IMPLEMENTASI SVM DAN ASOSIASI UNTUK SENTIMENT ANALYSIS DATA ULASAN THE PHOENIX HOTEL YOGYAKARTA PADA SITUS TRIPADVISOR

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana
Jurusan Statistika

Siti Rahmi Kurniasari
14 611 227

JURUSAN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS ISLAM INDONESIA
YOGYAKARTA
2018

HALAMAN PERSETUJUAN PEMBIMBING

TUGAS AKHIR

Judul : Implementasi SVM dan Asosiasi untuk Sentiment Analysis Data Ulasan The Phoenix Hotel Yogyakarta Pada Situs TripAdvisor

Nama Mahasiswa : Siti Rahmi Kurniasari

Nomor Mahasiswa : 14 611 227

TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK DIUJIKAN

Yogyakarta, 23 Maret 2018

Pembimbing

(Ayundyah Kesumawati, S.Si., M.Si.)

HALAMAN PENGESAHAN

TUGAS AKHIR

IMPLEMENTASI SVM DAN ASOSIASI UNTUK SENTIMENT ANALYSIS DATA ULASAN THE PHOENIX HOTEL YOGYAKARTA PADA SITUS TRIPADVISOR

Nama Mahasiswa : Siti Rahmi Kurniasari

Nomor Mahasiswa : 14 611 227

TUGAS AKHIR INI TELAH DIUJIKAN

PADA TANGGAL 23 Maret 2018

Nama Penguji	**Tanda tangan**
1. Ir. Ali Parkhan, M.T.	.....
2. Tuti Purwaningsih, S.Stat., M.Si.	.....
3. Ayundyah Kesumawati, S.Si., M.Si.	.....

Mengetahui,

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Drs. Allwar, M.Sc., Ph.D

KATA PENGANTAR

Assalamu'alaikum Warahmatullaahi Wabarakaatuh

Alhamdulillaahirabbil'aalamiin, puji syukur kehadirat Allah SWT yang telah melimpahkan rahmat dan hidayah-Nya selama melaksanakan Tugas Akhir sehingga dapat terselesaikan. Shalawat serta salam tercurah kepada Nabi Muhammad SAW beserta keluarga dan para pengikut-pengikutnya. Tugas akhir yang berjudul "Implementasi SVM dan Asosiasi untuk Sentiment Analysis Data Ulasan The Phoenix Hotel Yogyakarta Pada Situs TripAdvisor" ini sebagai salah satu syarat untuk memperoleh gelar sarjana Jurusan Statistika di Universitas Islam Indonesia. Selama menulis Tugas Akhir, penulis telah banyak mendapat bimbingan dan bantuan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis bermaksud menyampaikan ucapan terima kasih kepada:

1. Kedua orang tua tercinta, Bapak Dr. Abrar, M.Pd dan Ibu Srinangsi Makalalag, S.Pd., M.Pd. yang selalu memberikan semangat, doa, dan dukungan disetiap langkah penulis.
2. Adik penulis, Darwan Abrar yang sedang menempuh kuliah di Fakultas Hukum Universitas Islam Indonesia yang selalu mendukung dan memberi semangat, serta Keluarga Besar yang selalu setia menemani dan mendoakan yang terbaik.
3. Bapak Drs. Allwar, M.Sc, Ph.D selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia, Yogyakarta beserta seluruh jajarannya.
4. Bapak Dr. RB. Fajriya Hakim, S.Si, M.Si, selaku Ketua Jurusan Statistika beserta seluruh jajarannya.
5. Ibu Ayundyah Kesumawati, S.Si, M.Si, yang sangat berjasa dalam penyelesaian Tugas Akhir ini dan selalu memberi bimbingan selama penulisan Tugas Akhir ini.

6. Dosen-dosen Statistika Universitas Islam Indonesia yang telah mendidik dan memberikan ilmunya kepada penulis serta selalu menginspirasi.
7. Keluarga di Yogyakarta, Dr. Rasida Hatam, M.Si. selaku orang tua wali, terima kasih atas semua kebaikan, bimbingan, pelajaran, dan yang banyak membantu selama masa studi di Universitas Islam Indonesia semoga semua urusan dipermudahkan oleh Allah SWT.
8. Keluarga Besar mutiara-mutiara Bunda Hj. Uga Wiranto, S.H., M.Sc. (SPIT10NEDI), kalian semua istimewa, yang selalu memberikan doa, semangat, berbagi cerita dan pengalaman. Semoga tali silaturahmi di antara kita tidak pernah terputus.
9. Teman-teman satu bimbingan tugas akhir (bimbingan Ibu Ayun) Suci, Reny, Dwi, Ayu, Maulida, Elin, Riza, Molydah, Gustiara, Syauqi, Afifah, Achmad, Elsa, Dian, dan Yayan yang selalu berbagi ilmu, berbagi cerita, dan pengalaman.
10. Sahabat-sahabat yang luar biasa hebatnya dan teman seperjuangan, Annisa, Dhea, Ditia, Dwi, Khusnul, Maulida, Rabi, Reny, Zarina, Ayu yang selalu berbagai ilmu dan pengalaman.
11. Sahabat Statistika 2014, keluarga besar Ikatan Keluarga Statistika (IKS) yang sudah banyak memberikan semangat dan dukungan selama penulisan tugas akhir ini.
12. Kakak-kakak dan Adik-adik Purna Wirabhakti Forum Komunikasi Alumni Daerah Istimewa Yogyakarta - Jawa Tengah, yang selalu memberikan doa, dukungan, dan semangat selama penulisan tugas akhir ini.
13. Teman-teman KKN unit 123 Dusun Kalimanggis dan Dusun Gowok, Desa Wadas, Kecamatan Bener, Purworejo, Naufal, Dixi, Fira, Tata, Dila, Panji, Anwar, Indra, suka dan duka yang telah dilalui bersama tidak akan pernah terlupakan.
14. Semua pihak yang tidak dapat penulis sebutkan satu per satu, terima kasih.

Penulis menyadari bahwa tugas akhir ini masih jauh dari sempurna, oleh karena itu segala kritik dan saran yang bersifat membangun selalu penulis harapkan. Semoga tugas akhir ini dapat bermanfaat bagi penulis khususnya dan bagi semua yang membutuhkan. Akhir kata, semoga Allah SWT selalu melimpahkan rahmat serta hidayah-Nya kepada kita semua, Amin amin ya robbal „alamiin

Wassalamu'alaikum Warahmatullaahi Wabarakaatuh.

Yogyakarta, 6 Februari 2018

Siti Rahmi Kurniasari

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya yang sebelumnya pernah diajukan untuk memperoleh gelar kesarjanaan di suatu perguruan tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 6 Februari 2018

Penulis

IMPLEMENTASI SVM DAN ASOSIASI UNTUK SENTIMENT ANALYSIS DATA ULASAN THE PHOENIX HOTEL PADA SITUS TRIPADVISOR

Siti Rahmi Kurniasari
Program Studi Statistika Fakultas Matematika dan Ilmu Penetahuan Alam
Universitas Islam Indonesia

INTISARI

Web scraping digunakan untuk mendapatkan data secara online pada halaman website yaitu mengumpulkan data ulasan pengunjung The Phoenix Hotel Yogyakarta - MGallery Collection yang bersumber dari situs TripAdvisor. Proses pengklasifikasian data ulasan akan dilakukan dengan machine learning menggunakan metode Support Vector Machine (SVM). Selanjutnya hasil klasifikasi akan dianalisis dengan metode Text Mining, konsep utamanya adalah dengan melakukan ekplorasi seluas-luasnya dan ekstraksi dengan data yangsangat banyak dan terus bertambah. Sehingga ditemukan sebuah fakta dan informasi yang dianggap penting dan dapat berguna untuk berbagai bidang keperluan. Klasifikasi dengan metode SVM menunjukkan tingkat akurasi sebesar 96,07% pada ulasan berbahasa Inggris dan sebesar 84,77% untuk ulasan berbahasa Indonesia. Secara umum, dengan metode text mining diperoleh informasi bahwa lebih banyak pengunjung yang memberikan penilaian positif daripada pengunjung yang memberikan penilaian negatif. Pengunjung banyak memberikan penilaian positif diantaranya tentang penilaian kamar hotel yang mewah dan terdapat balkon, staff yang ramah dan membantu, menu sarapan yang disajikan secara prasmanan yang banyak variasi makanan yang dinilai sempurna dan pemilihan sebaran atau tatanan makanan yang tepat. Sedangkan beberapa penilaian negatif pengunjung daiantaranya adalah layanan yang digolongkan dikecewakan tentu menjengkelkan dan kekurangan staf dalam sisi tingkat perawatan. Selanjutnya hasil ulasan negatif tersebut dibuat dalam diagram fishbone untuk pemecahan masalah.

