau ciri pada suatu dokumen atau tidak menyampaikan pesan apapun secara signifikan pada teks atau kalimat. Contoh *stopword* adalah "yang", "dan", "di","dari" dan seterusnya (Putri, 2016).
e. *Tokenizing* merupakan proses memecah sebuah kalimat menjadi kata-kata atau memutus urutan string menjadi potongan-potongan seperti kata-kata berdasarkan tiap kata yang menyusunnya. Tahap selanjutnya yaitu penguraian teks yang semula berupa kalimat yang berisi kata-kata lalu memotong string berdasarkan tiap kata yang menyusunnya, proses tokenisasi mengandalkan karakter spasi pada dokumen teks untuk melakukan pemisahan (Putri, 2016).

2.5 *Term Frequency – Invers Document Frequency* (TF-IDF)

Term Frequency – Invers Document Frequency (TF-IDF) merupakan salah satu proses ekstraksi fitur dengan memberikan nilai pada masing-masing kata. Untuk mengetahui seberapa penting sebuah kata mewakili sebuah kalimat, akan dilakukan pembobotan atau perhitungan. Pemberian skor dalam TF-IDF berdasarkan frekuensi munculnya kata dalam dokumen (Pravina, Cholissodin, & Adikara, 2018). Nilai TF-IDF dapat ditemukan dengan persamaan (2.2) dan (2.3).

$$w_{ji} = tf_{ji} \times idf \tag{2.2}$$

$$idf = \log\left(\frac{n}{df_j}\right) \tag{2.3}$$

Keterangan :

w_{ji} = bobot dari kata *j* pada *review* ke-*i*

tf_{ji} = jumlah kemunculan kata *j* pada *review* ke-*i*

n = jumlah seluruh *review*

df_j = banyaknya *review* yang mengandung kata *j*

Tabel 2.1 Ilustrasi Perhitungan TF dan IDF

***Review* ke-**	**Kata**			
	bersih	**nyaman**	**senang**	**ramah**
1	1	2	0	1
2	1	2	0	1
3	0	1	1	0
4	2	1	2	0
5	0	0	1	0
Document Frequency	3	4	3	2
Inverse Document Frequency (IDF)	0,221	0,096	0,221	0,397

Perhitungan dengan persamaan (2.3) yang disajikan pada Tabel 2.1 diperoleh nilai IDF untuk kata “bersih” sebesar 0,221, kata “nyaman” sebesar 0,096, kata “senang” sebesar 0,221 dan kata “ramah” sebesar 0,397. Setelah mendapatkan hasil dari *Document*

Frequency (DF) dan *Inverse Document Frequency* (IDF) dari setiap kata maka langkah selanjutnya dilakukan perhitungan dengan *Term Frequency Inverse Document Frequency* (TF-IDF). Ilustrasi hasil perhitungan TF-IDF ditampilkan pada Tabel 2.2 sesuai dengan persamaan (2.2).

Tabel 2.2 Ilustrasi Perhitungan TF-IDF

Review	**Kata**			
ke-	**bersih**	**nyaman**	**senang**	**ramah**
1	0,221	0,193	0	0,397
2	0,221	0,193	0	0,397
3	0	0,096	0,221	0
4	0,443	0,096	0,443	0
5	0	0	0,221	0

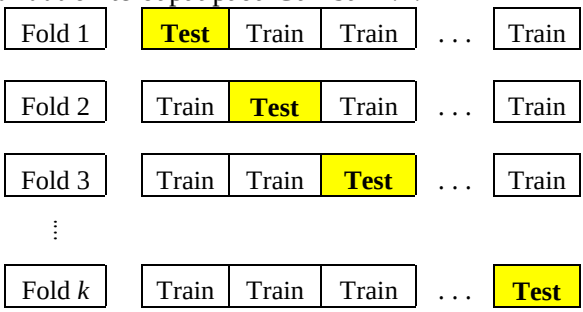
Tabel 2.2 menunjukkan bahwa TF-IDF dari kata "bersih" pada *review* ke 1 dan 2 masing-masing sebesar 0,221 serta pada *review* ke 4 sebesar 0,443. Nilai dari TF-IDF nantinya yang akan digunakan dalam analisis menggunakan metode *Naïve Bayes Classifier* dan Regresi Logistik Biner.

2.6 *K-Fold Cross Validation*

Cross validation merupakan salah satu teknik untuk menilai/memvalidasi keakuratan sebuah model yang dibangun berdasarkan dataset tertentu. Model yang dibuat nantinya bertujuan untuk melakukan prediksi atau klasifikasi terhadap suatu data baru yang belum ada di dalam dataset. Data yang digunakan dalam proses pembangunan model disebut data latih (*training*), sedangkan data yang akan digunakan untuk memvalidasi model disebut sebagai data tes (*testing*). *K-Fold Cross Validation* adalah salah satu metode dari *cross validation* yang digunakan untuk mempartisi data menjadi data *training* dan *testing*, dimana setiap data mendapat kesempatan menjadi data *testing* (Refaeilzadeh, Tang, & Liu, 2009).

Dalam *K-Fold Cross Validation* data dibagi menjadi *k* buah segmen yang memiliki rasio yang sama atau hampir sama lalu dilakukan *training* dan validasi sebanyak *k* kali dengan tiap perulangannya mengambil satu segmen berbeda sebagai data *testing*

atau validasi dan *k*-1 segmen sisanya sebagai data *training* kemudian diambil nilai rata-rata dari setiap hasil iterasi.prosedur ini diulangi *k* kali sehingga setiap partisi digunakan untuk *testing* tepat satu kali. Nilai *k* yang sering digunakan adalah 10 karena dapat menghasilkan estimasi error yang paling baik dan membagi data menjadi proporsi seimbang (Witten, Frank, & Hall, 2011). Ilustrasi pembagian data *training-testing* menggunakan *K-Fold Cross Validation* terdapat pada Gambar 2.1.



Gambar 2.1 Ilustrasi Pembagian Data *Training-Testing*

2.7 *Naïve Bayes Classifier* (NBC)

Naïve Bayes Classifier (NBC) merupakan model penyederhanaan dari algoritma Bayes yang cocok dalam pengklasifikasian teks atau dokumen. Metode ini digunakan untuk mencari nilai probabilitas tertinggi untuk mengklasifikasi data uji pada kategori yang paling tepat (Feldman & Sanger, 2007). Keunggulan dari NBC adalah memiliki akurasi yang relatif tinggi. Metode NBC menempuh dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan (*training*) dan tahap klasifikasi (*testing*) (Hamzah, 2012). Menurut Siang (2005) konsep dasar yang digunakan pada NBC adalah Teorema Bayes yang mana teorema ini mengacu pada konsep probabilitas bersyarat dan disajikan pada rumus persamaan (2.4).

$$P(Y \mid V) = \frac{P(V \mid Y)P(Y)}{P(V)} \tag{2.4}$$

Melalui aturan Teorema Bayes pada rumus (2.4), jika *review* dianggap sebagai v_i maka diasumsikan memiliki koleksi *review* dengan $V = \{v_i \mid i = 1, 2, ... \mid V \mid\} = \{v_1, v_2, ..., v_{|V|}\}$, dan koleksi kategori sentimen dengan $Y = \{y_k \mid k = 1, 2\} = \{y_1, y_2\}$ (Murnawan & Sinaga, 2017). Klasifikasi NBC dilakukan dengan cara mencari probabilitas $P(Y=y_k \mid V=v_i)$ yaitu probabilitas kategori y_k jika diketahui v_i dimana v_i dipandang sebagai kata-kata dalam *review* yaitu $x_1, x_2, ..., x_m$ dimana x_1 adalah kata pertama, x_2 adalah kata kedua dan seterusnya. Sehingga dapat dinyatakan pada persamaan (2.5).

$$P(Y \mid V) = \frac{P(x_1, x_2, ..., x_m \mid y_k) P(y_k)}{P(x_1, x_2, ..., x_m)} \tag{2.5}$$

Pada saat pengklasifikasian *review* teks, maka pendekatan Bayes akan memilih kategori yang memiliki probabilitas paling tinggi. Adapun persamaan V_{MAP} terdapat pada persamaan (2.6).

$$V_{MAP} = \arg\max_{y_k = Y} \frac{P(x_1, x_2, ..., x_m \mid y_k) P(y_k)}{P(x_1, x_2, ..., x_m)} \tag{2.6}$$

karena nilai $P(x_1, x_2, ..., x_m)$ konstan, maka persamaan (2.6) menjadi persamaan (2.7)

$$V_{MAP} = \arg\max_{y_k = Y} P(x_1, x_2, ..., x_m \mid y_k) P(y_k) \tag{2.7}$$

dengan mengasumsikan bahwa setiap kata dalam $x_1, x_2, ..., x_m$ adalah *independent* maka $P(x_1, x_2, ..., x_m \mid y_k)$ pada persamaan (2.7) dapat ditulis sebagai persamaan (2.8)

$$P(x_1, x_2, ..., x_m \mid y_k) = \prod_{j=1}^{m} P(x_j \mid y_k) \tag{2.8}$$

Sehingga persamaan (2.8) dapat ditulis menjadi persamaan (2.9)

$$V_{MAP} = \arg\max_{y_k = Y} P(y_k) \prod_{j=1}^{m} P(x_j \mid y_k) \tag{2.9}$$

Nilai $P(y_k)$ dihitung pada saat *training*, didapat dengan rumus persamaan (2.10).

$$P(y_k) = \frac{|docs\ k|}{|contoh|} \quad (2.10)$$

dengan, $|docs\ k|$ merupakan jumlah *review* yang memiliki kategori *k* dalam *training.* Sedangkan $|contoh|$ merupakan jumlah seluruh *review* sampel yang digunakan dalam proses *training.* Setiap probabilitas kata x_j untuk setiap kategori dihitung pada saat *training* pada persamaan (2.11).

$$P(x_j \mid y_k) = \frac{m_j + 1}{|m + kosakata|} \quad (2.11)$$

dimana m_j adalah jumlah kemunculan kata x_j dalam *review* berkategori y_k sedangkan *m* adalah banyaknya seluruh kata dengan kategori y_k dan |*kosakata*| adalah banyaknya kata dalam data *training.*

2.8 Regresi Logistik

Regresi logistik merupakan suatu metode analisis data yang digunakan untuk mencari hubungan antara variabel respon (*y*) yang bersifat dikotomus (dua kategori) atau polikotomus (lebih dari dua kategori) dengan satu atau lebih variabel prediktor (*x*) yang berskala kategori atau kontinyu (Hosmer & Lemeshow, 2000). Misalkan sekumpulan *j* variabel prediktor ditunjukkan sebagai **x** = $(x_1, x_2, ..., x_m)$. Bentuk logit dari regresi logistik multivariabel diberikan pada persamaan (2.14).

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m \quad (2.12)$$

model regresi logistiknya dengan *j* adalah banyaknya variabel prediktor dituliskan pada persamaan (2.15)

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \quad (2.13)$$

Metode yang sederhana namun *powerful* dalam *machine learning* adalah regresi logistik, metode ini bukan sebuah proses *black-boxes* yang menambah daya tariknya untuk digunakan. Regresi logistik tidak membutuhkan asumsi normalitas, linieritas dan homogenitas varians untuk variabel prediktornya (Khempila & Boonjing, 2010). Selain itu, regresi logistik membutuhkan ukuran sampel yang cukup besar (datascienceplus, 2017).

Regresi logistik tidak memerlukan asumsi normalitas, heteroskedastisitas, dan autokorelasi karena variabel respon yang terdapat pada regresi logistik merupakan variabel dikotomi (0 dan 1), sehingga residualnya tidak memerlukan ketiga pengujian tersebut. Salah satu asumsi penting yang harus dipenuhi adalah tidak adanya multikolinieritas. Multikolinieritas merupakan kondisi terdapatnya hubungan linier yang tinggi antar masing-masing variabel prediktor dalam model regresi. Jika terdapat multikolinieritas, maka nilai *standard error* akan membesar seiring dengan tingkat kolinieritas antar variabel prediktor yang cenderung meningkat sehingga mengakibatkan suatu variabel tidak signifikan (Hosmer & Lemeshow, 2000). Salah satu cara untuk mendeteksi multikolinieritas adalah *Variance Inflation Factor* (VIF). Jika nilai VIF > 10 maka menunjukkan adanya multikolinieritas pada data.

2.8.1 Estimasi Parameter

Estimasi parameter dapat dilakukan dengan menggunakan metode *Maximum Likelihood Estimation* (MLE). Fungsi probabilitas mengikuti distribusi *bernoulli* dimana setiap pengamatan (x_i, y_i) ditunjukkan pada persamaan (2.14).

$$f(y_i) = \pi(x_i)^{y_i}[(1-\pi(x_i))]^{1-y_i} \tag{2.14}$$

Apabila antar pengamatan diasumsikan independen maka fungsi *likelihood* dari pengamatan yang independen ditunjukkan pada persamaan (2.15).

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \tag{2.15}$$

β merupakan vektor yang berisi nilai β . Fungsi *likelihood* **(β)** kemudian diubah ke persamaan *ln* dituliskan pada persamaan (2.16).

$$L(\boldsymbol{\beta}) = \ln\left[l(\boldsymbol{\beta})\right] = \sum_{i=1}^{n}\{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\} \quad (2.16)$$

Untuk mencari nilai **β** yang memaksimumkan *L*(**β)** selanjutnya *L*(**β)** diturunkan terhadap β_0 dan β_1 lalu disamadenganlah nol sehingga didapatkan persamaan pada (2.17) dan (2.18)

$$\sum\left[y_i - \pi(x_i)\right] = 0 \quad (2.17)$$

$$\sum x_i\left[y_i - \pi(x_i)\right] = 0 \quad (2.18)$$

Nilai **β** yang didapatkan pada persamaan (2.17) dan (2.18) merupakan hasil dari MLE yang menghasilkan nilai $\hat{\boldsymbol{\beta}}$ sehingga untuk memprediksi nilai pada regresi logistik dapat dituliskan dengan persamaan (2.19) (Hosmer & Lemeshow, 2000).

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i) \quad (2.19)$$

2.8.2 Pengujian Serentak

Pengujian regresi logistik secara serentak bertujuan untuk mengetahui apakah model telah signifikan dengan hipotesis sebagai berikut.

H_0 : $\beta_1 = \beta_2 = \ldots = \beta_m = 0$

H_1 : minimal ada satu $\beta_j \neq 0, j = 1,2,...,m$

Statistik uji yang digunakan dalam pengujian ini terdapat dalam persamaan (2.20) dimana statistik uji G^2 mengikuti distribusi *Chi-square* dengan derajat bebas sama dengan banyaknya parameter. Tolak H_0 dengan taraf signifikan sebesar α bila nilai $G^2 > \chi^2_{(\alpha,p)}$ (Hosmer & Lemeshow, 2000).

$$G^2 = 2\left\{\sum_{i=1}^{n}\left[y_i \ln(\hat{\pi}_i) + (1-y_i)\ln(1-\hat{\pi}_i)\right] - \left[n_1 \ln(n_1) + n_0 \ln(n_0) - n\ln(n)\right]\right\} \quad (2.20)$$

dimana n_0 merupakan jumlah data dari variabel respon yang berkategori 0, n_1 merupakan jumlah data dari variabel respon yang berkategori 1, dan *n* merupakan jumlah keseluruhan data.

2.8.3 Pengujian Parsial

Pengujian parsial dilakukan untuk mengetahui signifikan atau tidaknya setiap parameter variabel prediktor dengan hipotesis sebagai berikut.

$H_0 : \beta_j = 0, j = 1,2,...,m$

$H_1 : \beta_j \neq 0, j = 1,2,...,m$

Statistik uji yang digunakan dalam pengujian ini terdapat dalam persamaan (2.21) dimana dengan taraf signifikan sebesar α maka akan tolak H_0 jika $W > \chi^2_{(\alpha,1)}$.

$$W = \left(\frac{\hat{\beta}_j}{SE\left(\hat{\beta}_j\right)} \right)^2 \tag{2.21}$$

dimana β_j adalah estimasi parameter ke-*j* dan $SE(\beta_j)$ adalah standar eror parameter ke-*j*.

2.9 Performa Klasifikasi

Pengukuran ketepatan klasifikasi dilakukan untuk mengevaluasi seberapa besar kemampuan suatu metode dalam mengklasifikasikan dokumen ke dalam kelas yang tepat. Pengujian akurasi dalam pengujian ini menggunakan *confusion matrix* yang terdapat pada Tabel 2.3.