Kata Kunci: *Analisis Sentimen, Fishbone, Machine Learning, The Phoenix Hotel, SVM, Text Mining, TripAdvisor, Web Scraping.*

IMPLEMENTATION OF SVM AND ASSOCIATION FOR SENTIMENT ANALYSIS OF DATA REVIEW PHOENIX HOTEL YOGYAKARTA ON TRIPADVISOR SITES

Siti Rahmi Kurniasari

Statistics Department, Faculty of Mathematics and Natural Sciences

Islamic University of Indonesia

ABSTRACT

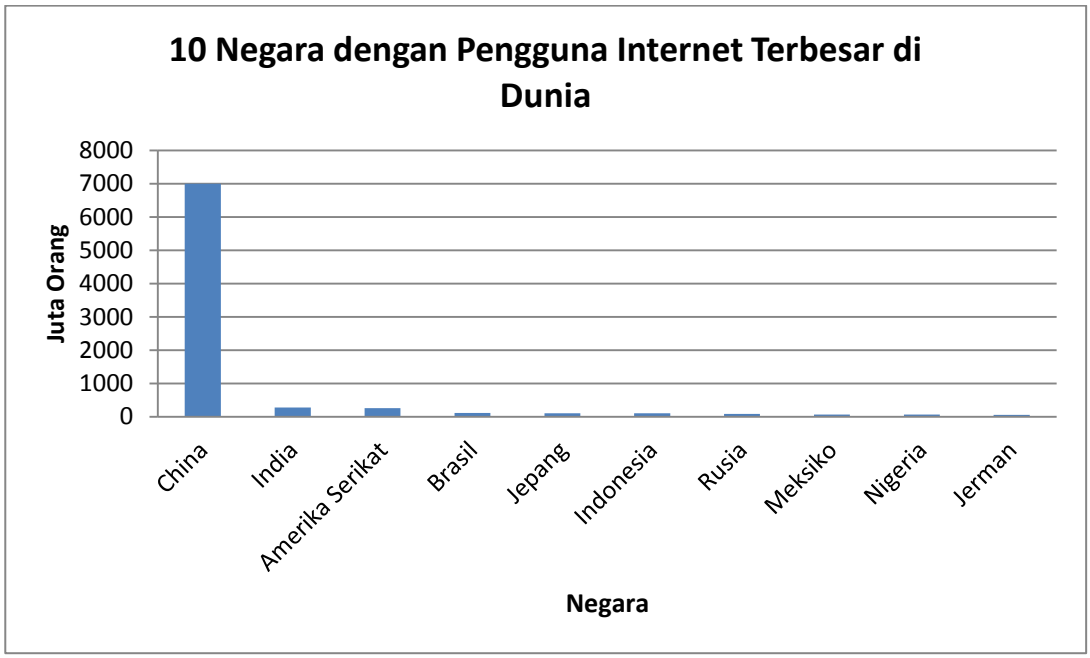
Web scraping is used to get the data online on the website page that is collecting the data of the visitor reviews The Phoenix Hotel Yogyakarta - MGallery Collection sourced from the TripAdvisor site. The process of classifying the review data will be done with machine learning using Support Vector Machine (SVM) method. Subsequently the results of classification will be analyzed by Text Mining method, the main concept is to conduct the widest exploration and extraction with data very much and continues to grow. So found a fact and information that is considered important and can be useful for various areas of need. Classification using the SVM method shows an accuracy of 96.07% in English-language reviews and 84.77% for Indonesian-language reviews. In general, with the text mining method obtained information that more visitors who provide a positive assessment than visitors who provide negative ratings. Many visitors gave positive ratings such as the assessment of luxurious hotel rooms and balconies, friendly and helpful staff, a buffet-style breakfast menu with a large variety of well-judged food and proper selection of spreads or food arrangements. While some of the negative ratings of visitors in between are the services classified as disappointing it is certainly annoying and lack of staff in terms of the level of care. Furthermore, the negative review results are made in the fishbone diagram for troubleshooting.

Keywords: *Sentiment Analysis, Fishbone, Machine Learning, The Phoenix Hotel, SVM, Text Mining, TripAdvisor, Web Scraping.*

BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Kemajuan di bidang teknologi, *computer*, dan telekomunikasi telah mendukung perkembangan teknologi internet. Internet (*interconnection network*) yang diartikan jaringan yang saling terhubung, kini telah menjadi kebutuhan primer bagi banyak orang. Penggunaan internet yang gila-gilaan di seluruh dunia bikin semua hal berubah dengan cepat di era digital ini. Dari 3,8 miliar orang, 2,9 miliar aktif menggunakan media sosial (Ismarani, 2017). Menurut data yang dikutip dari Emarketer.com dan CIA World Factbook (Ericson, 2017), 10 Negara dengan penggunaan internet terbanyak di dunia sebagimana diilustrasikan dalam bentuk **Gambar 1.1** berikut ini:



Gambar 1.1 Negara dengan Pengguna Internet Terbesar di Dunia

Meningkatnya kebutuhan akan informasi mendorong manusia untuk mengembangkan teknologi baru agar pengolahan data dan informasi dapat dilakukan dengan mudah dan cepat. Salah satunya adalah akan mempermudah dan mempercepat proses pengolahan data, mencari informasi dan lainnya (Josi, Abdillah, & Suryayusra, 2014).

Berbagai kemudahan dan manfaat yang diperoleh pelaku bisnis dari adanya internet, salah satunya adalah bagi perusahaan jasa yang bergerak di bidang pariwisata seperti hotel (Hartono, 2014). Industri perhotelan saat ini telah dimudahkan dengan banyak jasa layanan *online* yang bergerak dibidang pariwisata yang dapat membantu dalam melakukan promosi dan pemasaran. Banyaknya situs-situs *online* yang memberikan pelayanan berupa *booking online*, baik tiket pesawat maupun hotel menjadi sebuah pendorong bagi dunia pariwisata, akomodasi, dan juga bidang lainnya yang memerlukan servis pelayanan dari situs-situs tersebut. Semua bisa dilakukan dengan hanya sebuah akses dan klik melalui sebuah situs di internet. Beberapa situs layanan *online* tersebut yang terkenal diantaranya adalah *pegipegi.com, tiket.com, traveloka.com, wego.co.id, tripAdvisor.com, agoda.com, expedia.co.id*, dan lain sebagainya.

Salah satu lanyanan *online* pariwisata yang banyak digunakan masyarakat dunia saat ini adalah *TripAdvisor. TripAdvisor* adalah salah satu situs wisata terbesar di dunia yang membantu wisatawan mengoptimalkan potensi setiap perjalanan. *TripAdvisor* menawarkan sarana dari jutaan wisatawan serta berbagai pilihan dan fitur perencanaan wisata dengan *link* praktis ke alat bantu pemesanan yang memeriksa ratusan situs *web* untuk menemukan harga hotel terbaik. Situs *web TripAdvisor* merupakan komunitas wisata terbesar di dunia, menjangkau rata-rata 390 juta pengunjung untuk setiap bulannya, serta menampilkan 435 juta ulasan dan opini tentang 6,8 juta akomodasi, restoran, dan objek wisata yang beroperasi di 49 pasar di seluruh dunia (TripAdvisor, 2016).

Setiap bulan, lebih dari 260 juta konsumen mencari informasi tentang pembelian wisata menggunakan jasa *TripAdvisor.com* dan menempatkan *TripAdvisor* di peringkat 2 dunia di bawah *booking.com* (detiktravel, 2013), sehingga saat ini jasa layanan *TripAdvisor* digunakan oleh banyak hotel yang ada di dunia sebagai media pemasaran untuk memudahkan pengunjung dalam melakukan reservasi hotel secara *online.*

Daerah yang potensial dengan sektor pariwisata di Indonesia memiliki peluang besar menjadi target destinasi banyak wisatawan, baik wisatawan domestik maupun mancanegara. Salah satu daerah di Indonesia yang terkenal

dengan potensi wisatanya adalah provinsi Daerah Istimewa Yogyakarta. Yogyakarta selain di kenal sebagai kota pendidikan juga di kenal dengan kota yang memiliki potensi besar di bidang pariwisata. Daerah Istimewa Yogyakarta yang relatif aman dan nyaman dengan keramah-tamahan masyarakatnya, menjadikan Yogyakarta banyak diminati orang/wisatawan untuk berkunjung. Setiap tahunnya jumlah kunjungan wisatawan baik wisatawan mancanegara (wisman) maupun wisatawan nusantara (wisnus) yang datang terus meningkat (Tribunjogja, 2016).

Hotel sebagai salah satu sarana akomodasi tempat menginap sementara bagi para tamu tentu sangat berperan penting dalam menampung jumlah wisatawan yang datang ke Daerah Istimewa Yogyakarta. Tanpa adanya akomodasi pariwisata, maka industri wisata akan mengalami kesulitan dalam perkembangannya. Pertumbuhan jumlah wisatawan diikuti dengan pertumbuhan jumlah hotel berbintang. Pertumbuhan hotel berbintang yang demikian pesat akan mempertajam tingkat persaingan usaha, sehingga berbagai upaya promosi dan pemasaran terus dilakukan oleh masing-masing industri perhotelan. Mengingat perkembangan teknologi yang semakin canggih, upaya promosi dan pemasaran hotel saat ini banyak dilakukan secara *online*, salah satunya dengan memanfaatkan jasa layanan *TripAdvisor*.

The Phoenix Hotel Yogyakarta - MGallery Collection merupakan hotel bersejarah di Daerah Istimewa Yogyakarta yang bekerjasama dengan *TripAdvisor* dalam melakukan promosi dan pemasarannya. Phoenix Hotel yang didirikan oleh Mr. Kwik Djoen Eng pada tahun 1918 sebagai rumah tinggal (Prasetyo, 2018) adalah hotel yang terletak di Jantung Kota Yogyakarta dengan gaya bagunan campuran Kebudayaan Jawa dan Cina. Tahun 1942 ketika Jepang datang ke Hotel tersebut dikuasai oleh Jepang dan berganti nama menjadi Yamato Hotel. Tahun 1945 Hotel kembali ke pemiliknya dan berganti nama menjadi Hotel Merdeka. Tahun 1993 Hotel itu kembali berganti nama menjadi "Phoenik Heritage Hotel" Hotel Phoenix sampai sekarang ini (Prasetyo, 2018).

Menurut *General Manager* The Phoenix Hotel (Evrard, 2016), hotelnya mempunyai kekhasan tersendiri dalam melayani tamu. Hotel bintang 5 (lima) dengan kapasitas 143 kamar tesebut mengusung budaya Yogyakarta. Tersedia 4 (empat) tipe kamar yang dapat dipilih oleh wisatawan yang menginap. Kekuatan lain dari hotel ini adalah arsitektur bangunan yang kental dengan corak *colonial* Belanda. Berdasarkan Sistem Informasi Data Statistik Hotel Yogyakarta dengan Wisman terbanyak tahun 2017, Hotel Phoenix menempati urutan ke 4 (empat) jumlah pengunjung wisman (Risyanto, 2018). Berikut ini adalah 10 (sepuluh) Hotel di Yogyakarta dengan pengunjung wisman terbanyak sepanjang tahun 2017 sebagimana diilustrasikan dalam bentuk **Gambar 1.1** berikut ini:

Gambar 1.2 Hotel dengan Wisman Terbanyak Tahun 2017 di Yogyakarta

Mempertahankan tingkat hunian dengan semakin banyak hotel yang dibagun, maka dipelukan suatu usaha untuk menarik perhatian tamu agar memilih hotel tertentu sebagai akomodasi pilihan saat berlibur. Oleh karenanya, kepuasan pelanggan juga perlu diperhatikan untuk menjaga stabilitas kedatangan wisatawan dan menjadikan mereka sebagai *repeater* atau pelanggan tetap bagi suatu hotel.