Tabel 2.3 *Confusion Matrix*

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	*TP*	*FN*
Negatif	*FP*	*TN*

Confusion matrix terdiri dari empat kondisi yaitu *True Positive* (*TP)* yaitu jumlah kalimat bersentimen positif yang tepat diprediksi masuk dalam kelas positif, *False Positive (FP)* yaitu jumlah kalimat bersentimen negatif yang terprediksi masuk dalam kelas positif, *False Negatif (FN)* yaitu jumlah kalimat bersentimen

positif yang terprediksi masuk dalam kelas negatif dan *True Negative* (*TN*) yaitu jumlah kalimat yang tepat diprediksi masuk dalam kelas negatif. Pengukuran yang dapat digunakan untuk menghitung performa klasifikasi diantaranya *accuracy*, *sensitivity* dan *specificity* (Hotho, Numberger, & Paas, 2005) yang ditunjukkan pada persamaan (2.22), (2.23) dan (2.24).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.22}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{2.23}$$

$$specificity = \frac{TN}{TN + FP} \tag{2.24}$$

Accuracy adalah jumlah proporsi prediksi yang benar. *Precision* adalah proporsi jumlah dokumen teks yang relevan terkenali diantara semua dokumen teks yang terpilih oleh sistem. *Recall* adalah proporsi jumlah dokumen teks ynag relevan terkenali diantara semua dokumen teks relevan yang ada pada koleksi. Persamaan untuk *precision* dan *recall* ditunjukkan pada persamaan (2.25) dan (2.26).

$$precision = \frac{TP}{TP + FP} \tag{2.25}$$

$$recall = \frac{TP}{TP + FN} \tag{2.26}$$

Sedangkan untuk data *imbalanced* maka pengukuran ketepatan klasifikasi dapat menggunakan perhitungan *Area Under Curve* (AUC) untuk memperoleh pengukuran ketepatan klasifikasi lebih lanjut (Chawla, 2005). Perhitungan *Area Under Curve* (AUC) dapat dilakukan seperti persamaan (2.27).

$$AUC = \frac{1}{2}\left(sensitivity + specificity\right) \tag{2.27}$$

Nilai AUC terletak pada interval 0 hingga 1, sehingga semakin mendekati 1 maka nilai AUC semakin bagus (Zaky & Meira, 2014). Namun dalam praktiknya, nilai AUC bervariasi

antara 0,5 hingga 1. Tabel 2.4 menunjukkan selang ketepatan klasifikasi yang baik dengan perhitungan AUC yang telah didapatkan (Bekkar, Djemaa, & Alitouche, 2013)

Tabel 2.4 Interpretasi Nilai AUC

Nilai AUC	Keterangan
0,5 – 0,6	Kurang
0,6 – 0,7	Cukup
0,7 – 0,8	Baik
0,8 – 0,9	Sangat Baik
0,9 – 1	Sempurna

2.10 *Syntetic Minority Oversampling Technique* (SMOTE)

Class imbalance adalah sebuah permasalahan yang biasanya ditemukan pada dataset yang mana distribusi antara kelas mayoritas dan minoritas tidak seimbang. Salah satu cara untuk memperbaiki data yang tidak seimbang adalah membuat data menjadi seimbang yang salah satu caranya adalah *oversampling* pada *minority class*. Metode *Syntetic Minority Oversampling Technique* (SMOTE) bekerja dengan mencari *k nearest neighbors* (yaitu ketetanggaan terdekat data sebanyak *k*) untuk setiap data di kelas minoritas, setelah itu dibuat data sintesis sebanyak persentase duplikasi data minor (*percentage oversampling*, N%) yang diinginkan dan *K-Nearest Neighbors* (KNN) yang dipilih secara acak (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

$$N\% = \frac{\text{jumlah data kelas mayoritas}}{\text{jumlah data kelas minoritas}} \times 100\% \tag{2.28}$$

Nearest neighbor dipilih berdasarkan jarak Euclidian antara kedua data. Misalkan diberikan data dengan *p* dimensi yaitu $X^T = [x_1, x_2, \ldots, x_n]$ dan $Y^T = [y_1, y_2, \ldots, y_n]$ maka jarak *euclidean* $d(x,y)$ secara umum sebagai berikut.

$$x_{knn} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} \tag{2.29}$$

Pembangkitan data *syntetic* dilakukan dengan menggunakan persamaan (2.30)

$$x_{syn} = x_i + (x_{knn} - x_i) \times \delta \tag{2.30}$$

keterangan :

x_{syn} adalah data sintesis hasil dari replikasi

x_i adalah data yang akan direplikasi

x_{knn} adalah data dari kelas minor yang memiliki jarak terdekat dari data yang akan direplikasi

δ adalah bilangan random antara 0 dan 1

Data ilustrasi dengan menerapkan metode SMOTE dengan menggunakan 10 sampel disajikan pada Tabel 2.5.

Tabel 2.5 Data Ilustrasi Metode SMOTE

No	Y	bersih	nyaman	senang	ramah
1	0	0,3358	0,3763	0,3050	0,2354
2	0	0	0,1977	0	0,2313
3	0	0,2460	0,4948	0	0
4	1	0	0,2936	0	0
5	1	0	0	0	0,2218
6	1	0,1632	0	0	0,4436
7	1	0	0,1656	0	0,1938
8	1	0	0	0	0,3979
9	1	0	0,6508	0,1529	0
10	1	0,1795	0,3884	0	0

Berdasarkan Tabel 2.5 maka data yang akan direplikasi yaitu data pada kelas minoritas yaitu klasifikasi dengan kategori 0 (Y=0). Jumlah data minoritas sebanyak 3, sedangkan jumlah data mayoritas (Y=1) sebanyak 7, sehingga nilai persentase SMOTE yang akan digunakan adalah (7/3) x 100% = 233,33%.

Selanjutnya dilakukan replikasi 1 kali pada setiap data minoritas dan tetangga data lalu dari data yang direplikasi akan dipilih salah satu yang merupakan tetangga data terdekat (x_{knn}). Pada ilustrasi ini ditentukan dengan menentukan tetangga terdekat (x_{knn}) diawali dengan perhitungan antara *review* 1 dengan *review* 2 lalu membandingkannya dengan perhitungan antara *review* 1 dengan *review* 3.

Review 1 dengan *review* 2

$$d\left(\begin{bmatrix}0,3358\\0,3763\\0,3050\\0,2354\end{bmatrix},\begin{bmatrix}0\\0,1977\\0\\0,2313\end{bmatrix}\right)=\sqrt{(0,3358-0)^2+\ldots+(0,2354-0,2313)^2}=0,4875$$

Review 1 dengan *review* 3

$$d\left(\begin{bmatrix}0,3358\\0,3763\\0,305\\0,2354\end{bmatrix},\begin{bmatrix}0,2460\\0,4948\\0\\0\end{bmatrix}\right)=\sqrt{(0,3358-0,2460)^2+\ldots+(0,2354-0)^2}=0,4129$$

Berdasarkan hasil perhitungan, didapatkan nilai jarak terdekat ada di perhitungan *review* 1 dengan *review* 3. Sehingga tetangga terdekat (x_{knn}) yang digunakan adalah *review* 3. Perhitungan data sintesis dengan persamaan (2.30) dengan menggunakan δ = 0,5 sebagai berikut.

$$x_{syn}=\begin{bmatrix}0,3358\\0,3763\\0,3050\\0,2354\end{bmatrix}+\left(\begin{bmatrix}0,2460\\0,4948\\0\\0\end{bmatrix}-\begin{bmatrix}0,3358\\0,3763\\0,3050\\0,2354\end{bmatrix}\right)\times 0,5=\begin{bmatrix}0,2909\\0,4355\\0,1525\\0,1177\end{bmatrix}$$

Maka data sintesis yang diperoleh untuk replikasi pertama yakni data $bersih_{syn}$ = 0,2909; $nyaman_{syn}$ = 0,4355; $senang_{syn}$ = 0,1525; $ramah_{syn}$ = 0,1177 dan Y_{syn} = 1 dan seterusnya untuk observasi yang direplikasi sebanyak persentase *oversampling* yang telah dihitung sebelumnya.

2.11 *Word Cloud*

Word cloud merupakan sebuah sistem yang memunculkan susunan kata sebagai citra visual terkait frekuensi kemunculan kata dalam suatu teks verbal. Menurut McNaught dan Lam (2010) visualisasi *word cloud* dari teks akan memudahkan pengamat dalam melihat gagasan sehingga dapat menjadi alat bantu dalam melakukan analisis terhadap sebuah wacana tertulis. Maka dengan

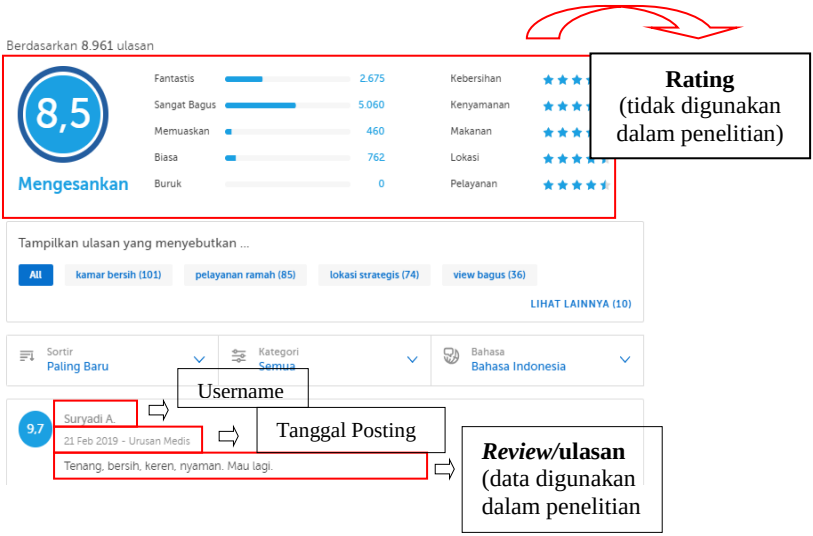
menggunakan *word cloud,* gambaran frekuensi kata-kata dapat ditampilkan dalam bentuk yang menarik dan informatif. Semakin sering satu kata digunakan, maka semakin besar ukuran kata tersebut ditampilkan dalam *word cloud.*

Gambar 2.2 *Wordcloud*

2.12 *Traveloka*

Traveloka merupakan salah satu *online travel agent* di Indonesia yang didirikan pada tahun 2012 oleh Ferry Unardi, Derianto Kusuma, dan Albert Zhang. Situs ini menyediakan layanan pencarian dan pemesanan hotel, tiket pesawat, tiket kereta api, tiket bus dan travel, antar jemput bandara, destinasi wisata, pembelian pulsa san paket internet, pembayaran tagihan listrik PLN, BPJS Kesehatan, Telkom serta PDAM. Para pengguna situs ini juga memberikan keuntungan kepada para *user*nya dengan memberikan berbagai macam promo mulai dari tiket pesawat, hotel, tiket kereta api, tagihan dan isi ulang hingga aktivitas dan rekreasi. Sebagai salah satu *online travel agent*, maka konsumen dapat juga dapat menemukan rute penerbangan dan hotel terbaik secara cepat, aman dengan harga termurah tanpa biaya apapun. Pada awalnya situs *Traveloka* berfungsi sebagai mesin pencari yang digunakan membandingkan harga tiket pesawat dari berbagai situs yang ada, lalu pada tahun 2014 situs *Traveloka* mulai menyediakan jasa pemesanan hotel. Pengunjung yang ingin melakukan reservasi hotel bisa melihat *review* yang diberikan oleh para pengguna layanan *Traveloka* mengenai kualitas hotel hingga pelayanannya (Traveloka, 2019).

Gambar 2.3 Logo *Traveloka*

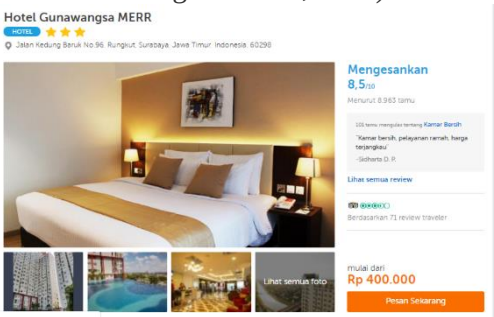


Gambar 2.4 Tampilan *Review* Pelanggan *Traveloka*

2.13 Hotel Gunawangsa MERR

Merupakan hotel yang berlokasi strategis dekat dengan kawasan industri Rungkut yang mengusung *one stop living mixed* dimana terdapat sekolah, hotel, bisnis, toko dan kantor di area ini. Hotel ini memiliki total kamar sebanyak 135 kamar yang terdiri dari 109 kamar Deluxe, 6 kamar Eksekutif, 12 Junior Suite dan 8

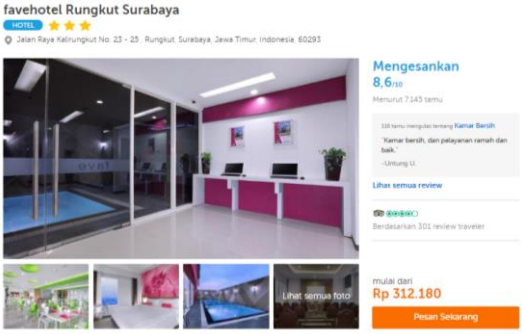
Family Suite dengan pemandangan kota dan kolam renang yang luar biasa (Hotel Gunawangsa MERR, 2016).



Gambar 2.5 Hotel Gunawangsa MERR

2.14 Favehotel Rungkut Surabaya

Hotel ini merupakan sebuah hotel bintang tiga yang berlokasi di kawasan pusat industri Surabaya Selatan yang mana lokasi ini merupakan daerah populer dengan menawarkan rekreasi, organisasi bisnis, bank, univesitas serta hanya 2 menit menuju Transmart Surabaya. Hotel ini dilengkapi dengan 173 kamar modern (favehotel Rungkut, 2019).



Gambar 2.6 Favehotel Rungkut Surabaya

(Halaman ini sengaja dikosongkan)

BAB III
METODOLOGI PENELITIAN

3.1 Sumber Data

Data yang digunakan dalam penelitian ini berasal dari data ulasan/*review* hotel bintang tiga di Surabaya yaitu Hotel Gunawangsa MERR dan Favehotel Rungkut Surabaya yang terdapat pada situs *website Traveloka. Review* tersebut terdata mulai Mei tahun 2016 hingga April tahun 2019.

3.2 Variabel Penelitian

Berdasarkan sumber data yang diperoleh berupa *review* maka data tersebut harus dilakukan *filter* dengan mengambil data yang mengandung sentimen. Variabel penelitian yang digunakan terdapat pada Tabel 3.1.

Tabel 3.1 Variabel Penelitian

Variabel	Keterangan	Skala Data
y	Sentimen (Positif/Negatif) 0 = Sentimen Negatif 1 = Sentimen Positif	Nominal
w	Bobot kemunculan kata kunci	Rasio

Struktur data yang digunakan dalam penelitian ini setelah dilakukan praproses data teks terdiri dari variabel respon (*y*) yang mana 0 dikategorikan sebagai sentimen negatif dan 1 dikategorikan sebagai sentimen posititif. Sedangkan variabel prediktor dalam analisis sentimen ini merupakan bobot kemunculan kata kunci yang diperoleh dari hasil TF-IDF yang dinotasikan dengan (*w).* Struktur data penelitiannya dapat dilihat pada Tabel 3.2.

Tabel 3.2 Struktur Data Setelah *Pre-Processing*

No	*Review* (d_i)	Klasifikasi Sentimen (y)	Kata kunci (w_1)	Kata kunci (w_2)	...	Kata kunci (w_m)
1	d_1	y_1	$w_{1,1}$	$w_{1,2}$	...	$w_{1,m}$
2	d_2	y_2	$w_{2,1}$	$w_{2,2}$	...	$w_{2,m}$
...	...	...	...	...	...	...
...	...	...	...	...	...	...
n	d_n	y_n	$w_{n,1}$	$w_{n,2}$		$w_{n,m}$

3.3 Langkah Analisis

Langkah analisis yang digunakan pada penelitian ini adalah sebagai berikut.

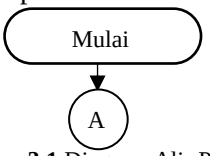
1. Melakukan *web scraping* pada *review* pengunjung Hotel Gunawangsa MERR dan Favehotel Rungkut Surabaya di situs *Traveloka.*
2. Melakukan klasifikasi sentimen berdasarkan *lexicon based* untuk mengelompokkan data *review* menjadi kategori positif atau negatif.
3. Melakukan *preprocessing* data
   a. Melakukan *case folding*, yaitu mengubah semua teks dengan huruf kecil (non kapital).
   b. Melakukan *cleansing*, yaitu menghapus angka, *emoticons*, *punctuation* pada *review.*
   c. Melakukan *stemming*, untuk mendapatkan kata dasar menggunakan algoritma *confix-stripping stemmer.*
   d. Melakukan proses *stopwords*
   e. Melakukan *tokenizing* untuk memecah kalimat *review* menjadi kata per kata
   f. Mengubah kalimat *review* kedalam bentuk bobot kata menggunakan TF-IDF
4. Melakukan visualisasi dengan *Word Cloud*
5. Data *review* dibagi menjadi data *training* dan data *testing* dengan menggunakan 10-*fold cross validation*
6. Klasifikasi data menggunakan *Naïve Bayes Classifier*

a. Menghitung probabilitas y_k pada data *training*, dimana y_k merupakan kategori sentimen, yaitu y_1 = negatif dan y_2 = positif.
b. Menghitung probabilitas kata x_j pada kategori y_k .
c. Menghitung probabilitas *Naïve Bayes Classifier* disimpan dan digunakan untuk tahap *testing.*
d. Menghitung probabilitas tertinggi dari kategori sentimen yang diujikan (V_{MAP}).
e. Mencari nilai V_{MAP} paling maksimum dan memasukkan kalimat tersebut pada kategori dengan V_{MAP} maksimum.
f. Menghitung performa klasifikasi dari model yang terbentuk
g. Mengulang langkah a sampai f dengan menggunakan data SMOTE.

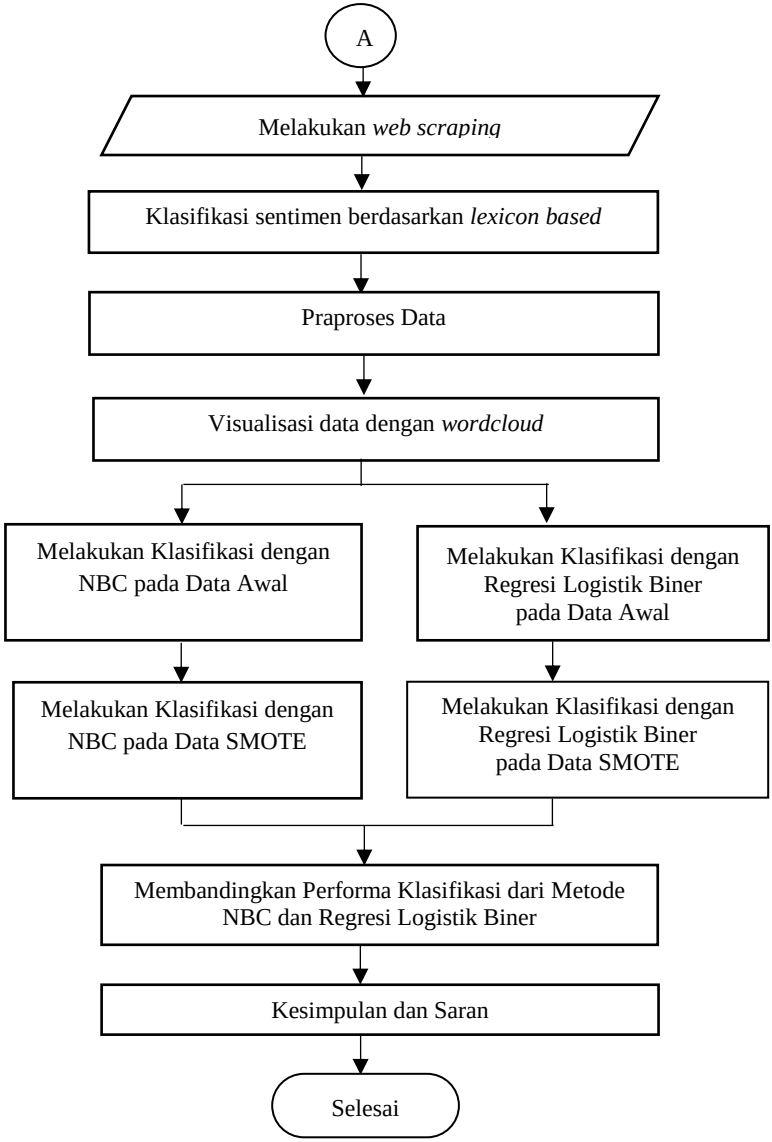
6. Klasifikasi data menggunakan Regresi Logistik Biner
   a. Melakukan pemeriksaan asumsi multikolinieritas.
   b. Melakukan uji signifikansi parameter secara serentak dan parsial.
   c. Menentukan model dari regresi logistik biner.
   d. Menghitung performa klasifikasi dari model yang terbentuk.
   e. Mengulang langkah a sampai d dengan menggunakan data SMOTE.
7. Menentukan metode terbaik dengan membandingkan performansi metode NBC dan regresi logistik biner berdasarkan nilai *Accuracy Under Curve* (AUC) pada data awal dan data SMOTE
8. Membuat kesimpulan dan saran.

3.4 Diagram Alir

Diagram alir pada penelitian ini adalah sebagai berikut.



Gambar 3.1 Diagram Alir Penelitian



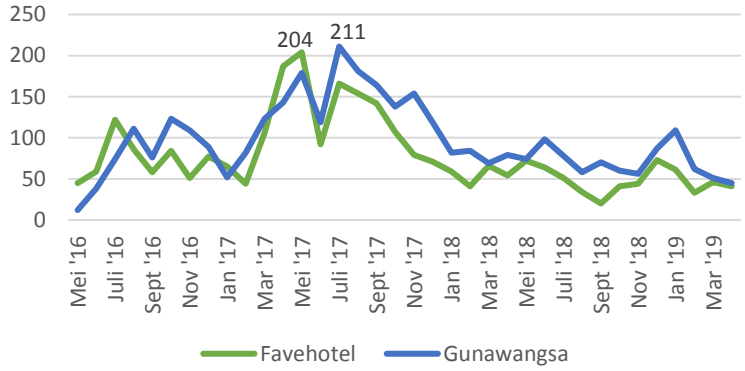
Gambar 3.2 Diagram Alir Penelitian (Lanjutan)

BAB IV
ANALISIS DAN PEMBAHASAN

Bab ini membahas tentang hasil analisis berdasarkan data *review* pengguna situs *Traveloka* terhadap Pelayanan di Hotel Gunawangsa MERR Rungkut dan Hotel Favehotel Rungkut. Klasifikasi sentimen yang digunakan berdasarkan *lexicon based.* Metode yang digunakan adalah *Naive Bayes Classifier* (NBC) dan Regresi Logistik Biner (RLB).

4.1 Karakteristik Data

Fitur *review* yang terdapat pada situs *Traveloka* disediakan untuk para penggunanya untuk mempermudah menyampaikan pendapat terhadap hotel terkait dimana pada penelitian ini yaitu Hotel Gunawangsa MERR dan Favehotel Rungkut. Gambar 4.1 menunjukkan jumlah *review* pengunjung hotel yang menggunakan aplikasi *Traveloka* pada bulan Mei 2016 hingga April 2019.

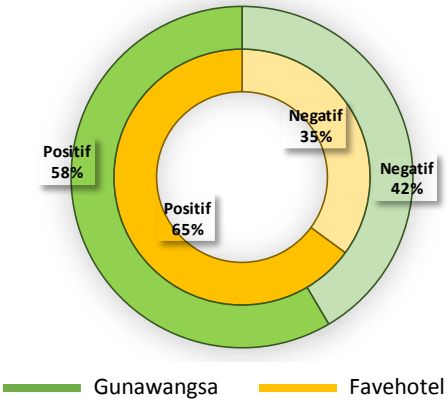


Gambar 4.1 Tren Jumlah *Review* Pengunjung Hotel

Gambar 4.1 menunjukkan jumlah *review* pengunjung hotel yang menggunakan aplikasi *Traveloka* yang cenderung stabil pada Januari 2018. Pengunjung yang memberikan *review* terbanyak pada Favehotel terjadi pada bulan Mei 2017, kebanyakan dari mereka yang menginap karena ada urusan bisnis. Sedangkan pada Hotel Gunawangsa, pengunjung yang memberikan *review*

terbanyak terjadi pada bulan Juli 2017. Faktor yang mempengaruhi banyaknya pengunjung karena adanya musim liburan sekolah.

Berdasarkan data yang terkumpul kemudian diseleksi dan *review* yang mengandung sentimen akan dilakukan analisis lebih lanjut. Berikut adalah perbandingan jumlah *review* yang mengandung sentimen positif dan negatif pada kedua hotel menggunakan klasifikasi berdasarkan *lexicon based*.



Gambar 4.2 Persentase Sentimen Positif dan Negatif

Berdasarkan Gambar 4.2 menunjukkan bahwa data yang digunakan terdiri dari sentimen negatif dan positif. Persentase sentimen positif pada kedua hotel lebih banyak daripada sentimen negatif. Berdasarkan klasifikasi menggunakan *lexicon* maka persentase sentimen positif pada Favehotel sebesar 65% dan sentimen negatif sebesar 35% sedangkan pada Hotel Gunawangsa MERR masing-masing sebesar 58% untuk sentimen positif dan sentimen negatif sebesar 42%.

4.2 Praproses Teks

Data *review* para pengunjung Hotel Gunawangsa MERR dan Favehotel Rungkut pada situs *Traveloka* yang telah terkumpul dari hasil *scraping* kemudian dilakukan filter dengan mengambil

review yang mengandung sentimen. Data yang mengandung sentimen kemudian dilakukan praproses teks yang meliputi *case folding*, *cleansing*, *stemming*, *stopwords* dan *tokenizing*. Berikut adalah hasil dari praproses teks yang ditampilkan pada Tabel 4.1.

Tabel 4.1 Tahapan Praproses Teks

Tahapan	**Hasil Praproses**
Data *review*	Senang sekali menginap 2 malam di favehotel rungkut Surabaya, beruntung sekali saya mendapatkan kamar atas dengan view malam hari sungguh menakjubkan.
Case folding	senang sekali menginap malam di favehotel rungkut surabaya, beruntung sekali saya mendapatkan kamar atas dengan view malam hari sungguh menakjubkan
Cleansing	senang sekali menginap malam di favehotel rungkut surabaya beruntung sekali saya mendapatkan kamar atas dengan view malam hari sungguh menakjubkan
Stemming	senang sekali inap malam di favehotel rungkut surabaya, untung sekali saya dapat kamar atas dengan view malam hari sungguh takjub
Stopwords dan *Tokenizing*	[‘senang’,’inap’,’malam’,‘untung’,’view’,’sungguh’, ’takjub’]

Hasil dari proses *stopword* dan *tokenizing* dipakai sebagai kata kunci dari data *review* yang akan dianalisis dengan menggunakan metode NBC dan RLB. Berikut merupakan hasil perhitungan frekuensi tiap kata kunci pada data *review* Hotel Gunawangsa MERR dan Favehotel Rungkut.

Tabel 4.2 *Count Vectorizer* pada Data *Review*

Review	Variabel Prediktor							
ke-	acara	bandara	cepat	cocok	…	kedap	lokasi	senang
1	0	0	0	0	…	0	1	1
2	0	0	0	1	…	1	0	0
3	0	1	0	1	…	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	0	0	0	0	…	0	0	0

Tabel 4.2 menunjukkan perhitungan frekuensi pada kata kunci dari setiap hotel. Data *review* pada Favehotel memiliki jumlah kata kunci sebanyak 428 kata sedangkan data *review* Hotel Gunawangsa memiliki jumlah kata kunci sebanyak 523 kata. Setelah dilakukan praproses pada data teks maka akan ditampilkan 10 kata kunci dengan frekuensi kemunculan tertinggi pada data *review* Favehotel dan Hotel Gunawangsa pada Tabel 4.3.

Tabel 4.3 Frekuensi Kata Kunci untuk Masing-masing Hotel

Favehotel		**Gunawangsa**	
Kata Kunci	**Frekuensi**	**Kata Kunci**	**Frekuensi**
Bersih	555	Bersih	646
Servis	458	Nyaman	543
Nyaman	411	Servis	495
Bagus	279	Kolam	396
Transmart	273	Bagus	378
Ramah	194	Renang	321
Lokasi	201	Breakfast	250
Puas	193	Inap	233
Panas	187	Toilet	221
Inap	174	Ramah	216

4.3 Visualisasi *Word Cloud*

Visualisasi dengan *Word Cloud* digunakan untuk mengetahui variabel prediktor (kata) yang sering muncul pada data *review*. Data yang digunakan pada *Word Cloud* adalah data *review* yang telah dibedakan berdasarkan sentimennya yaitu positif atau negatif sehingga dapat diketahui kata-kata yang sering muncul pada setiap sentimen tentang masing-masing hotel. Ukuran *font* pada *word cloud* menunjukkan frekuensi kemunculan kata. Jadi, semakin besar ukuran *font* berarti semakin besar pula frekuensi kemunculan kata tersebut. Hasil visualisasi dengan *word cloud* untuk masing-masing kategori sentimen pada Favehotel dan Hotel Gunawangsa ditunjukkan pada Gambar 4.3 dan Gambar 4.4.

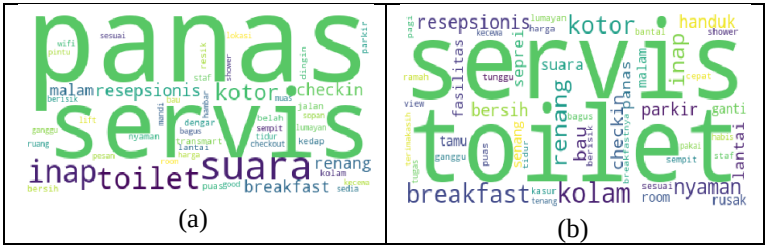
Berdasarkan Gambar 4.3 maka kata kunci terbanyak yang mengarah pada Favehotel dengan sentimen positif terbesar yaitu kata ‘bersih’ dan ‘nyaman’. Pengunjung Favehotel menganggap bahwa kamar yang ada di Favehotel telah bersih dan suasana dalam

kamar hotel yang nyaman. Sama seperti Favehotel, pada Hotel Gunawangsa sentimen positif terbesar adalah kata ‘bersih’ dan ‘nyaman’.

(a) (b)

Gambar 4.3 Visualisasi *Word Cloud* Sentimen Positif
(a) Favehotel (b) Gunawangsa

Sedangkan pada Gambar 4.4 menunjukkan bahwa kata kunci terbanyak yang mengarah pada Favehotel dengan sentimen negatif terbesar yaitu kata ‘panas’ dan ‘servis’. Pengunjung Favehotel menganggap bahwa *Air Conditioner* (AC) yang ada di kamar tidak berfungsi dengan maksimal sehingga penghuni didalamnya merasa kepanasan dan beberapa servis atau pelayanan yang diberikan masih kurang memuaskan. Sedangkan pada Gunawangsa sentimen negatif terbesar adalah kata ‘servis’ dan ‘toilet’. Hal ini karena servis yang diberikan masih kurang maksimal dan keadaan di dalam toilet dalam kamar yang cenderung kotor.



(a) (b)

Gambar 4.4 Visualisasi *Word Cloud* Sentimen Negatif
(a) Favehotel (b) Gunawangsa

4.4 Metode Klasifikasi *Naïve Bayes Classifier* (NBC)

Metode NBC merupakan metode pengklasifikasian statistik yang dapat digunakan untuk memprediksi peluang keanggotaan suatu kelas. Proses ini akan dilakukan pada masing-masing hotel dengan data *review* yang telah dilakukan pengkategorian sentimen. Selanjutnya data akan dibagi menjadi *training* dan *testing.* Pemilihan model terbaik didasarkan pada rata-rata *Area Under Curve* (AUC), *accuracy, precision,* dan *recall.* Rata-rata diperoleh dari masing-masing *subset fold* dengan metode pembagian data *cross validation* 10 *fold* dengan perbandingan data *training* dan *testing* sebesar 90:10.

4.4.1 Metode NBC dengan Data Awal

Berdasarkan metode pembagian data *training* dan *testing* dengan metode *cross validation* maka data *review* pada Favehotel sebanyak 1776 dan Hotel Gunawangsa sebanyak 2172 maka masing-masing dibagi menjadi 90% data *training* dan 10% data *testing* dengan 10 macam kombinasi (*subset fold*). Jumlah keseluruhan data sentimen pada Favehotel dan Hotel Gunawangsa ditunjukkan pada Tabel 4.4.

Tabel 4.4 Jumlah Data Sentimen

Data	*Training*		*Testing*	
	Positif	**Negatif**	**Positif**	**Negatif**
Favehotel	1038	561	115	62
Gunawangsa	1143	811	127	91

Tabel 4.4 menunjukkan perbandingan jumlah data *training* dan *testing.* Selanjutnya dilakukan pemilihan *subset* terbaik dengan metode NBC menggunakan data awal pada *testing* berdasarkan nilai AUC yang tertinggi yang ditunjukkan pada Tabel 4.5.

Berdasarkan Tabel 4.5 pada *subset* ke-4 mempunyai performa klasifikasi yang terbaik pada Favehotel yaitu nilai AUC sebesar 0,88 sedangkan pada Hotel Gunawangsa mempunyai performa klasifikasi pada *subset* ke-1 dengan nilai AUC sebesar 0,84. Sehingga dengan kombinasi *training* dan *testing* data pada *subset* ini akan dilakukan pemodelan.

Tabel 4.5 Pemilihan *Subset* Terbaik NBC Data Awal Berdasarkan AUC

Subset 10-Fold	**Favehotel**	**Gunawangsa**
Subset ke-1	0,86	0,84
Subset ke-2	0,79	0,78
Subset ke-3	0,80	0,79
Subset ke-4	0,88	0,81
Subset ke-5	0,83	0,80
Subset ke-6	0,75	0,78
Subset ke-7	0,81	0,76
Subset ke-8	0,82	0,81
Subset ke-9	0,77	0,77
Subset ke-10	0,80	0,72
Rata-rata	**0,81**	**0,79**

Model klasifikasi yang didapatkan dari data *training* digunakan untuk menghitung nilai probabilitas pada tiap kategori sentimen. Berdasarkan pemilihan *subset* terbaik pada masing-masing hotel maka model klasifikasi metode NBC menggunakan data awal disajikan pada Tabel 4.6.

Tabel 4.6 Model Klasifikasi NBC Data Awal

Hotel	**Klasifikasi**	**Model**
Favehotel	Positif	$0,65 \times 0,00107^{(f1)} \times \ldots \times 0,00177^{(f428)}$
	Negatif	$0,35 \times 0,00571^{(f1)} \times \ldots \times 0,00752^{(f428)}$
Gunawangsa	Positif	$0,59 \times 0,00052^{(f1)} \times \ldots \times 0,00129^{(f523)}$
	Negatif	$0,41 \times 0,00084^{(f1)} \times \ldots \times 0,00258^{(f523)}$

Berdasarkan model klasifikasi pada Tabel 4.6 maka selanjutnya dilakukan pengklasifikasian dengan memasukkan frekuensi kata pada data baru disetiap masing-masing klasifikasi positif dan negatif. Sebagai contoh penerapannya seperti pada Tabel 4.7.

Berdasarkan Tabel 4.7 dapat diketahui peluang positif dan negatif sehingga suatu *review* masuk pada klasifikasi positif atau negatif. Data *testing* pertama pada Favehotel memiliki peluang klasifikasi positif sebesar $1,29 \times 10^{-5}$ sedangkan peluang untuk masuk pada klasifikasi negatif sebesar $1,02 \times 10^{-5}$ sehingga dapat

diputuskan bahwa data *review* pertama masuk dalam klasifikasi positif. Begitu juga data *testing* pertama pada Gunawangsa memiliki peluang klasifikasi positif sebesar $1{,}67x10^{-5}$ sedangkan peluang untuk masuk pada klasifikasi negatif sebesar $2{,}90x10^{-5}$ sehingga dapat diputuskan bahwa data *review* pertama masuk dalam klasifikasi negatif.