Sejalan dengan perkembangan media *online* atau teknologi yang cepat, komentar atau keluhan-keluhan tamu saat ini bisa dibuat dan dilihat oleh banyak orang melalui media sosial maupun media *online* lainnya (Wati, 2015), salah satunya adalah *TripAdvisor*. Situs *TripAdvisor* menyediakan berbagai fasilitas yang dapat memudahkan pengunjung untuk memperoleh informasi secara rinci tentang hotel yang akan dikunjungi, baik lokasi, jumlah kamar, dan berbagai fasilitas lain yang tersedia. Selain itu, situs ini banyak dikunjungi karena menyediakan informasi mengenai ulasan-ulasan para wisatawan. Setiap pengunjung dapat memberikan ulasan baik berupa saran, kritik, maupun penilaian lain terhadap hotel yang pernah dikunjungi. Dengan adanya fasilitas tersebut, calon pengunjung hotel akan lebih mudah mendapatkan informasi hotel berdasarkan ulasan-ulasan pengunjung sebelumnya, sehingga dapat dijadikan referensi untuk mengambil keputusan dalam memilih hotel yang sesuai dengan keinginannya.

Selain bermanfaat untuk pengunjung, ulasan-ulasan tersebut juga dapat bermanfaat bagi pihak hotel untuk mengetahui kepuasan pelanggan dengan cara melihat persepsi pengunjung melalui komentar atau ulasan secara *online*, yakni melalui ulasan positif maupun negatif, sehingga hotel tidak perlu melakukan survei kepuasan pelanggan secara manual.

Jumlah data ulasan pengunjung yang masuk ke situs *TripAdvisor* terus bertambah seiring berjalannya waktu, hal ini mengakibatkan sulitnya pihak hotel dalam memperoleh informasi secara keseluruhan dari semua ulasan, karena akan membutuhkan waktu yang lama untuk membaca satu persatu setiap ulasan yang masuk pada halaman situs *TripAdvisor*. Dengan banyaknya data ulasan yang masuk, maka diperlukan sebuah teknik khusus untuk mengumpulkan data secara *online* dan sebuah analisis yang dapat membantu dalam memperoleh informasi dari sejumlah data dalam skala besar. *Web scraping* menjadi solusi alternatif yang dapat digunakan untuk mendapatkan data secara *online* pada halaman *website* sebelum data tersebut diolah dan dianalisis.

Web Scarping adalah proses pengambilan sebuah dokumen semi-terstruktur dari Internet, umumnya berupa halaman *web* dalam bahasa *markup* seperti *HTML* atau *XHTML*, dan menganalisis dokumen tersebut untuk diambil

data dan dipergunakan untuk kepentingan lainnya (Turland, 2010). Pada penelitian ini, *web scraping* digunakan untuk mengumpulkan data ulasan pengunjung The Phoenix Hotel Yogyakarta-MGallery Collection yang bersumber dari situs *TripAdvisor*.

Data ulasan pengunjung dari situs *TripAdvisor* dapat dijadikan sebagai sumber informasi untuk mengetahui opini dan persepsi pengunjung terhadap hotel. Dengan mengetahui persepsi pengunjung, pihak hotel dapat mengetahui apa yang menjadi kelebihan dan kekurangan dari hotel tersebut. Pihak hotel akan terus berupaya mempertahankan kualitas jika dinilai baik dan dapat melakukan evaluasi perbaikan dan peningkatan jika terdapat penilaian yang buruk dari konsumen, sehingga diharapkan okupansi dapat terus meningkat. Oleh sebab itu, pihak hotel juga perlu mengetahui dan memahami segala bentuk penilaian pengunjung terhadap hotel, semua hal tentang hotel yang paling sering menjadi pokok bahasan dalam ulasan akan menjadi informasi penting bagi pihak hotel sebagai acuan dalam upaya meningkatkan kualitas dan pelayanan hotel. Dengan menggunakan beberapa metode *data mining*, penulis mencoba melakukan klasifikasi data ulasan pengunjung berdasarkan sentimen positif dan sentimen negatif.

Klasifikasi data ulasan berdasarkan jenis sentimen akan mempermudah pihak hotel dalam mendapatkan informasi persepsi pengunjung. Proses pengklasifikasian data ulasan akan dilakukan dengan *machine learning* menggunakan metode *Support Vector Machine (SVM). SVM* adalah metode *machine learning* yang bekerja atas prinsip *Structural Risk Minimization (SRM)* dengan tujuan menemukan *hyperplane* terbaik yang memisahkan dua buah kelas pada *input space* (Susilowati, 2015). Pemilihan metode *SVM* didasarkan pada beberapa penelitian terdahulu yang membuktikan bahwa *SVM* memiliki performansi yang cukup baik dalam melakukan klasifikasi dokumen berupa teks. Setelah melakukan klasifikasi, penulis mencoba mengekstrak dan mengeksplorasi seluas-luasnya informasi apa yang ada pada setiap klasifikasi sentimen positif maupun sentimen negatif yang sekiranya dianggap penting untuk digunakan pada berbagai keperluan. Pada proses ektraksi dan eksplorasi informasi, penulis

menggunakan statistik deskriptif dan asosiasi antar kata untuk menemukan topik yang akan sering dibicarakan oleh pengunjung.

Karena itu peneliti berharap, penelitian ini mampu mengklasifikasikan teks ulasan dengan baik sehingga nantinya informasi yang diperoleh di dalamnya dapat diekstraksi dengan baik serta penyajian informasi dari data yang diamati dapat memberikan informasi yang berguna bagi pihak hotel dan pihak-pihak lain yang membutuhkannya.

1.2 Rumusan Masalah

Berdasarkan latar belakang, maka permasalahan yang akan dikaji dalam penelitian ini adalah sebagai berikut:

1. Bagaimana cara mengimplementasikan teknik *web scraping* untuk mendapatkan data ulasan The Phoenix Hotel Yogyakarta-MGallery Collection dari situs *TripAdvisor*?
2. Sejauh mana gambaran umum data ulasan tentang The Phoenix Hotel Yogyakarta - MGallery Collection berdasarkan situs *TripAdvisor*?
3. Bagaimana implementasi metode *SVM* dalam mengklasifikasikan sentimen pada data ulasan The Phoenix Hotel Yogyakarta - MGallery Collection?
4. Informasi apa saja yang didapatkan dalam setiap klasifikasi yang telah dilakukan dengan menggunakan metode *SVM* pada The Phoenix Hotel Yogyakarta - MGallery Collection?
5. Berdasarkan diagram *fishbone,* faktor-faktor apa saja yang harus di lakukan untuk memperbaiki hasil dari ulasan negatif yang didapat?

1.3 Batasan Masalah

Untuk menghindari permasalahan yang lebih luas dan agar tujuan pembahasan semakin terarah maka dilakukan batasan masalah dalam penulisan ini dibatasi sebagai berikut:

1. Penelitian ini menggunakan data ulasan pengunjung The Phoenix Hotel Yogyakarta-MGallery Collection berbahasa Inggris dan berbahasa Indonesia

yang bersumber dari *website www.tripadvisor.com* yang tercatat sejak 2011 hingga tahun 2017.

2. Penulis menggunakan bantuan *software Microsoft Excel* 2016, *R 3.4.3* dan *Xmind 7* untuk melakukan proses analisis data.

1.4 Tujuan Penelitian

Menjawab permasalahan penelitian tersebut, maka penelitian ini mempunyai tujuan secara umum dikemukakan sebagai berikut:

1. Mengimplementasikan teknik *Web Scraping* untuk mendapatkan data ulasan The Phoenix Hotel Yogyakarta - MGallery Collection dari situs *TripAdvisor*.
2. Mengetahui gambaran umum data ulasan tentang The Phoenix Hotel Yogyakarta - MGallery Collection berdasarkan situs *TripAdvisor*.
3. Mengimplementasikan metode *SVM* dalam mengklasifikasikan sentimen pada data ulasan pengunjung The Phoenix Hotel Yogyakarta - MGallery Collection.
4. Mendapatkan informasi penting dalam setiap klasifikasi pada The Phoenix Hotel Yogyakarta - MGallery Collection.
5. Mendapatkan informasi faktor-faktor apa saja yang harus di lakukan untuk memperbaiki hasil dari ulasan negatif yang didapat

1.5 Manfaat Penelitian

Hasil penelitian ini diharapkan dapat memberikan manfaat kepada pihak-pihak yang terkait. Adapun manfaat yang diharapkan antara lain:

1. Memperoleh data ulasan dari situs *TripAdvisor* dengan menggunakan teknik *web scraping*.
2. Mengetahui gambaran secara umum data ulasan tentang The Phoenix Hotel Yogyakarta - MGallery Collection berdasarkan situs *TripAdvisor*.

3. Pengklasifikasian data ulasan tentang The Phoenix Hotel Yogyakarta-MGallery Collection untuk memudahkan pihak hotel dalam mengetahui persepsi pengunjung dalam bentuk opini negatif dan opini positif, sehingga dapat dijadikan sebagai acuan dalam upaya menjaga kualitas dan memperbaiki kekurangan serta evaluasi ke arah yang lebih baik.

BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian tentang Teknik *Web Scraping* dan Klasifikasi Sentimen Menggunakan Metode *Support Vector Machine* dan Asosiasi sudah banyak dilakukan oleh peneiti-peneliti sebelumnya. Penelitian terdahulu sangatlah penting bagi penulis sebagai kajian untuk mengetahui keterkaitan antara penelitian terdahulu dengan penelitian yang akan di lakukan, untuk menghindari terjadinya tindakan duplikasi yang dilakukan oleh penulis. Tujuan dari tinjauan pustaka ini adalah untuk menunjukkan bahwa penelitian yang dilakukan penulis sangatlah bermanfaat dan mempunyai arti penting sehingga dapat diketahui kontribusi penelitian terhadap ilmu pengetahuan. Berikut beberapa ulasan tentang penelitian-penelitian terdahulu yang pernah dilakukan sebelumnya berkenaan dengan data dan metode yang digunakan. Beberapa jurnal dan penelitian yang penulis jadikan sebagai acuan adalah sebagai berikut.

Penelitian tentang Hotel Royal Ambarrukmo sebelumnya pernah dilakukan oleh Puspita (2016), dalam tugas akhirnya penelitian ini membahas tentang strategi bauran produk yang diterapkan di Hotel Royal Ambarrukmo dan mengetahui kesesuaian produk hotel dengan tamu. Penelitian ini merupakan penelitian deskriptif kualitatif dengan tiga informan tamu bisnis, empat staff Hotel Royal Ambarrukmo, dan satu orang abdi dalem Ambarrukmo sebagai narasumber yang memiliki keterkaitan dengan bahasan penelitian. Hasil penelitian menunjukan bahwa tamu bisnis yang berkunjung ke hotel Royal Ambarrukmo terdiri dari dua jenis tamu yaitu tamu bisnis perseorangan dan tamu bisnis kelompok. Melihat dari hasil wawancara, dapat diambil kesimpulan bahwa tamu bisnis mengelompokkan kamar sebagai *core products* (produk inti). Sedangkan produk hotel yang lain, pengelompokannya berbeda-beda sesuai dengan kebutuhan tamu.

Penelitian yang dilakukan oleh Abtohi (2017) menerapkan teknik *web scraping* dan klasifikasi sentimen dengan menggunakan metode *Support Vector Machine*. Klasifikasi sentimen kemudian dianalisis dengan metode *Text Mining* menggunakan asosiasi kata, konsep utamanya adalah dengan melakukakan ekplorasi seluas-seluasnya dan ekstraksi informasi dari sejumlah data yang cukup besar. Klasifikasi dengan metode *SVM* menunjukkan tingkat akurasi sebesar 95,27% pada ulasan berbahasa Inggris dan sebesar 95,00% untuk ulasan berbahasa Indonesia. Dari penelitian tersebut disimpulkan bahwa teknik *web scraping* dapat digunakan sebagai cara alternatif untuk mendapatkan dari halaman *website* dan dapat mempermudah serta mempercepat proses pegambilan data dalam skala besar secara otomatis di internet.

Dewantoro (2016) dalam Tugas Akhirnya menggunakan teknik *web scraping* pada proses *topic modeling*. Proses *scraping website* dilakukan dengan menggunakan bahasa pemrograman Python 2.7.11. Teknik scraping digunakan untuk mengambil artikel dari portal berita yang kemudian diolah menggunakan topic modeling. *Topic modeling* yang digunakan adalah *Latent Dirichlet Allocation (LDA). LDA* adalah model probabilitas generatif dari koleksi data diskrit seperti kumpulan-kumpulan teks. Pada proses scraping, perancangan sistem dilakukan dengan identifikasi kelas *tag HTML*. *Tag HTML* yang digunakan yaitu tag yang mengapit judul, isi dan tanggal berita untuk kemudian dibuatkan *scraping template*. Data yang diperoleh kemudian diolah dengan pemodelan *LDA*, sehingga dapat diketahui topik-topik yang sering muncul dari portal berita nasional. Sistem yang dibuat dapat memproses *web scraping* dari portal berita Kompas dan kemudian disimpan ke *file CSV*.

Penelitian tentang analisis sentimen *review* hotel sebelumnya pernah dilakukan oleh Indrayuni (2016). Dalam penelitiannya digunakan metode *Support Vector Machine (SVM)* untuk mengklasifikasikan ulasan positif dan negatif. Selanjutnya digunakan metode *Particle Swarm Optimization (PSO)* sebagai seleksi fitur untuk meningkatkan akurasi analisis sentimen. Evaluasi dilakukan menggunakan 10-*fold cross validation*. Hasil penelitian menunjukkan peningkatan nilai akurasi sebesar 5.61% untuk algoritma *Support Vector Machine*

dari 91.33% menjadi 96.94% setelah penerapan seleksi *fitur Particle Swarm Optimization*.

Elango dan Narayanan (2014) melakukan penelitian analisis sentimen terhadap data ulasan hotel yang diperoleh dari situs *TripAdvisor* menggunakan *machine learning* dengan pendekatan beberapa model probabilistik yaitu *Naïve Bayes* (NB), *Support Vector Machine* (SVM), *Laplace Smoothing* dan *Semantic Orientation* (SO) yang digunakan untuk mengklasifikasikan ulasan. Hasil penelitian menunjukkan bahwa model *Naïve Bayes* mampu memprediksi lebih baik daripada metode *SVM*.

Rianto (2016) melakukan penelitian untuk melakukan analisis sentimen terhadap gubernur DKI Jakarta. Sumber data yang digunakan berasal dari situs berita detik.com dan kompas.com yang selanjutnya digunakan menjadi pembelajaran model klasifikasi sentimen analisis. Dengan variasi-variasi proses prapemprosesan didapatkan model yang terbaik yaitu dengan menggunakan *SVM* dengan proses prapemprosesan *cleansing, case folding* dan eliminasi KBBI. Dengan performa *precision* sebesar 65,61%, *recall* sebesar 65,36% dan *F-measure* 65,06%.

Nugeraha (2015) melakukan penelitian menggunakan *machine learning* untuk melakukan klasifikasi teks pornografi berbahasa indonesia. Metode penelitian yang digunakan adalah *Support Vector Machine* (SVM) dan *Naïve Bayes Classifier (NBC)* yang banyak penelitian klasifikasi teks menunjukkan performa yang baik, dengan jumlah data yang digunakan pada penelitian adalah sebanyak 200 data latih dan 186 data uji. Penelitian tersebut berupaya membandingkan dan memilih model klasifikasi yang lebih akurat daripada penelitian sebelumnya. Terdapat banyak faktor yang mempengaruhi tingkat akurasi, beberapa diantaranya adalah koleksi data, banyak fitur yang digunakan, *corpus category*, metode pra-proses, dan pemilihan algoritma klasifikasi. Penelitian ini juga melakukan pengujian berbagai metode pra-proses teks. Hasil penelitian menunjukkan bahwa akurasi klasifikasi dipengaruhi oleh metode pra-proses yang diterapkan. Hal ini berlaku baik pada metode klasifikasi *SVM* ataupun *NBC*. Akurasi klasifikasi tertinggi yang dihasilkan adalah sebesar 97.85%.

Ulwan (2016) melakukan penelitian menggunakan *machine learning* dengan metode *Support Vector Machine (SVM)* mengklasifikasikan data teks laporan masyarakat yang diperoleh dari situs LAPOR!. Data teks yang tidak terstruktur (*unstructured data*) tersebut diklasifikasikan menjadi tiga kelas yaitu Aspirasi, Keluhan, dan Pertanyaan. Selanjutnya hasil klasifikasi dianalisis dengan metode *Text Mining.* Hasil klasifikasi menunjukkan tingkat akurasi sebesar 96.7%. Secara umum metode *Text Mining* menunjukkan hasil ekstraksi informasi pada kelas aspirasi adalah terkait penertiban terhadap psk, pkl, asap, merokok, *busway,* dan pembagian bantuan masyarakat. Pada kelas keluhan masyarakat mengeluhkan tentang pembagian BLSM atau KPS yang tidak merata, masalah macet, layanan Telkom yang buruk, serta *busway* yang sering bermasalah. Sedangkan pada kelas pertanyaan yang menjadi hal yang sering ditanyakan adalah masalah BLSM dan KPS serta seputar informasi mengenai agama, BPJS, beasiswa, sertifikasi, dan tunjangan.

Penelitian yang dilakukan oleh Nur dan Santika (2011) menerapkan pendekatan *SVM* dalam melakukan klasifikasi sentimen pada dokumen berbahasa Indonesia ke dalam kelas positif dan negatif. Data yang digunakan berasal dari sumber situs jejaring sosial *Twitter* berbahasa Indonesia. Untuk mendapatkan data sentimen dilakukan dengan menggunakan kata kunci berupa ikon emosi “:)” dan “:(“. Proses analisis klasifikasi dilakukan dengan menggunakan aplikasi *open source* WEKA. Hasil penelitian diperoleh kesimpulan bahwa *SVM* memiliki tingkat ketelitian yang lebih baik dengan rata-rata persentase kebenaran sebesar 73,70% dibandingkan dengan *Naïve Bayes* yang mempunyai rata-rata persentase kebenaran sebesar 67,24%.

Kurniawan, et al. (2012) melakukan penelitian terhadap artikel berita di media *website* yang mengelompokkan berita secara manual. Hal ini akan menjadi masalah apabila jumlah artikel berita yang akan dimuat di *website* memiliki jumlah yang banyak, karena dapat memakan waktu dan tenaga untuk mengelompokkannya. Sehingga diperlukan sebuah sistem yang dapat mengelompokkan artikel berita itu secara otomatis. Metode yang digunakan yaitu *Text Mining* dan *Naïve Bayes Classifier*. *Text mining* digunakan untuk

menemukan pola dari data berupa teks sehingga bisa dilakukan analisis lebih lanjut. Sedangkan *Naïve Bayes Classifier* digunakan untuk mengelompokkan artikel berita tadi sesuai dengan isi dari artikel tersebut. Penelitian ini menghasilkan pengelompokkan yang cukup akurat menggunakan metode *Naïve Bayes Classifier.*

Penelitian yang akan dilakukan adalah mengimplementasikan teknik *web scraping* untuk mengumpulkan data ulasan pengunjung The Phoenix Hotel Yogyakarta dari situs *TripAdvisor* untuk kemudian dilakukan analisis klasifikasi sentimen positif dan negatif menggunakan metode *Support Vector Machine (SVM).* Setelah diperoleh hasil klasifikasi, kemudian dilakukan proses visualiasasi, eksplorasi dan ekstraksi informasi dengan pendekatan *text mining* menggunakan asosiasi kata. Pada **Tabel 2.1** menjadi perbandingan penelitian sebelumnya dan penelitian yang akan dilakukan oleh penulis.

Tabel 2.1 Perbandingan penelitian sebelumnya dengan penelitian yang penulis lakukan.