Tabel 4.7 Probabilitas Klasifikasi NBC Data Awal

Hotel	*Review*	Probabilitas Positif	Probabilitas Negatif	Keputusan
Favehotel	1	$1,29\times10^{-5}$	$1,02\times10^{-5}$	Positif
	2	$4,69\times10^{-8}$	$1,99\times10^{-7}$	Negatif
	...	...	...	...
	177	$7,54\times10^{-4}$	$1,18\times10^{-4}$	Positif
Gunawangsa	1	$1,67\times10^{-5}$	$2,90\times10^{-5}$	Negatif
	2	$1,70\times10^{-6}$	$5,39\times10^{-6}$	Negatif
	...	...	...	...
	218	$1,22\times10^{-9}$	$1,57\times10^{-9}$	Negatif

Selanjutnya yang dihasilkan adalah *confusion matrix* untuk melakukan perhitungan ketepatan klasifikasi yang ditampilkan sebagai berikut.

Tabel 4.8 *Confusion Matrix* Favehotel dengan Metode NBC Data Awal

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	51	11
Negatif	8	107

Tabel 4.8 mengenai hasil *confusion matrix* untuk data Favehotel dapat diketahui bahwa ketepatan klasifikasi untuk data negatif yang diklasifikasikan benar negatif sebesar 107, sedangkan untuk data negatif yang diklasifikasikan positif sebesar 8. Pada data positif yang diklasifikasikan negatif sebesar 11 sedangkan untuk data positif yang diklasifikasikan benar positif sebesar 51. Sedangkan hasil *confusion matrix* untuk data Hotel Gunawangsa ditunjukkan pada Tabel 4.9.

Tabel 4.9 *Confusion Matrix* Gunawangsa dengan Metode NBC Data Awal

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	83	8
Negatif	30	97

Tabel 4.9 menunjukkan *confusion matrix* pada data *testing* yang selanjutnya dilakukan perhitungan ketepatan klasifikasi untuk data negatif yang diklasifikasikan benar negatif sebesar 97, sedangkan untuk data negatif yang diklasifikasikan positif sebesar 30. Pada data positif yang diklasifikasikan negatif sebesar 8 sedangkan untuk data positif yang diklasifikasikan benar positif sebesar 83.

4.4.2 Metode NBC dengan Data SMOTE

Pada sub bab karakteristik data dapat diketahui bahwa jumlah *review* positif pada Favehotel dan Hotel Gunawangsa memiliki persentase lebih banyak daripada persentase sentimen negatif. Kondisi tersebut menunjukkan bahwa data *imbalance.* Maka dari itu, data *review* akan dilakukan *oversampling* dengan *Syntetic Minority Oversampling Technique* (SMOTE), dimana jumlah data yang mengandung sentimen negatif akan disamakan dengan jumlah data yang mengandung sentimen positif. Berikut ditunjukkan perbandingan persentase sentimen positif dan negatif sebelum dan sesudah *oversampling.*

Tabel 4.10 Jumlah Data Sentimen (Y) Data *Training*

Data	**Awal**		**SMOTE**	
	Positif	**Negatif**	**Positif**	**Negatif**
Favehotel	1038	561	1038	1038
Gunawangsa	1143	811	1143	1143

Berdasarkan Tabel 4.10 didapatkan perbandingan jumlah data sentimen positif dan negatif pada data *training* dengan SMOTE adalah 50:50, sehingga dapat dikatakan bahwa data telah *balance* untuk dilakukan analisis menggunakan metode NBC dan RLB. Selanjutnya dilakukan pemilihan *subset* terbaik dengan metode NBC menggunakan data yang telah dilakukan *oversampling* SMOTE yang ditunjukkan pada Tabel 4.11.

Tabel 4.11 Pemilihan *Subset* Terbaik NBC Data SMOTE Berdasarkan AUC

Subset 10-Fold	**Favehotel**	**Gunawangsa**
Subset ke-1	0,87	0,85
Subset ke-2	0,79	0,81
Subset ke-3	0,82	0,80
Subset ke-4	0,90	0,81
Subset ke-5	0,85	0,81
Subset ke-6	0,78	0,78
Subset ke-7	0,84	0,76
Subset ke-8	0,82	0,82
Subset ke-9	0,79	0,79
Subset ke-10	0,80	0,74
Rata-rata	**0,83**	**0,80**

Berdasarkan Tabel 4.11 pada *subset* ke-4 mempunyai performa klasifikasi yang terbaik pada Favehotel yaitu nilai AUC sebesar 0,90 sedangkan pada Hotel Gunawangsa performa klasifikasi terbaik terdapat pada *subset* ke-1 sebesar 0,85. Model klasifikasi yang didapatkan digunakan untuk menghitung nilai probabilitas pada tiap kategori sentimen. Berdasarkan pemilihan *subset* terbaik pada masing-masing hotel maka model klasifikasi NBC menggunakan data SMOTE disajikan pada Tabel 4.12.

Tabel 4.12 Model Klasifikasi NBC Data SMOTE

Hotel	**Klasifikasi**	**Model**
Favehotel	Positif	$0,5 \times 0,00107^{(f1)} \times \ldots \times 0,00177^{(f428)}$
	Negatif	$0,5 \times 0,00574^{(f1)} \times \ldots \times 0,00771^{(f428)}$
Gunawangsa	Positif	$0,5 \times 0,00052^{(f1)} \times \ldots \times 0,00129^{(f523)}$
	Negatif	$0,5 \times 0,00082^{(f1)} \times \ldots \times 0,00237^{(f523)}$

Berdasarkan model klasifikasi pada Tabel 4.12 maka selanjutnya melakukan pengklasifikasian dengan memasukkan frekuensi kata pada data baru disetiap masing-masing klasifikasi positif dan negatif. Contoh penerapannya seperti pada Tabel 4.13.

Berdasarkan Tabel 4.13 dapat diketahui peluang positif dan negatif sehingga suatu *review* masuk pada klasifikasi positif atau negatif. Data *testing* pertama pada Favehotel memiliki peluang

klasifikasi positif sebesar 1,00x10^{-5} sedangkan peluang untuk masuk pada klasifikasi negatif sebesar 1,18x10^{-5} sehingga dapat diputuskan bahwa data *review* pertama masuk dalam klasifikasi negatif. Begitu juga pada data *testing* pertama pada Gunawangsa memiliki peluang klasifikasi positif sebesar 1,43x10^{-5} sedangkan peluang untuk masuk pada klasifikasi negatif sebesar 3,18x10^{-5} sehingga dapat diputuskan bahwa data *review* pertama masuk dalam klasifikasi negatif.

Tabel 4.13 Probabilitas Klasifikasi NBC Data SMOTE

Hotel	*Review*	Probabilitas Positif	Probabilitas Negatif	Keputusan
Favehotel	1	$1,00 \times 10^{-5}$	$1,18 \times 10^{-5}$	Negatif
	2	$3,61 \times 10^{-8}$	$2,88 \times 10^{-7}$	Negatif
	...	...	...	...
	177	$5,81 \times 10^{-4}$	$2,18 \times 10^{-4}$	Positif
Gunawangsa	1	$1,43 \times 10^{-5}$	$3,18 \times 10^{-5}$	Negatif
	2	$1,45 \times 10^{-6}$	$6,50 \times 10^{-6}$	Negatif
	...	...	...	...
	218	$1,04 \times 10^{-9}$	$2,18 \times 10^{-9}$	Negatif

Selanjutnya yang dihasilkan adalah *confusion matrix* untuk melakukan perhitungan ketepatan klasifikasi pada Favehotel yang ditampilkan pada Tabel 4.14.

Tabel 4.14 *Confusion Matrix* Favehotel dengan Metode NBC Data SMOTE

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	55	7
Negatif	9	106

Berdasarkan Tabel 4.14 dapat diketahui bahwa ketepatan klasifikasi untuk data positif yang diklasifikasikan benar positif sebesar 55, sedangkan untuk data positif yang diklasifikasikan negatif sebesar 7. Pada data negatif yang diklasifikasikan positif sebesar 9 sedangkan untuk data negatif yang diklasifikasikan benar

negatif sebesar 106. Sedangkan hasil *confusion matrix* untuk data Hotel Gunawangsa ditunjukkan pada Tabel 4.15.

Tabel 4.15 *Confusion Matrix* Gunawangsa dengan Metode NBC Data SMOTE

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	85	6
Negatif	30	97

Tabel 4.15 menunjukkan *confusion matrix* pada data *testing* yang selanjutnya dilakukan perhitungan ketepatan klasifikasi untuk data positif yang diklasifikasikan benar positif sebesar 85, sedangkan untuk data positif yang diklasifikasikan negatif sebesar 6. Pada data negatif yang diklasifikasikan positif sebesar 30 sedangkan untuk data negatif yang diklasifikasikan benar negatif sebesar 97.

Selanjutnya akan dilakukan perbandingan antara performa klasifikasi menggunakan data awal dan data yang telah dilakukan SMOTE pada Favehotel dan Hotel Gunawangsa dengan metode NBC. Perbandingan performa klasifikasi dilakukan dengan membandingkan rata-rata performa klasifikasi setiap *subset fold*, sehingga diperoleh dengan hasil ringkasan pada Tabel 4.16.

Tabel 4.16 Perbandingan Performa Klasifikasi NBC Data Awal dan SMOTE

	Data	**Hotel**	***Accuracy***	***Precision***	***Recall***	**AUC**
Training	Awal	Favehotel	0,86	0,86	0,75	0,83
		Gunawangsa	0,83	0,87	0,70	0,82
	SMOTE	Favehotel	0,87	0,91	0,83	0,87
		Gunawangsa	0,83	0,90	0,74	0,83
Testing	Awal	Favehotel	0,84	0,82	0,71	0,81
		Gunawangsa	0,81	0,84	0,67	0,79
	SMOTE	Favehotel	0,85	0,82	0,75	0,83
		Gunawangsa	0,81	0,84	0,70	0,80

Berdasarkan Tabel 4.16 dapat diketahui bahwa data hasil *oversampling* SMOTE pada *training* dan *testing* di kedua hotel menghasilkan rata-rata performa klasifikasi lebih baik daripada data awal.

4.5 Metode Klasifikasi Regresi Logistik Biner

Metode klasifikasi Regresi Logistik Biner (RLB) pada penelitian ini akan dibandingkan dengan metode klasifikasi NBC. Data yang digunakan pada penelitian ini adalah data *review* Favehotel dan Hotel Gunawangsa yang sebelumnya telah dilakukan pengkategorian sentimen. Selanjutnya data akan dibagi menjadi *training* dan *testing.* Data yang mengandung sentimen positif pada kedua hotel lebih banyak daripada sentimen negatifnya. Sehingga dilakukan *balancing* data menggunakan *oversampling* SMOTE. Berikut ditampilkan perhitungan dengan Data Awal dan data hasil SMOTE sehingga diperoleh model dan performa klasifikasi menggunakan metode Regresi Logistik Biner.

4.5.1 Metode RLB Data Awal

Berdasarkan metode pembagian data *training* dan *testing* dengan metode *cross validation* maka data *review* pada Favehotel sebanyak 1776 dan Gunawangsa sebanyak 2172 maka masing-masing dibagi menjadi 90% data *training* dan 10% data *testing* dengan 10 macam kombinasi (*subset fold*). Kemudian akan dilakukan pendeteksian apakah terjadi kasus multikolinieritas atau tidak pada data yang hasilnya ditampilkan pada Tabel 4.17.

Tabel 4.17 Pendeteksian Multikolinieritas (Favehotel)

Variabel	VIF	Keterangan	Variabel	VIF	Keterangan
acnya	4,236	Tidak Terjadi Multiko	internetnya	17,904	Multiko
agam	2,178	Tidak Terjadi Multiko	istimewa	162,728	Multiko
airport	1,000	Tidak Terjadi Multiko	istirahat	3,522	Tidak Terjadi Multiko
akses	1,773	Tidak Terjadi Multiko	istri	1,301	Tidak Terjadi Multiko
⋮	⋮	⋮	⋮	⋮	⋮
industri	1,791	Tidak Terjadi Multiko	wangi	24,390	Multiko
info	8,973	Tidak Terjadi Multiko	wastafel	1,000	Tidak Terjadi Multiko
interior	1,037	Tidak Terjadi Multiko	welcome	8,467	Tidak Terjadi Multiko
internet	1,234	Tidak Terjadi Multiko	wifi	7,055	Tidak Terjadi Multiko

Berdasarkan Tabel 4.17 dapat diketahui bahwa pada variabel internetnya, istimewa, dan wangi diindikasikan terjadi kasus multikolinieritas karena nilai VIF yang lebih dari 10. Selanjutnya untuk variabel yang terjadi multikolinieritas tidak dilanjutkan ke dalam analisis selanjutnya. Berikut ditampilkan pendeteksian

multikolinieritas pada data *review* Hotel Gunawangsa pada Tabel 4.18.

Tabel 4.18 Pendeteksian Multikolinieritas (Gunawangsa)

Variabel	VIF	Keterangan	Variabel	VIF	Keterangan
abis	1,950	Tidak Terjadi Multiko	jarang	1,552	Tidak Terjadi Multiko
acara	1,364	Tidak Terjadi Multiko	jaring	7,096	Tidak Terjadi Multiko
acnya	1,752	Tidak Terjadi Multiko	jatuh	1,000	Tidak Terjadi Multiko
aerobik	1,000	Tidak Terjadi Multiko	jelek	1,386	Tidak Terjadi Multiko
⋮	⋮	⋮	⋮	⋮	⋮
informasi	1,917	Tidak Terjadi Multiko	warna	1,791	Tidak Terjadi Multiko
istimewa	49,453	Multiko	wastafel	1,000	Tidak Terjadi Multiko
istirahat	1,289	Tidak Terjadi Multiko	wifi	1,430	Tidak Terjadi Multiko
jaga	2,073	Tidak Terjadi Multiko	wifinya	1,797	Tidak Terjadi Multiko

Tabel 4.18 manunjukkan bahwa variabel istimewa terjadi kasus multikolinieritas karena nilai VIF yang lebih dari 10. Selanjutnya untuk variabel yang terjadi multikolinieritas tidak dilanjutkan ke dalam analisis selanjutnya.

Setelah dilakukan pendeteksian multikolinieritas pada Favehotel dan Gunawangsa maka variabel yang digunakan pada pengujian signifikasnsi parameter adalah variabel yang tidak terjadi kasus multikolinieritas. Selanjutnya akan dilakukan pengujian serentak dengan hipotesis sebagai berikut.

H_0 : $\beta_1 = \beta_2 = \ldots = \beta_m = 0$

H_1 : minimal ada satu $\beta_j \neq 0$, j = 1,2,...,m

Berdasarkan hasil uji serentak pada data awal Favehotel yang dilampirkan pada Lampiran 10A menghasilkan keputusan Tolak H_0 karena nilai statistik uji G^2 sebesar 1800,53 lebih besar daripada $\chi^2_{(0,05;416)}$ sebesar 464,554. Sehingga dapat disimpulkan bahwa minimal terdapat satu variabel prediktor yang berpengaruh signifikan terhadap variabel respon.

Sedangkan hasil pengujian serentak pada data awal Hotel Gunawangsa yang dilampirkan pada Lampiran 11A menghasilkan keputusan Tolak H_0 karena nilai statistik uji G^2 sebesar 1493,4 lebih besar daripada $\chi^2_{(0,05;462)}$ sebesar 513,110. Sehingga dapat disimpulkan bahwa minimal terdapat satu variabel prediktor yang berpengaruh signifikan terhadap variabel respon.

Pengujian parsial dilakukan untuk mengetahui variabel prediktor mana yang memiliki pengaruh signifikan terhadap model dengan hipotesis sebagai berikut.

$H_0 : \beta_j = 0, j = 1,2,...,m$

$H_1 : \beta_j \neq 0, j = 1,2,...,m$

Hasil pengujian parsial menggunakan data awal pada Favehotel yang disajikan pada Lampiran 10A dan Gunawangsa yang disajikan pada Lampiran 11A menghasilkan keputusan Tolak H_0 karena nilai statistik uji *Wald* lebih besar daripada $\chi^2_{(0,05;1)}$ sebesar 3,841 sehingga dapat disimpulkan bahwa variabel prediktor tersebut berpengaruh signifikan terhadap model. Berdasarkan hasil estimasi parameter maka model regresi logistik binernya sebagai berikut.

$$\hat{\pi}(x) = \frac{\exp(0,38 - 3,99X_{21} - 7,46X_{33} + \ldots - 5,80X_{423})}{1 + \exp(0,38 - 3,99X_{21} - 7,46X_{33} + \ldots - 5,80X_{423})} \quad (4.1)$$

$$\hat{\pi}(x) = \frac{\exp(-0,02 - 2,55X_{29} - 3,51X_{41} + \ldots - 3,84X_{500})}{1 + \exp(-0,02 - 2,55X_{29} - 3,51X_{41} + \ldots - 3,84X_{500})} \quad (4.2)$$

Setelah mendapatkan model regresi logistik pada persamaan (4.1) dan (4.2) maka selanjutnya menghitung performa klasifikasi untuk data *testing* pada data awal *review* Favehotel dan Hotel Gunawangsa yang ditunjukkan pada Tabel 4.19.

Tabel 4.19 Performa Klasifikasi RLB Data Awal Berdasarkan AUC

Subset 10-Fold	**Favehotel**	**Gunawangsa**
Subset ke-1	0,88	0,86
Subset ke-2	0,79	0,81
Subset ke-3	0,82	0,81
Subset ke-4	0,80	0,82
Subset ke-5	0,86	0,84
Subset ke-6	0,81	0,76
Subset ke-7	0,84	0,81
Subset ke-8	0,79	0,84
Subset ke-9	0,80	0,78
Subset ke-10	0,78	0,79
Rata-rata	**0,82**	**0,81**

Berdasarkan Tabel 4.19 menunjukkan bahwa pada *subset* ke-1 mempunyai performa klasifikasi yang terbaik pada Favehotel yaitu nilai AUC sebesar 0,88 sedangkan nilai AUC Hotel Gunawangsa terbaik terdapat pada pada *subset* ke-1 sebesar 0,86. Hasil *confusion matrix* untuk Favehotel ditunjukkan pada Tabel 4.20.