JUDUL PENELITIAN	NAMA PENELITI, TAHUN	METODE	TUJUAN
Analisis Kesesuain Bauran Produk dengan Karateristik Tamu Bisnis di Hotel Royal Ambarrukmo Yogyakarta	Puspita, M., 2016	Deskriptif Kualitatif	Menganalisis kesesuaian produk yang disediakan pihak hotel terhadap karateristik tamu pengunjung
Implementasi Teknik *Web Scraping* dan Klasifikasi Sentimen	Abtohi, Slamet.,2017	*Support Vector Machine dan Assosiasi*	Menerapkan teknik *web scraping* dalam mengumpulkan data ulasan dan melakukan klasifikasi berdasarkan sentimen positif dan negatif
Implementasi Teknik *Web Scraping* pada proses *Topic Modeling* Portal Berita	Dewantoro,P. R., 2016	*Web Scraping* dan *LDA*	Mengetahui topik-topik pada portal berita dan mengidentifikasi topik yang paling sering muncul secara

			cepat
Analisa Sentimen *Review* Hotel Menggunakan Algoritma *Support Vector Machine* Berbasis *Particle Swarm Optimization*	Indrayuni, Elly., 2016	SVM, PSO	Mengklasifikasikan ulasan hotel berdasarkan sentimen
Sentiment Analysis for Hotel Reviews	Elango, V., dan Narayanan, G., 2014	*Naïve Bayes SVM, Laplace Smoothing Semantic Orientatiton*	Mengklasifikasikan ulasan hotel berdasarkan sentimen
Implementasi Perbandingan Metode Prapemprosesan Pada Analisis Sentimen Gubernur DKI Jakarta Menggunakan Metode *Support Vector Machine dan Naïve Bayes*	Rianto, B., 2016	*SVM, Naïve Bayes*	Menganalisis sentiment masyarakat terhadap Gubernur DKI Jakarta, Ahok pada Media Massa *online*
Proses Klasifikasi Teks Pornografi Berbahasa Indonesia berbasis *Machine Learning*	Nugeraha, F., 2015	*SVM, NBC*	Mengklasifikasi teks yang mengandung unsur pornografi
Pattern Recognition pada *Unstructured* Data Teks Menggunakan *Support Vector Machine dan Association*	Ulwan, M., 2016	*SVM,* Asosiasi Kata	Mengklasifikasi laporan masyarakat berdasarkan keluhan, aspirasi, dan pertanyaan pada situs LAPOR! dan ekstraksi informasi
Analisis Sentimen pada Dokumen Berbahasa Indonesia dengan Pendekatan *Support Vector Machine*	Nur, M., Y. dan Santika, D., 2011	*SVM, Naïve Bayes*, Pembobotan *(TP, TF, TF-IDF)*	Mengklasifikasi dokumen berbahasa Indonesia ke dalam sentiment positif dan negatif.
Klasifikasi Konten Berita dengan Metode *Text Mining*	Kurniawan, et.al., 2016	*Nave Bayes Classifer, Text Mining*	Mengklasifikasi konten berita berdasarkan kategori politik, ekonomi, olahraga dan *entertainment*

Penelitian yang penulis lakukan			
Implementasi *SVM* dan Asosiasi untuk *Sentiment Analysis* Data Ulasan The Phoenix Hotel Yogyakarta – MGallery Collection pada Situs TripAdvisor	Kurniasari, S., R. 2018	*Web Sraping, SVM* dan Asosiasi Kata	Menerapkan teknik *Web Scraping* dalam mengumpulkan data ulasan dan melakukan klasifikasi berdasarkan *sentiment* positif dan negatif.

BAB III
LANDASAN TEORI

3.1 Definisi Hotel

Hotel adalah suatu perusahaan yang dikelola oleh pemiliknya dengan menyediakan pelayanan makanan, minuman, dan fasilitas kamar untuk tidur kepada orang-orang yang sedang melakukan perjalanan dan mampu membayar dengan jumlah yang wajar sesuai dengan pelayanan yang diterima tanpa adanya perjanjian khusus (Sulastiyono, 2011).

Kata hotel memilki pengertian atau definisi yang cukup banyak, masing–masing orang berbeda dalam menguraikannya. Berikut ini adalah beberapa pengertian hotel: 1) Menurut Menteri Perhubungan, hotel adalah suatu bentuk akomodasi yang dikelola secara komersial, disediakan bagi setiap orang untuk memperoleh pelayanan penginapan berikut makan dan minum, 2) Menurut AHMA (*American Hotel and Motel Associations*), hotel adalah suatu tempat dimana disediakan penginapan, makanan, dan minuman, serta pelayanan lainnya, untuk disewakan bagi para tamu atau orang – orang yang tinggal untuk sementara waktu, 3) Menurut Webster, hotel adalah suatu bangunan atau lembaga yang menyediakan kamar untuk menginap, makanan dan minuman serta pelayanan lainnya untuk umum.

Dengan mengacu pada pengertian diatas dan untuk menertibkan perhotelan Indonesia, pemerintah menurunkan peraturan yang dituangkan dalam Surat Keputusan Menparpostel No.KM/37/PW.340/MPPT-86, tentang peraturan usaha dan penggolongan hotel, Bab 1, Ayat (b) dalam SK tersebut menyebutkan bahwa : “Hotel adalah suatu jenis akomodasi yang dipergunakan sebagian atau seluruh bangunan untuk menyediakan jasa penginapan, makanan dan minuman serta jasa penunjang lainnya bagi umum dan dikelola secara komersial” (Yunus, 2014).

3.2 The Phoenix Hotel Yogykarta - MGallery Collection

The Phoenix Hotel Yogyakarta - MGallery Collection terletak di Jalan Jenderal Sudirman no 9, Cokrodiningratan, Jetis, Kota Yogyakarta, Daerah Istimewa Yogyakarta 55233, Indonesia. Jalur utama yang berjarak 9,3 km dari Bandara Adi Sutjipto dan bisa ditempuh hanya membutuhkan waktu sekitar 30 menit sehingga lokasi hotel ini sangat mudah untuk di cari dan di temukan.

Hotel ini memiliki sejarah penuh warna sendiri. Bangunan utama awalnya dibangun pada tahun 1918 oleh Kwik Djoen Eng sebagai tempat tinggal pribadi. Dia adalah pendiri dan pemilik perusahaan perdagangan yang sukses, yang didirikan di Semarang pada tahun 1877. Ketika resesi ekonomi di tahun 1930an, rumahnya dilelang dan dibeli oleh Liem Djoen Hwat. Segera setelah itu, itu diubah menjadi hotel pensiunan oleh D.N.E. Franckle, seorang pengusaha Belanda yang menyewa rumah dari Hwatt. The Splendid Hotel-seperti bagaimana kemudian disebut ada sampai 1942 ketika Jepang menduduki Indonesia dan mengambil alih hotel. Laporan tidak resmi mengatakan bahwa itu terus berjalan sebagai hotel dengan nama Hotel Yamato.

Setelah keberangkatan Jepang di tahun 1945, hotel tersebut diberikan kembali kepada Liem Djoen Hwatt sebagai pemilik sah. pada tahun 1946-1949, Yogyakarta adalah ibu kota Republik Indonesia. Saat itu, rumah ini digunakan sebagai kediaman resmi Konsul China. Sejak 1951, rumah itu disewa oleh Direktorat Negara dan Pariwisata Negara dan difungsikan sebagai hotel dengan nama Hotel Merdeka, kata Kebebasan Indonesia. Pada tahun 1988, putra agung Hwatt-Sulaeman-mengambil kembali hotel dan memutuskan untuk terus menjalankannya sebagai hotel. Dia merenovasi bangunan berusia 70 tahun itu dan mengembalikannya ke bentuk aslinya. Selain itu, ia juga menambahkan gedung baru di sisi utara dan timur bangunan utama. Pada tahun 1993, ia membuka pintu pertamanya kepada pelanggan dengan nama baru Phoenix Heritage Hotel.

Bangunan asli hotel ini dinyatakan sebagai bangunan peninggalan pada tahun 1996 karena merupakan contoh sempurna dari Arsitektur Indische abad ke-19 yang terkenal, campuran Art Nouveau Eropa dengan kepentingan Kepulauan, Jawa dan unsur-unsur budaya Cina.

Setelah sepuluh tahun beroperasi, kepemilikan hotel kembali berubah. Imelda Sundoro Hosea, pengusaha yang berbasis di Solo, membeli hotel tersebut pada tahun 2003 dan telah berkomitmen untuk mempertahankan fungsinya sebagai hotel warisan. Tujuannya adalah untuk meningkatkannya sebagai hotel bintang 5 dan dijalankan oleh jaringan hotel internasional, oleh karena itu dia menugaskan Accor, jaringan hotel yang dapat diandalkan di seluruh dunia untuk mengambil alih manajemen. Setelah direnovasi secara ekstensif dan setelah ditutup selama hampir setahun, hotel tersebut dibuka kembali pada bulan Mei 2004 sebagai Grand Mercure Yogyakarta, hotel butik bintang 5 dengan 144 kamar. Pada tanggal 30 Maret 2009, secara resmi dirombak sebagai The Phoenix Hotel Yogyakarta dengan label MGallery sebuah koleksi hotel kelas atas Accor.

3.3 *Website*

Menurut Maulana Ilmar dalam skripsinya yang berjudul "Analisis dan Perancangan Sistem Informasi Berbasis *Website* pada SMA Negeri 1 Pemalang", menyatakan bahwa *website* merupakan halaman yang akan digunakan pada tampilan informasi, gerak pada gambar, video maupun gabungan dari keseluruhan tersebut untuk sifat yang tetap (statis) dan juga yang berubah-ubah (dinamis), akan membentuk suatu rangkaian yang saling terkait, dan dihubungkan menggunakan *link*. Pada desain *website*-nya menggunakan beberapa *software* aplikasi (perangkat lunak) seperti bahasa pemrograman *PHP* dan *HTML* (*Hypertext Markup Language*), *MySQL* yang digunakan untuk mengakses *database server*.

3.4 *HTML*

HTML (Hyper Text Markup Language) adalah sebuah bahasa *formatting* yang digunakan untuk membuat sebuah halaman *website*. Didalam dunia pemrograman berbasis *website (Web Programming), HTML* menjadi pondasi dasar pada halaman *website*. Sebuah *file HTML* di disimpan dengan ekstensi .html (*dot html*) dan dapat di eksekusi atau diakses menggunakan *web browser* (*Google Chrome, Mozilla Firefox, Opera,*

Safari dan lain-lain). *HTML* memiliki beberapa elemen yang tersusun dari *tag-tag* yang memiliki fungsinya masing-masing. Seperti *tag heading,* paragraf, pembuatan *form,* tombol, *list*, membuat *hyperlink* atau *link* yang menghubungkan antar halaman *website* dan banyak lagi (Hadi, 2017).

3.5 *CSS*

CSS merupakan singkatan dari "*Cascading Style Sheets*", sesuai dengan namanya *CSS* memiliki sifat "*style sheet language*" yang berarti bahasa pemrograman yang digunakan untuk *web design*. *CSS* adalah bahasa pemrograman yang di gunakan untuk men-*design* sebuah halaman *website*. Dalam men-*design* halaman *website*, *CSS* menggunakan penanda yang di kenal dengan *id* dan *class*. *CSS* dapat mengubah font, ukuran font, warna dan format font, mengatur ukuran *layout*, lebar, tinggi dan warna element, mengubah tampilan *form*, membuat halaman *website* yang *responsive* dan masih banyak lagi yang dapat di lakukan oleh *CSS* (Hadi, 2017).

3.6 *TripAdvisor*

TripAdvisor merupakansitus wisata terbesar di dunia yang membantu wisatawan mengoptimalkan potensi penuh setiap perjalanan. *TripAdvisor* menyediakan pendapat banyak orang bagi wisatawan untuk membantu mereka memutuskan tempat menginap, maskapai penerbangan, hal yang dapat dilakukan, dan tempat makan. Dengan lebih dari 570 juta ulasan dan opini tentang pilihan terbanyak di dunia untuk daftar wisata di seluruh dunia merangkumi 7,3 juta akomodasi, maskapai penerbangan, objek wisata, dan restoran. *TripAdvisor* juga membandingkan harga di lebih dari 200 situs pemesanan hotel agar wisatawan dapat menemukan harga terbaik untuk hotel yang tepat bagi mereka. Situs *web TripAdvisor* tersedia dalam 49 pasar, dan menaungi komunitas wisata terbesar di dunia dengan rata-rata 455 juta pengunjung setiap bulannya. (TripAdvisor, 2017)

3.7 *Unstructured Data*

Data tidak terstruktur adalah data yang tidak memiliki model data tetap, dan tidak diatur dengan cara yang ditentukan sebelumnya. Tanpa *preprocessing*, data tidak terstruktur tidak dapat disimpan dalam tabel. Contoh media sosial (*tweet*, *blog*, pos, dan lain-lain.), data *call center*, *email*, survei dengan pertanyaan terbuka, dan lain-lain. Data tidak terstruktur sangat terkait dengan tiga *V 'Big Data* :

1. *Volume* : Data tidak terstruktur biasanya membutuhkan lebih banyak ruangpenyimpanan daripada data terstruktur.
2. *Velocity* : Jumlah data tidak terstruktur meningkat lebih cepat daripada jumlah data terstruktur
3. *Variety* : Data tidak terstruktur dihasilkan pada sumber data yang sebelumnya belum dimanfaatkan, yang dapat mengungkapkan informasi pelanggan yang sangat pribadi (Wieringa, 2016).

3.8 *Web Scraping*

Web Scraping adalah proses pengambilan sebuah dokumen semi-terstruktur dari internet, umumnya berupa 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).

Josi, et al. (2014) dalam penelitiannya memaparkan bahwa *web scraping* memiliki sejumlah langkah, yaitu :

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 diatas, aplikasi *web scraper* dibuat untuk mengetomatisasi pengambilan informasi dari *website* yang ditentukan; dan

4. *Extracted Data and Package History* : Informasi yang didapat dari langkah 3 disimpan dalam tabel atau tabel-tabel *database.*

3.9 Analisis Deskriptif

Analisis deskriptif adalah suatu metode yang berfungsi untuk mendeskripsikanatau memberi gambaran terhadap objek yang diteliti melalui data atau sampel yang telah terkumpul sebagaimana adanya tanpa melakukan analisis dan membuat kesimpulan yang berlaku untuk umum. Dengan kata lain penelitian analisis deskriptif mengambil masalah atau memusatkan perhatian kepada masalah-masalah sebagaimana adanya saat penelitian dilaksanakan hasil penelitian yang kemudian diolah dan dianalisis untuk diambil kesimpulannya. (Sugiyono, 2009).

Analisis data dengan menerapkan metode deskriptif dinyatakan sebagai analisis statistik sederhana atau yang paling sederhana. Akan tetapi, hasil analisis statistik deskriptif tersebut dapat menjadi masukan yang sangat berharga untuk para pengambil keputusan, tergantung pada bentuk dan cara menyajikan hasil analisis tersebut (Agung, 2000).

Hasan (2004) menjelaskan: Analisis deskriptif adalah merupakan bentuk analisis data penelitian untuk menguji generalisasi hasil penelitian berdasarkan satu sample. Analisa deskriptif ini dilakukan dengan pengujian hipotesis deskriptif. Hasil analisisnya adalah apakah hipotesis penelitian dapat digeneralisasikan atau tidak. Jika hipotesis nol (H_0) diterima, berarti hasil penelitian dapat digeneralisasikan. Analisis deskriptif ini menggunakan satu variabel atau lebih tapi bersifat mandiri, oleh karena itu analisis ini tidak berbentuk perbandingan atau hubungan.

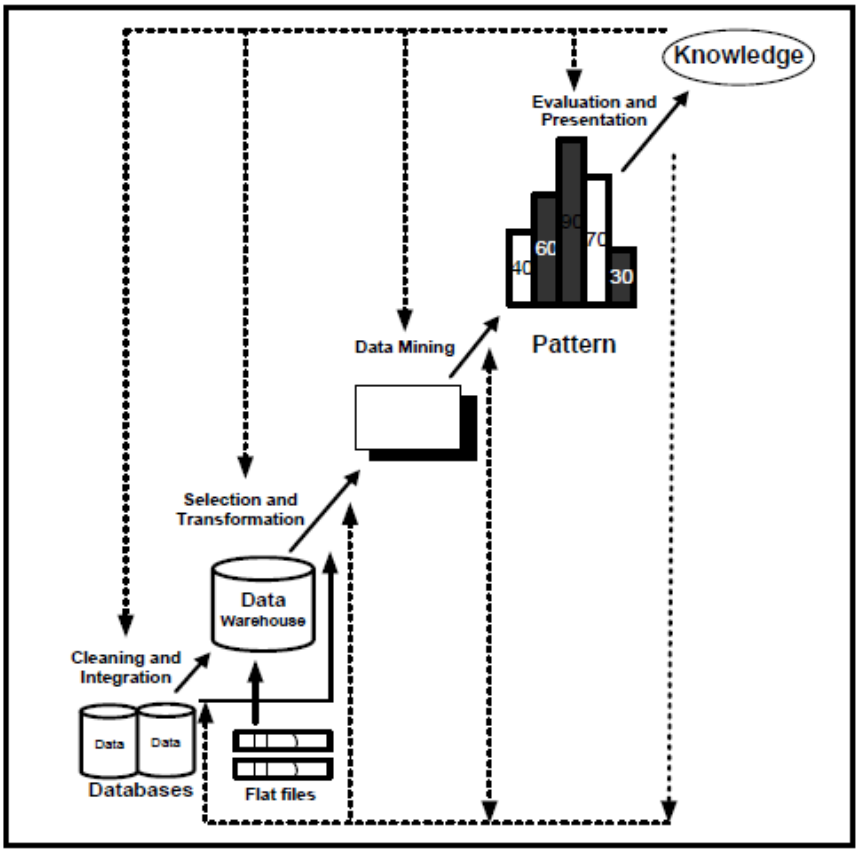
Selanjutnya Hasan (2001) menjelaskan bahwa Statistik deskriptif atau statistik deduktif adalah bagian dari statistik mempelajari cara pengumpulan data dan penyajian data sehingga muda dipahami. Statistik deskriptif hanya berhubungan dengan hal menguraikan atau memberikan keterangan-keterangan mengenai suatu data atau keadaan atau fenomena. Dengan kata statistik deskriptif berfungsi menerangkan keadaan, gejala, atau persoalan.

3.10 *Data Mining*

Secara sederhana *data mining* adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar (Davies, 2004). *Data mining* juga disebut sebagai serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data (Pramudiono, 2007). *Data mining*, sering juga disebut sebagai *knowledge discovery in database* (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar (Santosa, 2007).

Data mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam *database*, *data warehouse*, atau penyimpanan informasi lainnya. *Data mining* berkaitan dengan bidang ilmu – ilmu lain, seperti *database system*, *data warehousing*, statistik, *machine learning*, *information retrieval*, dan komputasi tingkat tinggi. Selain itu, *data mining* didukung oleh ilmu lain seperti *neural netwo*rk, pengenalan pola, *spatial data analysis*, *image database*, *signal processing* (Han, 2006). *Data mining* didefinisikan sebagai proses menemukan pola-pola dalam data. Proses ini otomatis atau seringnya semiotomatis. Pola yang ditemukan harus penuh arti dan pola tersebut memberikan keuntungan, biasanya keuntungan secara ekonomi. Data yang dibutuhkan dalam jumlah besar (Witten, 2005).

Sebagai suatu rangkaian proses, *data mining* dapat dibagi menjadi beberapa tahap yang diilustrasikan di **Gambar 3.1.** Tahap-tahap tersebut bersifat interaktif, pemakai terlibat langsung atau dengan perantaraan *knowledge base*.



Gambar 3.1 Tahap-Tahap *Data Mining* (Han, 2006)

Berdasarkan **Gambar 3.1**, tahap-tahap *data mining* ada 7 yaitu :

1. Pembersihan data (*Data Cleaning*)

   Pembersihan data merupakan proses menghilangkan noise dan data yang tidak konsisten atau data tidak relevan. Pada umumnya data yang diperoleh, baik dari *database* suatu perusahaan maupun hasil eksperimen, memiliki isian-isian yang tidak sempurna seperti data yang hilang, data yang tidak valid atau juga hanya sekedar salah ketik. Selain itu, ada juga atribut-atribut data yang tidak relevan dengan hipotesa *data mining* yang dimiliki. Data-data yang tidak relevan itu juga lebih baik dibuang. Pembersihan data juga akan mempengaruhi performasi dari teknik *data mining* karena data yang ditangani akan berkurang jumlah dan kompleksitasnya.