Tabel 4.20 *Confusion Matrix* Favehotel dengan Metode RLB Data Awal

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	53	10
Negatif	9	107

Berdasarkan Tabel 4.20 dapat diketahui bahwa ketepatan klasifikasi untuk data negatif yang diklasifikasikan benar negatif sebesar 107, sedangkan untuk data negatif yang diklasifikasikan positif sebesar 9. Pada data positif yang diklasifikasikan negatif sebesar 10 sedangkan untuk data positif yang diklasifikasikan benar positif sebesar 53. Hasil *confusion matrix* untuk Gunawangsa ditunjukkan pada Tabel 4.21.

Tabel 4.21 *Confusion Matrix* Gunawangsa dengan Metode RLB Data Awal

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	85	10
Negatif	22	101

Tabel 4.21 menunjukkan *confusion matrix* pada data *testing* yang selanjutnya dilakukan perhitungan ketepatan klasifikasi untuk data negatif yang diklasifikasikan benar negatif sebesar 101, sedangkan untuk data negatif yang diklasifikasikan positif sebesar 22. Pada data positif yang diklasifikasikan negatif sebesar 10 sedangkan untuk data positif yang diklasifikasikan benar positif sebesar 85.

4.5.2 Metode RLB Data SMOTE

Pada sub bab karakteristik data dapat diketahui bahwa jumlah *review* yang mengandung sentimen positif lebih banyak daripada negatif sehingga menunjukkan bahwa data *imbalance.* Maka dari itu, data *review* akan dilakukan *oversampling* dengan

Syntetic Minority Oversampling Technique (SMOTE), dimana jumlah data yang mengandung sentimen negatif akan disamakan dengan jumlah data yang mengandung sentimen positif. Setelah dilakukan *oversampling* dengan SMOTE kemudian dilakukan pendeteksian apakah terjadi kasus multikolinieritas atau tidak.

Berdasarkan pendeteksian multikolinieritas pada Tabel 4.17 dan Tabel 4.18 dapat diketahui bahwa variabel dengan nilai VIF yang lebih dari 10 diindikasikan terjadi kasus multikolinieritas sehingga variabel yang digunakan pada pengujian signifikansi parameter adalah variabel yang tidak terjadi kasus multikolinieritas. Selanjutnya akan dilakukan pengujian serentak dengan hipotesis sebagai berikut.

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_m = 0$

H_1 : minimal ada satu $\beta_j \neq 0$, $j = 1,2,...,m$

Berdasarkan hasil uji serentak pada Favehotel data SMOTE yang dilampirkan pada Lampiran 10B menghasilkan keputusan Tolak H_0 karena nilai statistik uji G^2 sebesar 1460,97 lebih besar daripada $\chi^2_{(0,05;323)}$ sebesar 365,912. Sehingga dapat disimpulkan bahwa minimal terdapat satu variabel prediktor yang berpengaruh signifikan terhadap variabel respon.

Sedangkan hasil pengujian serentak pada Hotel Gunawangsa data SMOTE yang dilampirkan pada Lampiran 11B menghasilkan keputusan Tolak H_0 karena nilai statistik uji G^2 sebesar 1780,8 lebih besar daripada $\chi^2_{(0,05;462)}$ sebesar 513,110. Sehingga dapat disimpulkan bahwa minimal terdapat satu variabel prediktor yang berpengaruh signifikan terhadap variabel respon.

Berdasarkan hasil pengujian serentak didapatkan kesimpulan bahwa minimal ada satu variabel prediktor yang berpengaruh signifikan terhadap variabel respon sehingga dapat dilanjutkan uji parsial. Pengujian parsial dilakukan untuk mengetahui variabel prediktor mana yang memiliki pengaruh signifikan terhadap model dengan hipotesis sebagai berikut.

$H_0 : \beta_j = 0$, $j = 1,2,...,m$

$H_1 : \beta_j \neq 0, j = 1,2,...,m$

Hasil pengujian parsial menggunakan data SMOTE pada Favehotel yang disajikan pada Lampiran 10B dan Gunawangsa yang disajikan pada Lampiran 11B menghasilkan keputusan Tolak H_0 karena nilai statistik uji *Wald* lebih besar daripada $\chi^2_{(0,05;1)}$ sebesar 3,841 sehingga dapat disimpulkan bahwa variabel prediktor tersebut berpengaruh signifikan terhadap model. Berdasarkan hasil estimasi parameter maka model regresi logistik binernya sebagai berikut.

$$\hat{\pi}(x) = \frac{\exp(-0,84 + 2,77X_{21} - 3,18X_{24} + \ldots + 3,47X_{423})}{1 + \exp(-0,84 + 2,77X_{21} - 3,18X_{24} + \ldots + 3,47X_{423})} \quad (4.3)$$

$$\hat{\pi}(x) = \frac{\exp(-0,86 - 5,66X_{12} + 2,97X_{29} + \ldots - 2,54X_{500})}{1 + \exp(-0,86 - 5,66X_{12} + 2,97X_{29} + \ldots - 2,54X_{500})} \quad (4.4)$$

Setelah mendapatkan model regresi logistik pada persamaan (4.3) dan (4.4) maka selanjutnya menghitung performa klasifikasi untuk data *testing* pada data SMOTE untuk Favehotel dan Hotel Gunawangsa yang ditunjukkan pada Tabel 4.22.

Tabel 4.22 Pemilihan *Subset* Terbaik RLB Data SMOTE Berdasarkan AUC

Subset 10-Fold	**Favehotel**	**Gunawangsa**
Subset ke-1	0,86	0,84
Subset ke-2	0,84	0,82
Subset ke-3	0,83	0,82
Subset ke-4	0,84	0,82
Subset ke-5	0,82	0,81
Subset ke-6	0,83	0,78
Subset ke-7	0,85	0,81
Subset ke-8	0,84	0,90
Subset ke-9	0,86	0,79
Subset ke-10	0,80	0,78
Rata-rata	**0,84**	**0,82**

Hasil klasifikasi metode RLB menggunakan data yang telah dilakukan *oversampling* SMOTE pada Tabel 4.22 menunjukkan bahwa pada *subset* ke-1 mempunyai performa klasifikasi yang terbaik pada Favehotel yaitu nilai AUC sebesar 0,86 sedangkan

pada Hotel Gunawangsa nilai AUC terbaiknya sebesar 0,90 berada pada *subset* ke-8. Hasil *confusion matrix* untuk Favehotel ditunjukkan pada Tabel 4.23.

Tabel 4.23 *Confusion Matrix* Favehotel dengan Metode RLB Data SMOTE

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	56	7
Negatif	19	97

Berdasarkan Tabel 4.23 dapat diketahui bahwa ketepatan klasifikasi untuk data positif yang diklasifikasikan benar positif sebesar 56, sedangkan untuk data positif yang diklasifikasikan negatif sebesar 7. Pada data negatif yang diklasifikasikan positif sebesar 19 sedangkan untuk data negatif yang diklasifikasikan benar negatif sebesar 97. Sedangkan hasil *confusion matrix* untuk data Hotel Gunawangsa sebagai berikut.

Tabel 4.24 *Confusion Matrix* Gunawangsa dengan Metode RLB Data SMOTE

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	69	8
Negatif	13	127

Tabel 4.24 menunjukkan *confusion matrix* pada data *testing* yang selanjutnya dilakukan perhitungan ketepatan klasifikasi untuk data positif yang diklasifikasikan benar positif sebesar 69, sedangkan untuk data positif yang diklasifikasikan negatif sebesar 8. Pada data negatif yang diklasifikasikan positif sebesar 13 sedangkan untuk data negatif yang diklasifikasikan benar negatif sebesar 127.

Selanjutnya akan dilakukan perbandingan antara performa klasifikasi menggunakan data awal dan data yang telah dilakukan SMOTE pada Favehotel dan Hotel Gunawangsa dengan metode RLB. Perbandingan performa klasifikasi dilakukan dengan membandingkan rata-rata performa klasifikasi setiap *subset fold*, sehingga diperoleh dengan hasil ringkasan pada Tabel 4.25.

Tabel 4.25 Perbandingan Performa Klasifikasi RLB Data Awal dan SMOTE

	Data	Hotel	*Accuracy*	*Precision*	*Recall*	AUC
Training	Awal	Favehotel	0,89	0,88	0,79	0,86
		Gunawangsa	0,87	0,84	0,85	0,87
	SMOTE	Favehotel	0,89	0,87	0,92	**0,89**
		Gunawangsa	0,88	0,85	0,91	**0,88**
Testing	Awal	Favehotel	0,85	0,82	0,72	0,82
		Gunawangsa	0,82	0,79	0,77	0,81
	SMOTE	Favehotel	0,84	0,73	0,85	**0,84**
		Gunawangsa	0,82	0,74	0,85	**0,82**

Berdasarkan Tabel 4.25 dapat diketahui bahwa data hasil *oversampling* SMOTE pada *training* dan *testing* di kedua hotel menghasilkan rata-rata performa klasifikasi lebih baik daripada data awal.

4.6 Pemilihan Metode Klasifikasi Terbaik

Hasil yang telah diketahui berdasarkan ketepatan klasifikasi pada metode *Naïve Bayes Classifier* (NBC) dan Regresi Logistik Biner (RLB) kemudian dilanjutkan dengan membandingkan antara kedua metode berdasarkan rata-rata performa *accuracy, precision, recall,* dan AUC.

Tabel 4.26 menunjukkan perbandingan metode NBC dan RLB untuk Favehotel Rungkut dan Hotel Gunawangsa MERR yang dapat disimpulkan bahwa metode RLB dengan SMOTE pada *training* dan *testing* lebih baik jika dibandingkan dengan metode NBC. Hal ini dapat diketahui dari hasil AUC yang tinggi daripada metode NBC. Pada Favehotel Rungkut nilai AUC pada data *training* sebesar 0,89 sedangkan pada Hotel Gunawangsa MERR pada data *training* nilai AUC sebesar 0,88. Sedangkan dengan menggunakan data *testing* maka nilai AUC pada Favehotel Rungkut sebesar 0,84 dan Hotel Gunawangsa MERR sebesar 0,82.

Tabel 4.26 Perbandingan Metode NBC dan RLB

Metode *Naïve Bayes Classifier* (NBC)						
	Data	**Hotel**	***Accuracy***	***Precision***	***Recall***	**AUC**
Training	Awal	Favehotel	0,86	0,86	0,75	0,83
		Gunawangsa	0,83	0,87	0,7	0,82
	SMOTE	Favehotel	0,87	0,91	0,83	0,87
		Gunawangsa	0,83	0,9	0,74	0,83
Testing	Awal	Favehotel	0,84	0,82	0,71	0,81
		Gunawangsa	0,81	0,84	0,67	0,79
	SMOTE	Favehotel	0,85	0,82	0,75	0,83
		Gunawangsa	0,81	0,84	0,70	0,80
Metode Regresi Logistik Biner (RLB)						
	Data	**Hotel**	***Accuracy***	***Precision***	***Recall***	**AUC**
Training	Awal	Favehotel	0,89	0,88	0,79	0,86
		Gunawangsa	0,87	0,84	0,85	0,87
	SMOTE	Favehotel	0,89	0,87	0,92	**0,89**
		Gunawangsa	0,88	0,85	0,91	**0,88**
Testing	Awal	Favehotel	0,85	0,82	0,72	0,82
		Gunawangsa	0,82	0,79	0,77	0,81
	SMOTE	Favehotel	0,84	0,73	0,85	**0,84**
		Gunawangsa	0,82	0,74	0,85	**0,82**

(Halaman ini sengaja dikosongkan)

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Hasil analisis yang telah dilakukan dapat diambil kesimpulan sebagai berikut.

1. Klasifikasi menggunakan *lexicon* maka persentase sentimen positif pada Favehotel sebesar 65% dan sentimen negatif sebesar 35% sedangkan pada Hotel Gunawangsa MERR masing-masing sebesar 58% untuk sentimen positif dan sentimen negatif sebesar 42%.
2. Visualisasi *word cloud* menunjukkan bahwa kata kunci terbanyak yang mengarah pada kedua hotel dengan sentimen positif terbesar yaitu kata ‘bersih’ dan ‘nyaman’. Sedangkan kata kunci terbanyak dengan sentimen negatif pada Favehotel adalah ‘panas’ dan ‘servis’ sedangkan pada Gunawangsa adalah ‘servis’ dan ‘toilet’.
3. Perbandingan metode antara *Naïve Bayes Classifier* (NBC) dan Regresi Logistik Biner (RLB) menunjukkan bahwa metode RLB dengan SMOTE lebih baik jika dibandingkan dengan metode NBC yang mana pada data *testing* untuk nilai AUC Favehotel sebesar 0,84 dan AUC Gunawangsa sebesar 0,82.

5.2 Saran

Saran yang dapat diberikan dari hasil penelitian ini adalah meskipun kebanyakan pengunjung telah memberikan *review* positif namun perlu diperhatikan juga saran dan *review* negatif dari para pengunjung. Selain itu, diharapkan menjadikan hasil penelitian ini sebagai informasi tambahan supaya pihak Favehotel dan Hotel Gunawangsa menjadikan permasalahan kamar yang panas karena AC dan beberapa servis yang kurang maksimal agar kedepannya lebih baik lagi. Berdasarkan hasil dari pemodelan menggunakan NBC dan RLB nantinya bisa digunakan untuk menentukan ketika ada *review* baru dimasukkan dalam klasifikasi

review positif atau negatif.

DAFTAR PUSTAKA

APJII. (2018). *Infografis Penetrasi dan Perilaku Pengguna Internet Indonesia* . Jakarta: APJII

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation Measures for Models Assessment over Imbalanced Data Sets. *Journal Information Engineering and Applications, 3*(10), 27-38.

BPS. (2018). *Kota Surabaya Dalam Angka.* Surabaya: Badan Pusat Statistik.

Buntoro, G. A., Adji, T. B., & Purnamasari, A. E. (2014). Sentiment Analysis Twitter dengan Kombinasi Lexicon Based dan Double Propagation. *The 6th Conference on Information Technology and Electrical Engineering* (hal. 39-43). Yogyakarta: Universitas Gadjah Mada.

Chawla, N. V., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

Chawla, N. V. (2005). *Data Mining and Knowledge Discovery Handbook.* USA: University of Notre Dame Press.

Datascienceplus. (2017). *Building A Logistic Regression in Python Step by Step*. Dipetik Maret 12, 2019, dari https://datascienceplus.com/building-a-logistic-regression-in-python-step-by-step/

Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment Analysis of Review Datasets using Naive Bayes and K-NN Classifier. *I.J. Information Engineering and Electronic Business, 8*(4), 54-62.

Favehotel Rungkut. (2019). *Welcome to favehotel Rungkut*. Dipetik Maret 6, 2019, dari https://www.favehotels.com/en/hotel/view/89/favehotel-rungkut

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data.* New York: Cambridge University Press.

Hamzah, A. (2012). Klasifikasi Teks dengan Naive Bayes Classifier (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis. *Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III* (hal. B-269 - B-277). Yogyakarta.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (*2nd Edition*). Canada: John Wiley & Sons, Inc.

Hotel Gunawangsa MERR. (2016). *Welcome*. Dipetik Maret 6, 2019, dari http://gunawangsahotel.com/merr/

Hotho, A., Numberger, A., & Paas, G. (2005). *A Brief Survey of Text Mining.* Kassel: University of Kassel.

Josi, A., Suryayusra, & Abdillah, L. (2014). Penerapan Teknik Web Scraping pada Mesin Pencari Artikel Ilmiah. *Jurnal Sistem Informasi, 5* (2), 159-164.

Kang, M., Ahn, J., & Lee, K. (2018). Opinion Mining Using Ensemble Text Hidden Markov Models For Text Classification. *Expert System with Aplication, 94*, 218-227.

Khempila, A., & Boonjing, V. (2010). Comparing Performance of Logistic Regression, Decision Trees and Neural Networks for Classifying Heart Disease Patients. *International Conference on Computer Information System and Industrial Management Aplications* (hal. 193-198). Poland: IEEE.

Korhonen, A., Seaghdha, D., Silins, I., Sun, L., Hogberg, J., & Stenius, U. (2012). Text Mining for Literature Review and Knowledge Discovery in Cancer Risk Assessment and Research. *PLoS ONE, 7*(4), 1-16.

Kurniasari, S. (2018). *Implementasi SVM dan Asosiasi untuk Sentiment Analysis Data Ulasan The Phoenix Hotel Yogyakarta pada Situs TripAdvisor*. Yogyakarta: Tugas Akhir Universitas Islam Indonesia.

Li, G., & Liu, F. (2010). A Clustering-Based Approach On Sentiment Analysis. *2010 IEEE International Conference*

on Intelligent Systems and Knowledge Engineering (hal. 331-337). China: IEEE.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Chicago: Morgan & Claypool Publishers.

McNaught, C., & Lam, P. (2010). Using Wordle as a Supplementary Research Tool. *The Qualitative Report, 15*(3), 630-643.

Mohammad, S., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada : Building the State-of-the-Art in Sentiment Analysis of Tweets. *Second Joint Conference on Lexical and Computational Semantics Volume 2 : Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (hal. 321-327). Atlanta, Georgia: Association for Computational Linguistics.

Murnawan, & Sinaga, A. (2017). Pemanfaatan Analisis Sentimen untuk Pemeringkatan Popularitas Tujuan Wisata. *Jurnal Penelitian Pos dan Informatika, 7*(2), 109-120.

Pravina, A. M., Cholissodin, I., & Adikara, P. P. (2018). Analisis Sentimen tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 3*(3), 2789-2797.

Putri, D. U. (2016). *Implementasi Inferensi Fuzzy Mamdani untuk Keperluan Sistem Rekomendasi Berita Berbasis Konten*. Yogyakarta : Skripsi Universitas Gadjah Mada.

Putri, N. A., & Alamsyah, A. (2017). Opinion Mining of TripAdivisor Review Towards Five-Star Hotels in Bandung City. *e-Proceding of Management, 4*(1), 572-575.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *Encyclopedia of Database System* (hal. 532-538). United States: Springer.

Siang, J. (2005). *Jaringan Syaraf Tiruan & Pemrogamannya Menggunakan MATLAB.* Yogyakarta: ANDI.

Tala, F. (2003). *A Study of Stemming Effect on Information Retrieval in Bahasa Indonesia.* Amsterdam: Institute for Logic, Language, and Computation, Universiteit van Amsterdam.

Tempo.co. (2018). *Surabaya Dinobatkan Jadi Kota Terbaik untuk Pengembangan Pariwisata*. Dipetik Februari 28, 2019, dari https://travel.tempo.co/read/1109134/surabaya-dinobatkan-jadi-kota-terbaik-untuk-pengembangan-pariwisata/full&view=ok#

Traveloka. (2019). *Tentang Kami*. Dipetik Februari 27, 2019, dari https://m.traveloka.com/about

Turland, M. (2010). *php|architect's Guide to Web Scraping.* Canada: Marco Tabini & Associates, Inc.

Wardhani, N. K., Rezkiani, Kurniawan, S., Setiawan, H., Gata, G., Tohari, S., Gata, W., Wahyudi, M. (2018). Sentiment Analysis Article News Coordinator Minister of Maritime Affairs Using Algorithm Naive Bayes and Support Vector Machine with Particle Swarm Optimization. *Journal of Theoretical and Applied Information Technology*, *96*(24), 8365-8378.

Weiss, S., Indurkhya, N., Zhang, T., & Damerau, F. (2005). *Text Mining Predictive Methods for Analyzing Unstructured Information.* New York: Spinger.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining : Practical Machine Learning Tools and Techniques* (*3rd Edition*). United States: Morgan Kaufmann.

Yuniarto. (2009). *Klasifikasi Angkatan Kerja Propinsi Bengkulu Menggunakan Metode CART dan Regresi Logistik.* Surabaya: Tesis Institut Teknologi Sepuluh Nopember.

Zaky, M., & Meira, J. (2014). *Data Mining and Analysis : Foundations and Algorithms.* New York: Cambridge University Press.

LAMPIRAN

Lampiran 1. Data *Review* Hotel

Rating	***Account***	**Date**	***Review***
9.7	Ditta D.	Apr 30, 2019 - Family vacation	Senang sekali menginap 2 malam di fave hotel rungkut Surabaya, beruntung sekali saya mendapatkan kamar atas dengan view malam hari sungguh menakjubkan.
...	...	...	...
8.5	Daniel K. S.	Apr 24, 2019 - Family vacation	Kamar bersih, staf ramah, ada welcome drink, parkir, dekat dengan pusat perbelanjaan.
8.5	Adjeng A. J.	Apr 24, 2019 - Business travel	Tempatnya nyaman dan dekat dengan teansmart mini. Bisa di tempuh dengan berjalan kaki selama kurang lebih 5 menit.

Lampiran 2. Data Hasil Klasifikasi

No	Klasifikasi	Text
1	1	senang sekali menginap 2 malam di favehotel rungkut surabaya, beruntung sekali saya mendapatkan kamar atas dengan view malam hari sungguh menakjubkan.
...	...	...
1751	1	kamar bersih, staf ramah, ada welcome drink, parkir, dekat dengan pusat perbelanjaan.
1752	1	tempatnya nyaman dan dekat dengan transmart mini. bisa di tempuh dengan berjalan kaki selama kurang lebih 5 menit.

Lampiran 3. *Syntax Preprocessing Data*

```
import pandas as pd
import string
import nltk
from nltk.tokenize import word_tokenize
import re
import sys
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory

data=pd.read_excel("D:/TUGAS AKHIR LJ/DATAF/datafave_excel.
xlsx")
text = data['text']
print(data)

#Case Folding
datalower = []
for line in text:
  a = line.lower()
  datalower.append(a)

#Menghapus Angka
dataclearangka = []
for line in datalower:
    result = re.sub("\d"," ", line)
    dataclearangka.append(result)
    print(result)

#Menghapus Emoticon
dataclearemoticon = []
for line in dataclearangka:
    result = re.sub(r'<.*?>'," ", line)
    dataclearemoticon.append(result)
    print(result)

#Menghapus Punctuation
dataclearpuntuation = []
for line in dataclearemoticon:
    result = re.sub(r"[^\w\s]"," ", line)
    dataclearpuntuation.append(result)
    print(result)

#Stemming
factory = StemmerFactory()
stemmer = factory.create_stemmer()
data_stemmed = map(lambda x: stemmer.stem(x), datalower)
databersih = map(lambda x: x.translate(str.maketrans('','',
string.punctuation)), data_stemmed)
databersih = list(databersih)
```

Lampiran 3. *Syntax Preprocessing Data* (Lanjutan)

```
#Stopwords and Tokenizing
stopwords = open('D:/TUGAS AKHIR LJ/stopword.txt',
'r').read()
favedata = []
favefinal = []
df = []

for line in databersih:
   word_token=nltk.word_tokenize(line)
   word_token=[word for word in word_token if not word
in stopwords and not word[0].isdigit()]
   favefinal.append(word_token)
   df.append(" ".join(word_token))
for l in favefinal:
    favedata+= l
final_fave={v: favedata.count(v) for v in set(favedata)}

import csv
with open ('D:/TUGAS AKHIR LJ/DATAF/final_fave.csv','w',
newline="") as csv_file:
   writer = csv.writer(csv_file)
   for key, value in final_fave.items():
writer.writerow([key, value])

#Count Vectorize
from pandas import DataFrame
from sklearn.feature_extraction.text import
CountVectorizer
Y = data['klasifikasi']
Y_A = pd.DataFrame(Y)
vectorizer = CountVectorizer(min_df=1)
X = vectorizer.fit_transform(df)
X_ = DataFrame(X.A,columns=vectorizer.get_feature_names
())
X_.to_csv("D:/TUGAS AKHIR LJ/DATAF/cv_favehotel.csv")

#TF-IDF
from pandas import DataFrame
from sklearn.feature_extraction.text import
TfidfTransformer
tfidf = TfidfTransformer(use_idf=True).fit_transform(X_)
tfidf_nya = (tfidf.toarray())
X_nya = tfidf_nya
tf=DataFrame(tfidf.A,columns=vectorizer.get_feature_name
s())
print (tf)
tf.to_csv("D:/TUGAS AKHIR LJ/DATAF/TFIDF_favehotel.csv")
```

Lampiran 4. *Syntax Word Cloud*

```
#Word Cloud Gabung
import pandas as pd
import string
import nltk
from nltk.tokenize import word_tokenize
import re
import sys
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory

data = pd.read_excel("D:/TUGAS AKHIR LJ/DATAF/datafave
_excel.xlsx")
Review = data['text']
print(data)

#Split Data

data_pos = data [ data['klasifikasi'] == 1 ]
data_pos = data_pos ['text']

data_neg = data [ data['klasifikasi'] == 0 ]
data_neg = data_neg ['text']

#Menghapus Angka
dataclearangka_pos = []
for line in data_pos:
    result = re.sub("\d"," ", line)
    dataclearangka_pos.append(result)

dataclearangka_neg = []
for line in data_neg:
    result = re.sub("\d"," ", line)
    dataclearangka_neg.append(result)

#Menghapus Emoticon
dataclearemoticon_pos = []
for line in dataclearangka_pos:
    result = re.sub(r'<.*?>'," ", line)
    dataclearemoticon_pos.append(result)

dataclearemoticon_neg = []
for line in dataclearangka_neg:
    result = re.sub(r'<.*?>'," ", line)
    dataclearemoticon_neg.append(result)
```

Lampiran 4. *Syntax Word Cloud* (Lanjutan)

```
#Menghapus Punctuation
dataclearpuntuation_pos = []
for line in dataclearemoticon_pos:
    result = re.sub(r"[^\w\s]"," ", line)
    dataclearpuntuation_pos.append(result)

dataclearpuntuation_neg = []
for line in dataclearemoticon_neg:
    result = re.sub(r"[^\w\s]"," ", line)
    dataclearpuntuation_neg.append(result)

#Case Folding
datalower_pos=[]
for line in dataclearemoticon_pos:
 a=line.lower()
 datalower_pos.append(a)

datalower_neg=[]
for line in dataclearemoticon_neg:
 a=line.lower()
 datalower_neg.append(a)

#Stemming
factory=StemmerFactory()
stemmer=factory.create_stemmer()
datastemmed_pos=map(lambda x: stemmer.stem(x),
datalower_pos)
databersih_pos=map(lambda x: x.translate(str.maketrans('',
'',string.punctuation)), datastemmed_pos)
databersih_pos=list(databersih_pos)

factory=StemmerFactory()
stemmer=factory.create_stemmer()
datastemmed_neg=map(lambda x: stemmer.stem(x),
datalower_neg)
databersih_neg=map(lambda x: x.translate(str.maketrans('',
'',string.punctuation)), datastemmed_neg)
databersih_neg=list(databersih_neg)

#Stopwords and Tokenizing
stopwords = open('D:/TUGAS AKHIR LJ/stopword.txt',
'r').read()
favedata_pos = []
favefinal_pos = []
df_pos = []
```

Lampiran 4. *Syntax Word Cloud* (Lanjutan)

```
for line in databersih_pos:
   wt_pos = word_tokenize(line)
   wt_pos = [word for word in wt_pos if not word in
stopwords and not word[0].isdigit()]
   favefinal_pos.append(wt_pos)
   df_pos.append(" ".join(wt_pos))
for l in favefinal_pos:
    favedata_pos+= l
final_pos={v: favedata_pos.count(v) for v in set
(favedata_pos)}

favedata_neg = []
favefinal_neg = []
df_neg = []

for line in databersih_neg:
   wt_neg = word_tokenize(line)
   wt_neg = [word for word in wt_neg if not word in
stopwords and not word[0].isdigit()]
   favefinal_neg.append(wt_neg)
   df_neg.append(" ".join(wt_neg))
for l in favefinal_neg:
    favedata_neg+= l
final_neg={v: favedata_neg.count(v) for v in set
(favedata_neg)}

#Merubah Data str
a = str(df_pos)
positif =re.sub(r"'","",a)

b = str(df_neg)
negatif =re.sub(r"'","",b)

============================================================
                          Wordcloud
============================================================
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
%matplotlib inline
from subprocess import check_output
from wordcloud import WordCloud, STOPWORDS
mpl.rcParams['font.size']=12 #10
mpl.rcParams['savefig.dpi']=100 #72
mpl.rcParams['figure.subplot.bottom']=.1
```

Lampiran 4. *Syntax Word Cloud* (Lanjutan)

```
#WordCloud Positif
wordcloud = WordCloud(collocations = False,
                      background_color='white',
                      stopwords=stopwords,
                      max_words=50,
                      max_font_size=200,
                      random_state=42
                      ).generate(positif)
print(wordcloud)
fig = plt.figure(1)
plt.imshow(wordcloud)
plt.axis('off')
plt.show()
fig.savefig("D:/TUGAS AKHIR LJ/DATAF/wordpos.png", dpi=900)

#WordCloud Negatif
wordcloud = WordCloud(collocations = False,
                      background_color='white',
                      stopwords=stopwords,
                      max_words=50,
                      max_font_size=200,
                      random_state=42
                      ).generate(negatif)
print(wordcloud)
fig = plt.figure(1)
plt.imshow(wordcloud)
plt.axis('off')
plt.show()
fig.savefig("D:/TUGAS AKHIR LJ/DATAF/wordneg.png", dpi=900)
```

Lampiran 5. Analisis Klasifikasi Setiap Metode

```
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
import imblearn
import optunity
import optunity.metrics
from sklearn.metrics import classification_report,
confusion_matrix,accuracy_score
from __future__ import print_function
from sklearn.model_selection import train_test_split,
cross_val_score,StratifiedKFold,KFold,ShuffleSplit,Strati
fiedShuffleSplit
from sklearn.feature_selection import SelectPercentile,
f_classif
from sklearn.naive_bayes import BernoulliNB
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import classification_report
from sklearn.utils import shuffle
from sklearn import datasets
from sklearn.metrics import roc_curve, auc
from scipy import interp
from imblearn.over_sampling import SMOTE
from sklearn import metrics

========================
#Naive Bayes Classifier#
========================

Y_new = DataFrame.as_matrix(Y_A)
X_new, Y_new = X_nya, Y_new
kfold= StratifiedKFold(n_splits=10,shuffle=False)
kfold.get_n_splits(X_new)
kfold.get_n_splits(Y_new)
cl = BernoulliNB()
smote=SMOTE()

i=1
for train, test in kfold.split(X_new, Y_new):
    (X_new[train], X_new[test])
    (Y_new[train], Y_new[test])
    hitung = cl.fit(X_new[train], Y_new[train])
    predictNBC = cl.predict(X_new[test])
    fpr,tpr,_ =metrics.roc_curve(Y_new[test],predictNBC)
     auc_ = metrics.auc(fpr,tpr)
```

Lampiran 5. Analisis Klasifikasi Setiap Metode (Lanjutan)

```
    X_trainsmote,Y_trainsmote =
smote.fit_sample(X_new[train],Y_new[train])
    hitungsmote= cl.fit(X_trainsmote, Y_trainsmote)
     predictNBCsmote = cl.predict(X_new[test])
     fprsm,tprsm,_=metrics.roc_curve(Y_new[test],
predictNBCsmote)
     aucsm_ = metrics.auc(fprsm,tprsm)
     print ("----------FOLD KE = {:.0f}----------".format(i))
     np.savetxt("D:/TUGAS AKHIR LJ/DATAF/NBC/Xawal%s.csv"%i,
X_new[train], delimiter=",")
     np.savetxt("D:/TUGAS AKHIR LJ/DATAF/NBC/Yawal%s.csv"%i,
Y_new[train], delimiter=",")
     np.savetxt("D:/TUGAS AKHIR LJ/DATAF/NBC/Xsmote%s.csv"%i,
X_trainsmote, delimiter=",")
     np.savetxt("D:/TUGAS AKHIR LJ/DATAF/NBC/Ysmote%s.csv"%i,
Y_trainsmote, delimiter=",")
     np.savetxt("D:/TUGAS AKHIR LJ/DATAF/NBC/Xawaltest%s.csv"
%i, X_new[test], delimiter=",")
     np.savetxt("D:/TUGAS AKHIR LJ/DATAF/NBC/Yawaltest%s.csv"
%i, Y_new[test], delimiter=",")
    i=i+1
    print ("NBC Data Awal")
    print (confusion_matrix(Y_new[test], predictNBC))
    print (classification_report(Y_new[test], predictNBC))
    print ()
    print ("NBC data SMOTE")
    print (confusion_matrix(Y_new[test], predictNBCsmote))
    print (classification_report(Y_new[test], predictNBCsmote))
    print
("============================================================")
    print ("Accuracy Data Awal =
{:.2f}".format(accuracy_score(Y_new[test], predictNBC)))
    print ("Accuracy data SMOTE =
{:.2f}".format(accuracy_score(Y_new[test], predictNBCsmote)))
    print ("Area Under Curve ROC Data Awal =
{:.2f}".format(auc_))
    print ("Area Under Curve ROC data SMOTE =
{:.2f}".format(aucsm_))
    print
("============================================================")
```

Lampiran 5. Analisis Klasifikasi Setiap Metode (Lanjutan)

```
========================
#Regresi Logistik Biner#
========================

lr = LogisticRegression()

i=1
for train, test in kfold.split(X_new, Y_new):
    (X_new[train], X_new[test])
    (Y_new[train], Y_new[test])
     hitung = lr.fit(X_new[train], Y_new[train])
     predictBLR = lr.predict(X_new[test])
     fpr,tpr,_= metrics.roc_curve(Y_new[test],predictBLR)
     auc_ = metrics.auc(fpr,tpr)

     X_trainsmote,Y_trainsmote
smote.fit_sample(X_new[train],Y_new[train])
    hitungsmote = lr.fit(X_trainsmote,Y_trainsmote)
    predictBLRsmote = lr.predict(X_new[test])
    fprsm,tprsm,_ =
metrics.roc_curve(Y_new[test],predictBLRsmote)
    aucsm_ = metrics.auc(fprsm,tprsm)
    print ("----------FOLD KE = {:.0f}----------".format(i))
    np.savetxt("D:/TUGAS AKHIR LJ/DATAF/BLR/Xawal%s.csv"%i,
X_new[train], delimiter=",")
    np.savetxt("D:/TUGAS AKHIR LJ/DATAF/BLR/Yawal%s.csv"%i,
Y_new[train], delimiter=",")
    np.savetxt("D:/TUGAS AKHIR LJ/DATAF/BLR/Xsmote%s.csv"%i,
X_trainsmote, delimiter=",")
    np.savetxt("D:/TUGAS AKHIR LJ/DATAF/BLR/Ysmote%s.csv"%i,
Y_trainsmote, delimiter=",")
    np.savetxt("D:/TUGAS AKHIR LJ/DATAF/BLR/Xawaltest%s.csv"
%i, X_new[test], delimiter=",")
    np.savetxt("D:/TUGAS AKHIR LJ/DATAF/BLR/Yawaltest%s.csv"
%i, Y_new[test], delimiter=",")
    i=i+1
    print ("BLR Data Awal")
    print (confusion_matrix(Y_new[test], predictBLR))
    print (classification_report(Y_new[test], predictBLR))
    print ()
    print ("BLR DATA SMOTE")
    print (confusion_matrix(Y_new[test], predictBLRsmote))
    print (classification_report(Y_new[test], predictBLRsmote))
    print
("============================================================")
```