2. Integrasi data (*Data Integration*)

   Integrasi data merupakan penggabungan data dari berbagai *database* ke dalam satu *database* baru. Tidak jarang data yang diperlukan untuk *data mining* tidak

hanya berasal dari satu *database* tetapi juga berasal dari beberapa *database* atau file teks. Integrasi data dilakukan pada atribut-aribut yang mengidentifikasikan entitas-entitas yang unik seperti atribut nama, jenis produk, nomor pelanggan dan lainnya. Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya. Sebagai contoh bila integrasi data berdasarkan jenis produk ternyata menggabungkan produk dari kategori yang berbeda maka akan didapatkan korelasi antar produk yang sebenarnya tidak ada.

3. Seleksi Data (*Data Selection*)

   Data yang ada pada *database* sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari *database*. Sebagai contoh, sebuah kasus yang meneliti faktor kecenderungan orang membeli dalam kasus *market basket analysis,* tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan saja.

4. Transformasi data (*Data Transformation*)

   Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam *data mining*. Beberapa metode *data mining* membutuhkan format data yang khusus sebelum bisa diaplikasikan. Sebagai contoh beberapa metode standar seperti analisis asosiasi dan *clustering* hanya bisa menerima *input* data kategorikal. Karenanya data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa interval. Proses ini sering disebut transformasi data.

5. Proses *mining*

   Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.

6. Evaluasi pola (*Pattern Evaluation*)

   Untuk mengidentifikasi pola-pola menarik kedalam *knowledge based* yang ditemukan. Dalam tahap ini hasil dari teknik *data mining* berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai. Bila ternyata hasil yang diperoleh tidak sesuai hipotesa ada beberapa alternatif yang dapat diambil seperti menjadikannya

umpan balik untuk memperbaiki proses *data mining*, mencoba metode *data mining* lain yang lebih sesuai, atau menerima hasil ini sebagai suatu hasil yang di luar dugaan yang mungkin bermanfaat.

7. Presentasi pengetahuan (*Knowledge Presentation*)

   Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna. Tahap terakhir dari proses *data mining* adalah bagaimana memformulasikan keputusan atau aksi dari hasil analisis yang didapat. Ada kalanya hal ini harus melibatkan orang-orang yang tidak memahami *data mining*. Karenanya presentasi hasil *data mining* dalam bentuk pengetahuan yang bisa dipahami semua orang adalah satu tahapan yang diperlukan dalam proses *data mining*. Dalam presentasi ini, visualisasi juga bisa membantu mengkomunikasikan hasil *data mining* (Han, 2006).

3.11 *Machine Learning*

Machine Learning (ML) atau pembelajaran mesin merupakan pendekatan dalam *Artificial Intelligence* (AI) yang banyak digunakan untuk menggantikan atau menirukan perilaku manusia untuk menyelesaikan masalah atau melakukan otomatisasi. Sesuai namanya, ML mencoba menirukan bagaimana proses manusia atau makhluk cerdas belajar dan mengeneralisasi (Tanaka & Okutomi, 2014)

Purnamasari, et al. (2013) dalam bukunya memaparkan bahwa *Machine Learning* adalah cabang dari kecerdasan buatan, merupakan disiplin ilmu yang mencakup perancangan dan pengembangan algoritma yang memungkinkan komputer untuk mengembangkan perilaku yang didasarkan kepada data empiris, seperti dari sensor data pada basis data. Sistem pembelajaran dapat memanfaatkan contoh (data) untuk menangkap ciri yangdiperlukan dari probabilitas yang mendasarinya (yang tidak diketahui). Data dapat dilihat sebagai contoh yang menggambarkan hubungan antara variabel yang diamati. Fokus besar penelitian *Machine Learning* adalah bagaimana mengenali secara otomatis pola kompleks dan membuat keputusan cerdas berdasarkan data. Kesukarannya terjadi karena himpunan semua perilaku yang mungkin, dari semua masukan yang

dimungkinkan, terlalu besar untuk diliput oleh himpunan contoh pengamatan (data pelatihan). Karena itu *Machine Learning* harus merampatkan (generalisasi) perilaku dari contoh yang ada untuk menghasilkan keluaran yang berguna dalam kasus-kasus baru.

3.12 *Text Mining*

Text mining adalah proses penemuan akan informasi atau *trend* baru yang sebelumnya tidak terungkap dengan memproses dan menganalisa data dalam jumlah besar. Dalam menganalisa sebagian atau keseluruhan *unstructured text*, *text mining* mencoba untuk mengasosiasikan satu bagian *text* dengan yang lainnya berdasarkan aturan-aturan tertentu. Hasil yang di harapkan adalah informasi baru atau *"insight"* yang tidak terungkap jelas sebelumnya. (Adiwijaya, 2006).

Seperti halnya *data mining, text mining* juga menghadapi masalah yang sama, termasuk jumlah data yang besar, dimensi yang tinggi, data dan struktur yang terus berubah, dan data *"noise"*. Berbeda dengan *data mining* yang utamanya memproses *structured data*, data yang digunakan *text mining* pada umumnya dalam bentuk *unstructured*, atau minimal *semi-structured, text*. Akibatnya, *text mining* mempunyai tantangan tambahan yang tidak di temui di *data mining*, seperti struktur *text* yang *complex* dan tidak lengkap, arti yang tidak jelas dan tidak *standard*, dan bahasa yang berbeda ditambah translasi yang tidak akurat. (Adiwijaya, 2006).

Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber data yang digunakan pada *text mining* adalah kumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Adapun tugas khusus dari *text mining* antara lain yaitu pengkategorisasian teks (*text categorization*) dan pengelompokan teks (*text clustering*) (Hilwah, Kudus, & Sunendiari, 2017).

Berdasarkan ketidakaturan struktur data teks, maka proses *text mining* memerlukan beberapa tahap awal yang pada intinya adalah mempersiapkan agar teks dapat diubah menjadi lebih terstruktur. Adapun tahapan yang dilakukan secara umum dalam *text mining* terdapat empat proses pokok, yaitu:

1. *Spelling normalization*, merupakan perbaikan kata-kata yang salah eja atau disingkat dengan bentuk tertentu. Misalnya kata "tidak" memiliki banyak bentuk penulisan seperti tdk, gak, nggak, enggak, dan banyak lainnya.
2. *Case folding*, merupakan proses pengubahan huruf dalam dokumen menjadi satu bentuk, misalnya huruf kapital menjadi huruf kecil dan sebaliknya.
3. *Tokenizing*, merupakan proses pemisahan teks menjadi potongan kalimat dan kata yang disebut token.
4. *Filtering*, merupakan proses membuang kata-kata serta tanda-tanda yang tidak bermakna secara signifikan, seperti *hashtag* (#), *url*, tanda baca tertentu (*emoticon*), dan lainnya (Megawati, 2015).

3.13 Analisis Sentimen

Analisis sentimen atau biasa disebut *opinion mining* merupakan salah satu cabang penelitian *Text Mining*. *Opinion mining* adalah riset komputasional dari opini, sentimen dan emosi yang diekspresikan secara tekstual. Jika diberikan suatu set dokumen teks yang berisi opini mengenai suatu objek, maka *opinion mining* bertujuan untuk mengekstrak atribut dan komponen dari objek yang telah dikomentasi pada setiap dokumen dan untuk menentukan apakah komentar tersebut bermakna positif atau negative (Shelby, 2013). *Sentiment Analysis* dapat dibedakan berdasarkan sumber datanya, beberapa level yang sering digunakan dalam penelitian *Sentiment Analysis* adalah *Sentiment Analysis* pada level dokumen dan *Sentiment Analysis* pada level kalimat. Berdasarkan level sumber datanya *Sentiment Analysis* terbagi menjadi 2 kelompok besar yaitu (Clayton, 2011): *Coarse-grained Sentiment Analysis* dan *fined-grained Sentiment Analysis.*

Pada *Sentiment Analysis Coarse-grained*, *Sentiment Analysis* yang dilakukan adalah pada level dokumen. Secara garis besar fokus utama dari *Sentiment Analysis* jenis ini adalah menganggap seluruh isi dokumen sebagai sebuah sentiment positif atau sentiment negatif. *Fined-grained Sentiment Analysis* adalah *Sentiment Analysis* pada level kalimat. Fokus utama *fined-greined Sentiment Analysis* adalah menentukan sentimen pada setiap kalimat.

3.14 Pembobotan Kata

Pembobotan kata (*term weighting*) adalah proses pembobotan pada kata. Pembobotan dasar dilakukan dengan menghitung frekuensi kemunculan *term* dalam dokumen. Frekuensi kemunculan (*term frequency*) merupakan petunjuk sejauh mana *term* tersebut mewakili isi dokumen. Semakin besar kemunculan suatu *term* dalam dokumen akan memberikan nilai kesesuian yang semakin besar.

Faktor lain yang diperhatikan dalam pemberian bobot adalah kejarangmunculan kata (*term scarcity*) dalam koleksi. Kata yang muncul pada sedikit dokumen harus dipandang sebagai kata yang lebih penting (*uncommon tems*) daripada kata yang muncul pada banyak dokumen. Pembobotan akan memperhitungkan faktor kebalikan frekuensi dokumen yang mengandung suatu kata (*inverse document frequency*). Faktor terakhirnya adalah faktor normalisasi terhadap panjang dokumen. Dokumen dalam suatu koleksi memiliki karakteristik panjang yang beragam. Ketimpangan terjadi karena dokumen yang panjang akan cenderung mempunyai frekuensi kemunculan kata yang besar. Sehingga untuk mengurangi ketimpangan tersebut diperlukan faktor normalisasi dalam pembobotan (Mandala & Setiawan, 2002).

Hal yang perlu diperhatikan dalam pencarian informasi dari koleksi dokumen yang heterogen adalah pembobotan *term*. *Term* dapat berupa kata, frase atau unit hasil *indexing* lainnya dalam suatu dokumen yang dapat digunakan untuk mengetahui konteks dari dokumen tersebut. Karena setiap kata memiliki tingkat kepentingan yang berbeda dalam dokumen, maka untuk setiap kata tersebut diberikan sebuah indikator, yaitu *term weight* (Zafikri, 2008).