Lampiran 5. Analisis Klasifikasi Setiap Metode (Lanjutan)

```
     print ("Accuracy Data Awal =
{:.2f}".format(accuracy_score(Y_new[test], predictBLR)))
    print ("Accuracy data SMOTE =
{:.2f}".format(accuracy_score(Y_new[test], predictBLRsmote)))
    print ("Area Under Curve ROC Data Awal =
{:.2f}".format(auc_))
    print ("Area Under Curve ROC data SMOTE =
{:.2f}".format(aucsm_))
    print
("============================================================"
```

Lampiran 6. Perhitungan Manual NBC

Perhitungan Prior Positif
$P(y_{pos}) = \frac{\lvert docs_{positif} \rvert}{\lvert contoh \rvert} = \frac{1051}{1598} = 0,6491$
Perhitungan Prior Negatif
$P(y_{neg}) = \frac{\lvert docs_{negatif} \rvert}{\lvert contoh \rvert} = \frac{547}{1598} = 0,3508$
Perhitungan Posterior Positif
$P(x_1 \mid y_{pos}) = \frac{m_1 + 1}{\lvert m + kosakata \rvert} = \frac{1,7562 + 1}{\lvert 2126,447 + 428 \rvert} = 0,00107$ $P(x_2 \mid y_{pos}) = \frac{m_2 + 1}{\lvert m + kosakata \rvert} = \frac{0,7757 + 1}{\lvert 2126,447 + 428 \rvert} = 0,00069$ $\vdots$ $P(x_{428} \mid y_{pos}) = \frac{m_{428} + 1}{\lvert m + kosakata \rvert} = \frac{3,523 + 1}{\lvert 2126,447 + 428 \rvert} = 0,00177$
Perhitungan Posterior Negatif
$P(x_1 \mid y_{neg}) = \frac{m_1 + 1}{\lvert m + kosakata \rvert} = \frac{8,961 + 1}{\lvert 1314,012 + 428 \rvert} = 0,00571$ $P(x_2 \mid y_{neg}) = \frac{m_2 + 1}{\lvert m + kosakata \rvert} = \frac{1,577 + 1}{\lvert 1314,012 + 428 \rvert} = 0,00147$ $\vdots$ $P(x_{428} \mid y_{neg}) = \frac{m_{428} + 1}{\lvert m + kosakata \rvert} = \frac{12,100 + 1}{\lvert 1314,012 + 428 \rvert} = 0,00752$
Perhitungan *Review* ke-*i* Sentimen Positif
$P(y_{pos}) \prod_{j=1}^{428} P(x_j \mid y_{pos}) = 0,65 \times 0,00107^0 \times 0,00069^0 \times \ldots \times 0,00177^0 = 1,29 \times 10^{-5}$ $P(y_{pos}) \prod_{j=1}^{428} P(x_j \mid y_{pos}) = 0,65 \times 0,00107^0 \times 0,00069^0 \times \ldots \times 0,00177^0 = 4,69 \times 10^{-8}$ $\vdots$

$P\left(y_{pos}\right)\prod_{j=1}^{428} P\left(x_j \mid y_{pos}\right) = 0,66 \times 0,00107^{0} \times 0,00069^{0} \times \ldots \times 0,00177^{0} = 7,54 \times 10^{-4}$
Perhitungan *Review* ke-*i* Sentimen Negatif
$P\left(y_{pos}\right)\prod_{j=1}^{428} P\left(x_j \mid y_{pos}\right) = 0,35 \times 0,00571^{0} \times 0,00147^{0} \times \ldots \times 0,00752^{0} = 1,02 \times 10^{-5}$
$P\left(y_{pos}\right)\prod_{j=1}^{428} P\left(x_j \mid y_{pos}\right) = 0,35 \times 0,00571^{0} \times 0,00147^{0} \times \ldots \times 0,00752^{0} = 1,99 \times 10^{-7}$
$\vdots$
$P\left(y_{pos}\right)\prod_{j=1}^{428} P\left(x_j \mid y_{pos}\right) = 0,35 \times 0,00557^{0} \times 0,00151^{0} \times \ldots \times 0,00702^{0} = 1,18 \times 10^{-4}$

Lampiran 7A. Hasil NBC Data Awal (*Training* dan *Testing*) dengan 10-Fold

Training								
FOLD KE-	**Favehotel**				**Hotel Gunawangsa**			
	ACC	**P**	**R**	**AUC**	**ACC**	**P**	**R**	**AUC**
1	0,87	0,84	0,86	0,74	0,84	0,88	0,71	0,82
2	0,87	0,87	0,74	0,84	0,84	0,89	0,70	0,82
3	0,87	0,86	0,75	0,84	0,83	0,88	0,69	0,81
4	0,86	0,86	0,74	0,84	0,83	0,87	0,69	0,81
5	0,86	0,85	0,73	0,83	0,83	0,87	0,69	0,81
6	0,86	0,85	0,74	0,83	0,83	0,87	0,70	0,81
7	0,86	0,86	0,74	0,83	0,84	0,87	0,71	0,82
8	0,87	0,85	0,74	0,84	0,83	0,87	0,69	0,81
9	0,86	0,86	0,74	0,83	0,83	0,87	0,71	0,82
10	0,86	0,85	0,74	0,83	0,83	0,87	0,71	0,82
Rata-rata	**0,86**	**0,86**	**0,75**	**0,83**	**0,83**	**0,87**	**0,70**	**0,82**
Testing								
FOLD KE-	**Favehotel**				**Hotel Gunawangsa**			
	ACC	**P**	**R**	**AUC**	**ACC**	**P**	**R**	**AUC**
1	0,85	0,76	0,86	0,86	0,83	0,73	0,91	0,84
2	0,82	0,75	0,71	0,79	0,78	0,72	0,78	0,78
3	0,82	0,75	0,73	0,8	0,81	0,82	0,69	0,79
4	0,89	0,86	0,82	0,88	0,83	0,90	0,68	0,81
5	0,86	0,88	0,71	0,83	0,82	0,83	0,70	0,8
6	0,81	0,83	0,56	0,75	0,80	0,82	0,66	0,78
7	0,85	0,86	0,68	0,81	0,79	0,91	0,56	0,76
8	0,86	0,89	0,68	0,82	0,84	0,97	0,63	0,81
9	0,81	0,78	0,65	0,77	0,79	0,84	0,62	0,77
10	0,84	0,85	0,66	0,8	0,76	0,90	0,48	0,72
Rata-rata	**0,84**	**0,82**	**0,71**	**0,81**	**0,81**	**0,84**	**0,67**	**0,79**

Lampiran 7B. Hasil NBC Data SMOTE (*Training* dan *Testing*) dengan 10-Fold

Training								
FOLD KE-	**Favehotel**				**Hotel Gunawangsa**			
	ACC	**P**	**R**	**AUC**	**ACC**	**P**	**R**	**AUC**
1	0,87	0,87	0,91	0,83	0,83	0,91	0,74	0,83
2	0,88	0,92	0,84	0,88	0,83	0,91	0,74	0,83
3	0,87	0,92	0,83	0,87	0,83	0,91	0,73	0,83
4	0,86	0,91	0,80	0,86	0,82	0,89	0,73	0,82
5	0,87	0,91	0,83	0,87	0,83	0,90	0,74	0,83
6	0,87	0,91	0,81	0,87	0,83	0,91	0,74	0,83
7	0,87	0,92	0,81	0,87	0,83	0,90	0,75	0,83
8	0,88	0,91	0,83	0,88	0,82	0,89	0,73	0,82
9	0,87	0,92	0,81	0,87	0,83	0,90	0,75	0,83
10	0,86	0,91	0,80	0,86	0,83	0,90	0,74	0,83
Rata-rata	**0,87**	**0,91**	**0,83**	**0,87**	**0,83**	**0,90**	**0,74**	**0,83**
Testing								
FOLD KE-	**Favehotel**				**Hotel Gunawangsa**			
	ACC	**P**	**R**	**AUC**	**ACC**	**P**	**R**	**AUC**
1	0,87	0,76	0,90	0,87	0,83	0,74	0,93	0,85
2	0,81	0,73	0,73	0,79	0,80	0,74	0,82	0,81
3	0,83	0,75	0,78	0,82	0,82	0,82	0,71	0,8
4	0,91	0,86	0,89	0,9	0,83	0,90	0,69	0,81
5	0,88	0,87	0,76	0,85	0,82	0,83	0,74	0,81
6	0,83	0,85	0,63	0,78	0,80	0,79	0,68	0,78
7	0,87	0,87	0,74	0,84	0,80	0,91	0,57	0,76
8	0,86	0,86	0,71	0,82	0,84	0,94	0,67	0,82
9	0,82	0,78	0,69	0,79	0,81	0,85	0,67	0,79
10	0,84	0,85	0,66	0,8	0,77	0,86	0,53	0,74
Rata-rata	**0,85**	**0,82**	**0,75**	**0,83**	**0,81**	**0,84**	**0,70**	**0,80**

Lampiran 8. *Output* Pendeteksian Multikolinieritas (Favehotel)

Variabel	VIF	Keterangan	Variabel	VIF	Keterangan
acnya	4,236	Tidak Terjadi Multiko	internetnya	17,904	Multiko
agam	2,178	Tidak Terjadi Multiko	istimewa	162,728	Multiko
airport	1,000	Tidak Terjadi Multiko	istirahat	3,522	Tidak Terjadi Multiko
akses	1,773	Tidak Terjadi Multiko	istri	1,301	Tidak Terjadi Multiko
aktivitas	1,016	Tidak Terjadi Multiko	iya	6,051	Tidak Terjadi Multiko
alas	5,203	Tidak Terjadi Multiko	jaga	1,885	Tidak Terjadi Multiko
alat	8,974	Tidak Terjadi Multiko	jalan	2,102	Tidak Terjadi Multiko
alhamdulillah	1,411	Tidak Terjadi Multiko	jangkau	1,724	Tidak Terjadi Multiko
aneh	34,999	Multiko	jelek	35,184	Multiko
angkat	3,907	Tidak Terjadi Multiko	jendela	1,834	Tidak Terjadi Multiko
apek	1053,158	Multiko	jenis	1,700	Tidak Terjadi Multiko
aplikasi	11,319	Multiko	jual	3208,628	Multiko
arah	1,000	Tidak Terjadi Multiko	juanda	28,540	Multiko
⋮	⋮	⋮	⋮	⋮	⋮
harga	2,119	Tidak Terjadi Multiko	tv	5,145	Tidak Terjadi Multiko
harum	22,457	Multiko	twinbed	1,000	Tidak Terjadi Multiko
heater	7,445	Tidak Terjadi Multiko	ubaya	1,000	Tidak Terjadi Multiko
hemat	2,132	Tidak Terjadi Multiko	ukur	1,020	Tidak Terjadi Multiko
hibur	1,412	Tidak Terjadi Multiko	unjung	87,421	Multiko
hujan	4,665	Tidak Terjadi Multiko	variatif	1,410	Tidak Terjadi Multiko
inap	1,609	Tidak Terjadi Multiko	view	3,345	Tidak Terjadi Multiko
indah	1,000	Tidak Terjadi Multiko	viewnya	564,777	Multiko
industri	1,791	Tidak Terjadi Multiko	wangi	24,390	Multiko
info	8,973	Tidak Terjadi Multiko	wastafel	1,000	Tidak Terjadi Multiko
interior	1,037	Tidak Terjadi Multiko	welcome	8,467	Tidak Terjadi Multiko
internet	1,234	Tidak Terjadi Multiko	wifi	7,055	Tidak Terjadi Multiko

Lampiran 9. *Output* Pendeteksian Multikolinieritas (Gunawangsa)

Variabel	VIF	Keterangan	Variabel	VIF	Keterangan
abis	1,950	Tidak Terjadi Multiko	jarang	1,552	Tidak Terjadi Multiko
acara	1,364	Tidak Terjadi Multiko	jaring	7,096	Tidak Terjadi Multiko
acnya	1,752	Tidak Terjadi Multiko	jatuh	1,000	Tidak Terjadi Multiko
aerobik	1,000	Tidak Terjadi Multiko	jelek	1,386	Tidak Terjadi Multiko
agam	1,256	Tidak Terjadi Multiko	jendela	6,580	Tidak Terjadi Multiko
akses	2,727	Tidak Terjadi Multiko	jenis	1,169	Tidak Terjadi Multiko
alami	2,892	Tidak Terjadi Multiko	jorok	1,000	Tidak Terjadi Multiko
alas	4,178	Tidak Terjadi Multiko	jual	1,000	Tidak Terjadi Multiko
alat	2,589	Tidak Terjadi Multiko	juanda	1,000	Tidak Terjadi Multiko
alfamart	1,486	Tidak Terjadi Multiko	juara	1,769	Tidak Terjadi Multiko
alhamdulillah	1,000	Tidak Terjadi Multiko	junior	4,509	Tidak Terjadi Multiko
anak	1,373	Tidak Terjadi Multiko	jutek	35,258	Multiko
anang	4,323	Tidak Terjadi Multiko	kabel	1,579	Tidak Terjadi Multiko
⋮	⋮	⋮	⋮	⋮	⋮
huni	1,202	Tidak Terjadi Multiko	varian	1,246	Tidak Terjadi Multiko
inap	1,507	Tidak Terjadi Multiko	variasi	1,376	Tidak Terjadi Multiko
indah	1,225	Tidak Terjadi Multiko	variatif	1,300	Tidak Terjadi Multiko
info	1,736	Tidak Terjadi Multiko	via	3,218	Tidak Terjadi Multiko
informasi	1,917	Tidak Terjadi Multiko	view	2,003	Tidak Terjadi Multiko
istimewa	49,453	Multiko	viewnya	1,592	Tidak Terjadi Multiko
istirahat	1,289	Tidak Terjadi Multiko	voucher	2,112	Tidak Terjadi Multiko
jaga	2,073	Tidak Terjadi Multiko	wangi	1,205	Tidak Terjadi Multiko
jalan	3,713	Tidak Terjadi Multiko	warna	1,791	Tidak Terjadi Multiko
jamin	3,241	Tidak Terjadi Multiko	wastafel	1,000	Tidak Terjadi Multiko
jangkau	1,869	Tidak Terjadi Multiko	wifi	1,430	Tidak Terjadi Multiko
jarak	1,097	Tidak Terjadi Multiko	wifinya	1,797	Tidak Terjadi Multiko

Lampiran 10A. Hasil Uji Serentak dan Parsial Favehotel Data Awal

Uji Serentak

Null deviance: 2069.24 on 1596 degrees of freedom
Residual deviance: 268.71 on 1180 degrees of freedom
AIC: 1102.7

Uji Parsial					
	Estimate	**Std. Error**	**Wald**	$\chi^2_{(0,05;1)}$	**Keputusan**
(Intercept)	0,384	0,098			
bagus	3,992	0,704	32,120	3,841	Tolak H0
bau	-7,469	1,799	17,237	3,841	Tolak H0
cepat	3,407	1,156	8,687	3,841	Tolak H0
checkin	-2,723	0,836	10,624	3,841	Tolak H0
cuek	-4,460	1,609	7,684	3,841	Tolak H0
dingin	-2,939	0,889	10,936	3,841	Tolak H0
fasilitas	4,346	1,214	12,825	3,841	Tolak H0
ganggu	-6,392	2,278	7,871	3,841	Tolak H0
jalan	-3,108	0,800	15,085	3,841	Tolak H0
keliling	-9,388	4,096	5,253	3,841	Tolak H0
kotor	-6,071	1,448	17,587	3,841	Tolak H0
lambat	-5,862	2,200	7,102	3,841	Tolak H0
lift	-4,603	1,904	5,846	3,841	Tolak H0
lumayan	4,215	0,912	21,351	3,841	Tolak H0
mudah	3,558	1,674	4,520	3,841	Tolak H0
nyaman	5,126	0,736	48,455	3,841	Tolak H0
panas	-2,711	0,588	21,267	3,841	Tolak H0
ramah	5,324	1,093	23,705	3,841	Tolak H0

Lampiran 10A. Hasil Uji Serentak dan Parsial Favehotel Data Awal (Lanjutan)

Uji Parsial					
	Estimate	**Std. Error**	**Wald**	$\chi^2_{(0,05;1)}$	**Keputusan**
rapi	8,176	2,238	13,350	3,841	Tolak H0
resik	-3,322	0,971	11,700	3,841	Tolak H0
sempit	-3,234	1,098	8,676	3,841	Tolak H0
sopan	-2,779	1,008	7,606	3,841	Tolak H0
strategis	5,045	1,226	16,944	3,841	Tolak H0
suara	-8,319	1,604	26,917	3,841	Tolak H0
tidur	-2,616	1,162	5,067	3,841	Tolak H0
toilet	-5,060	1,067	22,497	3,841	Tolak H0
transmart	3,451	0,809	18,216	3,841	Tolak H0
tunggu	-4,631	1,898	5,951	3,841	Tolak H0
view	5,800	1,947	8,875	3,841	Tolak H0

Lampiran 10B. Hasil Uji Serentak dan Parsial Favehotel Data SMOTE

Uji Serentak

Null deviance: 2069.24 on 1596 degrees of freedom
Residual deviance: 608.27 on 1273 degrees of freedom
AIC: 1256.3

Lampiran 10B. Hasil Uji Serentak dan Parsial Favehotel Data SMOTE (Lanjutan)

Uji Parsial					
	Estimate	**Std. Error**	**Wald**	$\chi^2_{(0,05;1)}$	**Keputusan**
(Intercept)	-0,847	0,088			
bagus	2,775	0,358	59,965	3,841	Tolak H0
bandara	3,183	0,752	17,917	3,841	Tolak H0
bathtub	5,427	2,152	6,360	3,841	Tolak H0
better	-3,385	1,397	5,869	3,841	Tolak H0
cepat	2,410	0,702	11,799	3,841	Tolak H0
cocok	2,986	0,921	10,511	3,841	Tolak H0
cuek	-2,230	0,926	5,803	3,841	Tolak H0
dingin	-1,513	0,551	7,544	3,841	Tolak H0
fasilitas	3,334	0,599	30,956	3,841	Tolak H0
ganggu	-2,170	0,882	6,049	3,841	Tolak H0
hambar	-3,605	0,813	19,679	3,841	Tolak H0
hibur	2,930	1,033	8,048	3,841	Tolak H0
jalan	-1,898	0,602	9,921	3,841	Tolak H0
kelas	-8,408	2,601	10,450	3,841	Tolak H0
lokasi	1,477	0,535	7,629	3,841	Tolak H0
lumayan	3,516	0,487	52,047	3,841	Tolak H0
mall	2,181	0,836	6,805	3,841	Tolak H0
mini	3,401	1,209	7,914	3,841	Tolak H0
mudah	3,279	0,890	13,586	3,841	Tolak H0
murah	2,568	0,798	10,357	3,841	Tolak H0
nyaman	3,735	0,350	113,669	3,841	Tolak H0
panas	-1,404	0,388	13,084	3,841	Tolak H0
puas	1,784	0,378	22,242	3,841	Tolak H0

Lampiran 10B. Hasil Uji Serentak dan Parsial Favehotel Data SMOTE (Lanjutan)

Uji Parsial					
	Estimate	**Std. Error**	**Wald**	$\chi^2_{(0,05;1)}$	**Keputusan**
ramah	3,442	0,446	59,584	3,841	Tolak H0
rapi	3,206	1,028	9,726	3,841	Tolak H0
rekomendasi	2,197	0,613	12,853	3,841	Tolak H0
resik	-2,124	0,751	8,005	3,841	Tolak H0
selimut	-4,715	1,409	11,203	3,841	Tolak H0
sempit	-1,809	0,645	7,870	3,841	Tolak H0
senang	1,953	0,486	16,128	3,841	Tolak H0
sprei	-2,728	0,917	8,844	3,841	Tolak H0
strategis	2,336	0,474	24,246	3,841	Tolak H0
suara	-3,016	0,542	31,017	3,841	Tolak H0
terimakasih	2,249	0,615	13,373	3,841	Tolak H0
tingkat	2,664	0,599	19,767	3,841	Tolak H0
toilet	-2,779	0,564	24,246	3,841	Tolak H0
top	3,147	1,398	5,065	3,841	Tolak H0
transmart	2,428	0,421	33,285	3,841	Tolak H0
view	3,470	0,935	13,776	3,841	Tolak H0

Lampiran 10C. *Threshold* Data Awal

```
> datatest = read_excel("D:/TUGAS AKHIR LJ/DATAF/BARU/test
_pred_ori.xlsx",sheet=1,col_names=TRUE)
> roc_op=roc(datatest$pred,datatest$act)
> threshold=coords(roc_op, "best", ret=c("threshold"), bes
t.method="youden",transpose=FALSE);threshold
[1] 0.5
```

Lampiran 11A. Hasil Uji Serentak dan Parsial Gunawangsa Data Awal

Uji Serentak

Null deviance: 2649.4 on 1953 degrees of freedom
Residual deviance: 1156.0 on 1491 degrees of freedom
AIC: 2082

Uji Parsial					
	Estimate	**Std. Error**	**Wald**	$\chi^2_{(0,05;1)}$	**Keputusan**
(Intercept)	-0,020	0,082			
bagus	2,554	0,366	48,584	3,841	Tolak H0
bau	-3,518	0,630	31,144	3,841	Tolak H0
bekas	-3,795	1,727	4,827	3,841	Tolak H0
berisik	-3,551	0,893	15,812	3,841	Tolak H0
bersih	4,276	0,350	148,975	3,841	Tolak H0
breakfastnya	-1,355	0,622	4,741	3,841	Tolak H0
buruk	-3,291	0,896	13,487	3,841	Tolak H0
clean	-3,953	1,353	8,538	3,841	Tolak H0
cocok	3,011	0,750	16,141	3,841	Tolak H0
cuek	-3,446	1,028	11,228	3,841	Tolak H0
fasilitas	1,702	0,463	13,542	3,841	Tolak H0
fungsi	-3,877	1,309	8,776	3,841	Tolak H0
ganggu	-4,348	0,985	19,502	3,841	Tolak H0
good	-2,096	0,730	8,247	3,841	Tolak H0
hambar	-2,734	1,094	6,245	3,841	Tolak H0
kedap	-4,628	1,185	15,243	3,841	Tolak H0
kotor	-4,683	0,652	51,571	3,841	Tolak H0
lumayan	1,900	0,541	12,345	3,841	Tolak H0

Lampiran 11A. Hasil Uji Serentak dan Parsial Gunawangsa Data Awal (Lanjutan)

Uji Parsial					
	Estimate	**Std. Error**	**Wald**	$\chi^2_{(0,05;1)}$	**Keputusan**
mantap	2,340	0,584	16,064	3,841	Tolak H0
nunggu	-4,481	1,660	7,286	3,841	Tolak H0
puas	1,672	0,361	21,472	3,841	Tolak H0
ribet	-4,220	1,911	4,876	3,841	Tolak H0
salur	-2,986	1,386	4,645	3,841	Tolak H0
satpam	-3,052	1,269	5,781	3,841	Tolak H0
sempit	-2,594	0,651	15,877	3,841	Tolak H0
sopan	-2,012	0,983	4,192	3,841	Tolak H0
staf	1,653	0,656	6,357	3,841	Tolak H0
strategis	2,607	0,484	29,056	3,841	Tolak H0
tamu	-2,662	0,957	7,735	3,841	Tolak H0
tenteram	3,723	0,769	23,415	3,841	Tolak H0
terimakasih	1,576	0,674	5,462	3,841	Tolak H0
tingkat	2,001	0,534	14,040	3,841	Tolak H0
tunggu	-3,845	1,118	11,836	3,841	Tolak H0

Lampiran 11B. Hasil Uji Serentak dan Parsial Gunawangsa Data SMOTE

Uji Serentak

Null deviance: 3133.0 on 2259 degrees of freedom
Residual deviance: 1352.2 on 1797 degrees of freedom
AIC: 2278,2

Lampiran 11B. Hasil Uji Serentak dan Parsial Gunawangsa Data SMOTE (Lanjutan)

Uji Parsial					
	Estimate	**Std. Error**	**Wald**	$\chi^2_{(0,05;1)}$	**Keputusan**
(Intercept)	-0,866	0,078			
anak	-5,663	2,080	7,414	3,841	Tolak H0
bagus	2,928	0,382	58,797	3,841	Tolak H0
bandara	3,654	0,726	25,317	3,841	Tolak H0
bau	-2,633	0,577	20,853	3,841	Tolak H0
bekas	-3,046	1,516	4,035	3,841	Tolak H0
bersih	4,515	0,383	138,934	3,841	Tolak H0
buruk	-2,133	0,836	6,506	3,841	Tolak H0
clean	-3,070	1,208	6,458	3,841	Tolak H0
cocok	3,816	0,719	28,149	3,841	Tolak H0
cuek	-2,094	0,873	5,758	3,841	Tolak H0
fasilitas	2,199	0,472	21,716	3,841	Tolak H0
ganggu	-3,618	0,931	15,092	3,841	Tolak H0
good	-1,672	0,726	5,300	3,841	Tolak H0
hambar	-2,105	0,937	5,045	3,841	Tolak H0
kecuali	2,613	1,105	5,594	3,841	Tolak H0
kedap	-3,771	1,228	9,426	3,841	Tolak H0
keren	2,499	0,595	17,638	3,841	Tolak H0
kota	2,293	0,719	10,163	3,841	Tolak H0
kotor	-3,459	0,574	36,359	3,841	Tolak H0
kualitas	2,432	1,002	5,888	3,841	Tolak H0
lumayan	2,787	0,517	29,109	3,841	Tolak H0
mantap	2,615	0,600	19,025	3,841	Tolak H0
nikmat	3,800	1,221	9,677	3,841	Tolak H0

Lampiran 11B. Hasil Uji Serentak dan Parsial Gunawangsa Data SMOTE (Lanjutan)

Uji Parsial					
	Estimate	**Std. Error**	**Wald**	$\chi^2_{(0,05;1)}$	**Keputusan**
nunggu	-3,661	1,155	10,044	3,841	Tolak H0
nyaman	1,885	0,347	29,450	3,841	Tolak H0
puas	2,052	0,369	30,969	3,841	Tolak H0
ramah	2,125	0,456	21,701	3,841	Tolak H0
rekomendasi	1,475	0,477	9,560	3,841	Tolak H0
satpam	-2,473	1,179	4,395	3,841	Tolak H0
sedap	-3,880	1,818	4,555	3,841	Tolak H0
sempit	-1,956	0,603	10,522	3,841	Tolak H0
strategis	2,699	0,495	29,775	3,841	Tolak H0
sukses	3,059	1,312	5,432	3,841	Tolak H0
super	2,730	1,173	5,420	3,841	Tolak H0
tahan	2,836	1,070	7,023	3,841	Tolak H0
tamu	-2,820	0,883	10,193	3,841	Tolak H0
tenteram	3,339	0,865	14,890	3,841	Tolak H0
terimakasih	2,322	0,663	12,253	3,841	Tolak H0
tingkat	2,534	0,563	20,239	3,841	Tolak H0
tunggu	-2,548	0,993	6,581	3,841	Tolak H0

Lampiran 11C. *Threshold* Data Awal

```
> datatest = read_excel("D:/TUGAS AKHIR LJ/DATAG/BARU/test
_pred_ori.xlsx",sheet=1,col_names=TRUE)
> roc_op=roc(datatest$pred,datatest$act)
> threshold=coords(roc_op, "best", ret=c("threshold"), bes
t.method="youden",transpose=FALSE);threshold
[1] 0.5
```

Lampiran 12A. Hasil RLB Data Awal (*Training* dan *Testing*) dengan 10-Fold

Training								
FOLD KE-	**Favehotel**				**Hotel Gunawangsa**			
	ACC	**P**	**R**	**AUC**	**ACC**	**P**	**R**	**AUC**
1	0,88	0,87	0,78	0,86	0,86	0,84	0,84	0,86
2	0,89	0,88	0,79	0,87	0,87	0,84	0,84	0,86
3	0,89	0,88	0,78	0,86	0,87	0,84	0,84	0,86
4	0,89	0,88	0,80	0,87	0,87	0,83	0,84	0,86
5	0,88	0,87	0,79	0,86	0,87	0,84	0,84	0,87
6	0,89	0,88	0,78	0,86	0,88	0,85	0,86	0,87
7	0,88	0,87	0,78	0,86	0,87	0,84	0,86	0,87
8	0,89	0,88	0,79	0,86	0,86	0,84	0,84	0,86
9	0,89	0,88	0,78	0,86	0,87	0,84	0,85	0,87
10	0,89	0,88	0,79	0,87	0,87	0,84	0,85	0,87
Rata-rata	**0,89**	**0,88**	**0,79**	**0,86**	**0,87**	**0,84**	**0,85**	**0,87**
Testing								
FOLD KE-	**Favehotel**				**Hotel Gunawangsa**			
	ACC	**P**	**R**	**AUC**	**ACC**	**P**	**R**	**AUC**
1	0,89	0,85	0,84	0,88	0,85	0,79	0,89	0,86
2	0,82	0,78	0,68	0,79	0,80	0,71	0,85	0,81
3	0,84	0,80	0,75	0,82	0,81	0,83	0,80	0,81
4	0,84	0,82	0,68	0,8	0,81	0,89	0,75	0,82
5	0,88	0,87	0,77	0,86	0,84	0,81	0,84	0,84
6	0,84	0,83	0,69	0,81	0,79	0,77	0,65	0,76
7	0,87	0,87	0,74	0,84	0,82	0,74	0,77	0,81
8	0,83	0,83	0,65	0,79	0,87	0,85	0,75	0,84
9	0,82	0,76	0,73	0,8	0,79	0,70	0,73	0,78
10	0,82	0,80	0,65	0,78	0,83	0,79	0,68	0,79
Rata-rata	**0,85**	**0,82**	**0,72**	**0,82**	**0,82**	**0,79**	**0,77**	**0,81**

Lampiran 12B. Hasil RLB Data SMOTE (*Training* dan *Testing*) dengan 10-Fold

Training								
FOLD KE-	**Favehotel**				**Hotel Gunawangsa**			
	ACC	**P**	**R**	**AUC**	**ACC**	**P**	**R**	**AUC**
1	0,88	0,86	0,92	0,88	0,87	0,84	0,90	0,87
2	0,89	0,87	0,93	0,89	0,88	0,85	0,90	0,88
3	0,88	0,87	0,91	0,88	0,88	0,85	0,92	0,88
4	0,89	0,87	0,92	0,89	0,87	0,85	0,91	0,87
5	0,90	0,87	0,93	0,9	0,87	0,85	0,91	0,87
6	0,89	0,87	0,92	0,89	0,88	0,85	0,92	0,88
7	0,89	0,87	0,92	0,89	0,87	0,84	0,91	0,87
8	0,89	0,87	0,92	0,89	0,87	0,85	0,91	0,87
9	0,88	0,86	0,91	0,88	0,88	0,86	0,91	0,88
10	0,89	0,87	0,92	0,89	0,88	0,86	0,92	0,88
Rata-rata	**0,89**	**0,87**	**0,92**	**0,89**	**0,88**	**0,85**	**0,91**	**0,88**
Testing								
FOLD KE-	**Favehotel**				**Hotel Gunawangsa**			
	ACC	**P**	**R**	**AUC**	**ACC**	**P**	**R**	**AUC**
1	0,85	0,75	0,89	0,86	0,83	0,74	0,92	0,84
2	0,83	0,71	0,87	0,84	0,80	0,69	0,92	0,82
3	0,83	0,72	0,84	0,83	0,82	0,78	0,90	0,82
4	0,84	0,72	0,87	0,84	0,82	0,84	0,83	0,82
5	0,81	0,69	0,84	0,82	0,80	0,73	0,88	0,81
6	0,85	0,78	0,79	0,83	0,79	0,71	0,74	0,78
7	0,85	0,75	0,84	0,85	0,81	0,71	0,82	0,81
8	0,84	0,73	0,84	0,84	0,90	0,84	0,90	0,9
9	0,85	0,74	0,90	0,86	0,78	0,66	0,82	0,79
10	0,80	0,68	0,79	0,8	0,80	0,69	0,74	0,78
Rata-rata	**0,84**	**0,73**	**0,85**	**0,84**	**0,82**	**0,74**	**0,85**	**0,82**

Lampiran 13. Surat Pernyataan Sumber Data

SURAT PERNYATAAN

Saya yang bertanda tangan di bawah ini, mahasiswa Departemen Statistika FMKSD ITS:

Nama : Silvia Astri Rahmaningrum

NRP : 062117 4500 0011

menyatakan bahwa data yang digunakan dalam Tugas Akhir / Thesis ini merupakan data sekunder yang diambil dari Penelitian / Buku / Tugas Akhir/ Thesis / Publikasi lainnya yaitu:

Sumber : https://www.traveloka.com

Keterangan : Data *review* dengan *keywords* "Favehotel Rungkut Surabaya" dan "Hotel Gunawangsa MERR"

Surat Pernyataan ini dibuat dengan sebenarnya. Apabila terdapat pemalsuan data maka saya siap menerima sanksi sesuai aturan yang berlaku.

Mengetahui
Pembimbing Tugas Akhir

(Pratnya Paramitha Oktaviana, S.Si, M.Si)
NIP. 1300201405001

Surabaya, Juli 2019

(Silvia Astri Rahmaningrum)
NRP. 062117 4500 0011

*(coret yang tidak perlu)

BIODATA PENULIS

Penulis bernama lengkap Silvia Astri Rahmaningrum, biasa disapa Silvia, lahir di Tulungagung pada tanggal 05 Januari 1996. Penulis adalah putri bungsu dari 2 bersaudara pasangan Hari Widodo dan Sri Iswati. Penulis telah menempuh pendidikan formal yaitu di SDN Karangwaru I, SMPN 2 Tulungagung, SMAN 1 Kedungwaru, Statistika Bisnis ITS. Pada tahun 2017 diterima menjadi mahasiswa Departemen Statistika ITS Program Studi Lintas Jalur. Selama menjadi mahasiswa, penulis aktif dalam beberapa kegiatan kemahasiswaan di ITS, diantaranya menjadi pengurus Paduan Suara Mahasiswa (PSM ITS) periode 2015/2016 sebagai staf Departemen RT, pengurus PSMITS periode 2016/2017 sebagai Kepala Departemen RT. Selain itu, penulis juga pernah mengikuti pelatihan LKMM pra-TD, LKMM TD, LKMM TM IX FMIPA. Penulis pernah melakukan kerja praktik di PT. Steel Pipe Industry of Indonesia, Tbk. (SPINDO Unit IV) Pasuruan dan Balai Penelitian Jeruk dan Buah Subtropika (Balitjestro). Dengan motto "Man Jadda Wajada", penulis meyakini jika segala sesuatu dilakukan dengan bersungguh-sungguh dan sabar maka Allah akan menunjukkan jalan-Nya. Apabila pembaca ingin memberikan saran dan kritik serta diskusi lebih lanjut mengenai Tugas Akhir ini dapat disampaikan melalui email astriviaa@gmail.com.

(Halaman ini sengaja dikosongkan)