Menurut Zafikri (2008) *term weighting* atau pembobotan *term* sangat dipengaruhi oleh hal-hal berikut ini :

1. *Term Frequency* (TF) *factor*, yaitu faktor yang menentukan bobot *term* padasuatu dokumen berdasarkan jumlah kemunculannya dalam dokumen tersebut. Nilai jumlah kemunculan suatu kata (*term frequency*) diperhitungkan dalam pemberian bobot terhadap suatu kata. Semakin besar jumlah kemunculan suatu *term* (tf tinggi) dalam dokumen, semakin besar

pula bobotnya dalam dokumen atau akan memberikan nilai kesesuian yang semakin besar.

2. *Inverse Document Frequency* (IDF) *factor*, yaitu pengurangan dominansi *term* yang sering muncul di berbagai dokumen. Hal ini diperlukan karena *term* yang banyak muncul di berbagai dokumen, dapat dianggap sebagai *term* umum (*common term*) sehingga tidak penting nilainya. Sebaliknya faktor kejarangmunculan kata (*term scarcity*) dalam koleksi dokumen harus diperhatikan dalam pemberian bobot. Menurut Mandala (dalam Witten, 1999) 'Kata yang muncul pada sedikit dokumen harus dipandang sebagai kata yang lebih penting (*uncommon terms*) daripada kata yang muncul pada banyak dokumen. Pembobotan akan memperhitungkan faktor kebalikan frekuensi dokumen yang mengandung suatu kata (*Inverse Document Frequency).*

Metode TF-IDF merupakan metode pembobotan *term* yang banyak digunakan sebagai metode pembanding terhadap metode pembobotan baru. Pada metode ini, perhitungan bobot *term t* dalam sebuah dokumen dilakukan dengan mengalikan nilai *Term Frequency* dengan *Inverse Document Frequency*. Pada *Term Frequency* (TF), terdapat beberapa jenis formula yang dapat digunakan yaitu (Zafikri, 2008):

1. TF biner (*binery TF*), hanya memperhatikan apakah suatu kata ada atau tidak dalam dokumen, jika ada diberi nilai satu, jika tidak diberi nilai nol
2. TF murni (*raw TF*), nilai tf diberikan berdasarkan jumlah kemunculan suatu kata di dokumen. Contohnya, jika muncul lima kali maka kata tersebut akan bernilai lima.
3. TF logaritmik, hal ini untuk menghindari dominansi dokumen yang mengandung sedikit kata dalam *query*, namun mempunyai frekuensi yang tinggi.

   $$tf = 1 + \log(tf) \quad (3.1)$$

4. TF normalisasi, menggunakan perbandingan antara frekuensi sebuah kata dengan jumlah keseluruhan kata pada dokumen.

$$tf = 0{,}5 + 0{,}5x\left(\frac{tf}{\max tf}\right) \quad (3.2)$$

Inverse Document Frequency (idf) dihitung dengan menggunakan formula

$$idf_j = log\left(\frac{D}{df_j}\right) \quad (3.3)$$

Dimana

D : adalah jumlah semua dokumen dalam koleksi

Df_i : adalah jumlah dokumen yang mengandung *term* t_j

Dengan demikian rumus umum untuk *TF-IDF* adalah penggabungan dari formula perhitungan *raw TF* dan formula *IDF* dengan cara mengalikan nilai *Term Frequency (TF) dengan nilai Inverse Document Frequency (idf)* :

$w_{ij} = tf_{ij} \times idf_j$

$$w_{ij} = tf_{ij} \times log\left(\frac{D}{df_j}\right) \quad (3.4)$$

Keterangan :

w_{ij} : adalah bobot *term* t_j terhadap dokumen d_i

tf_{ij} : adalah jumlah kemunculan term t_j dalam dokumen d_i

D : adalah jumlah semua dokumen yang ada dalam *database*

df_i : adalah jumlah dokumen yang mengandung term t_j (minimal ada satu kata yaitu *term* t_j)

3.15 Klasifikasi

Klasifikasi adalah proses untuk menemukan model atau fungsi yang dapat menjelaskan atau membedakan konsep atau kelas data. Tujuan klasifikasi adalah untuk memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Proses klasifikasi biasanya dibagi menjadi dua fase yaitu *fase learning* dan *fase test*. Pada *fase learning*, sebagian data yang telah diketahui kelas datanya diumpankan untuk membentuk model perkiraan. Kemudian pada *fase test* model yang sudah terbentuk diuji dengan sebagian data lainnya untuk mengetahui akurasi dari model tersebut. Bila akurasinya mencukupi model ini dapat dipakai untuk prediksi kelas

data yang belum diketahui. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk menghitung jarak antara ciri-ciri citra *template* dan citra masukan (Ulwan, 2016).

3.15.1 Evaluasi Peforma Model

Model klasifikasi yang dibuat ialah pemetaan dari suatu baris data dengan keluaran sebuah hasil prediksi kelas/target dari data tersebut. Klasifikasi yang memiliki dua kelas sebagai keluarannya disebut dengan klasifikasi biner. Kedua kelas tersebut biasa direpresentasikan dalam {0,1}, {+1,-1} atau {*positive*; *negative*}.

Dalam proses evaluasi klasifikasi terdapat empat kemungkinan yang terjadi dari proses pengklasifikasian suatu baris data. Jika data positif dan diprediksi positif akan dihitung sebagai *true positive*, tetapi jika data itu diprediksi negatif maka akan dihitung sebagai *false negative*. Jika data negatif dan diprediksi negatif akan dihitung sebagai *true negative*, tetapi jika data tersebut diprediksi positif maka akan dihitung sebagai *false positive* (Fawcett, 2006). Hasil klasifikasi biner pada suatu dataset dapat direpresentasikan dengan matriks 2 x 2 yang disebut *confusion matrix*.

Tabel 3.1 *Confusion Matrix* Aktual

	Class	***Positive***	***Negative***
Prediksi	***Positive***	*True Positive (TP)*	*False Positive (FP)*
	Negative	*False Negative (FN)*	*True Negative (TN)*

Pada **Tabel 3.1** adalah gambaran dari *confusion matrix*. Terdapat beberapa rumus umum yang dapat digunakan untuk menghitung performa klasifikasi. Hasil dari nilai akurasi, presisi, dan *recall* biasa ditampilkan dalam persentase (Rianto, 2016).

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (3.5)$$

$$Presisi = \frac{TP}{TP+FP} \quad (3.6)$$

$$Recall = \frac{TP}{TP+FN} \quad (3.7)$$

3.15.2 *K-Fold Cross Validation*

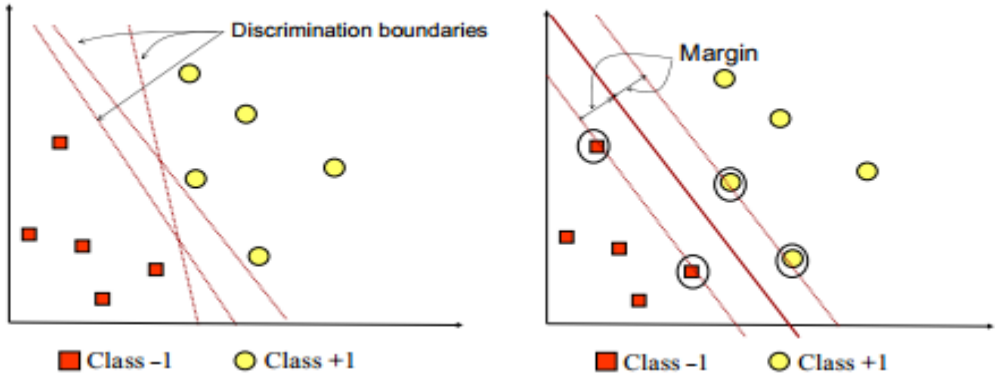
K-Fold Cross Validation merupakan salah satu teknik untuk melakukan estimasi tingkat kesalahan pengujian pemrosesan citra digital. Menurut Fauzie (2010) cara kerja *K-fold cross validation* yaitu dengan mengelompokkan data latih dan data uji yang saling terpisah, kemudian melakukan proses pengujian yang diulang sebanyak *K* kali.

Langkah dari *K-fold cross validation* (Pratiwi, 2010) antara lain: (1) Membagi data yang tersedia menjadi *K* kelompok. (2) Setiap K dibuat sejumlah T himpunan data yang memuat semua data latih kecuali yang berada di kelompok ke-k. (3) Mengerjakan algoritma yang dimiliki dengan sejumlah T data latih. (4) Pengujian algoritma menggunakan data pada kelompok *K* sebagai data uji. (5) Melakukan pencatatan hasil algoritma. Menurut Pratiwi (2010) keuntungan dari teknik *K-fold cross validation* yaitu menunjukkan bahwa semua elemen pada baris data digunakan untuk pelatihan sekaligus pengujian.

3.16 *Support Vector Machine*

Menurut Prasetyo (2012), metode SVM *(Support Vector Machine)* merupakan teori pembelajaran statistic dan dapat memberikan hasil yang lebih baik dari pada metode yang lain. SVM dapat bekerja dengan baik pada data dengan berdimensi set tinggi. Selain itu, SVM menggunakan teknik kernel dan hanya sejumlah data yang terpilih yang berkontribusi untuk membangun model klasifikasi. Hal tersebut menjadi kelebihan SVM, karena tidak semua data latih akan dilihat untuk dilibatkan dalam setiap iterasi pelatihannya. Konsep dasar dari SVM yaitu membentuk *hyperplane (maximal margin hyperplane)*.

3.16.1 SVM Pada Data Terpisah Secara Linear



(a) **(b)**

(Sumber: Riska, Cahyani, & Rosadi, 2015)

Gambar 3.2. (a) *Decision boundary* yang mungkin dan (b) *Decision boundary* dengan *margin* maksimal

Gambar 3.2 memperlihatkan beberapa *pattern* yang merupakan anggota dari dua buah kelas: +1 dan –1. *Pattern* yang tergabung pada kelas –1 disimbolkan dengan warna merah (kotak), sedangkan *pattern* pada kelas +1, disimbolkan dengan warna kuning (lingkaran). Problem klasifikasi dapat diterjemahkan dengan usaha menemukan *hyperplane* yang memisahkan antara kedua kelompok tersebut. Berbagai alternatif garis pemisah *(discrimination boundaries)* ditunjukkan pada gambar a.

Hyperplane pemisah terbaik antara kedua *class* dapat ditemukan dengan mengukur *margin hyperplane* tersebut dan mencari titik maksimalnya. *Margin* adalah jarak antara *hyperplane* tersebut dengan pattern terdekat dari masing-masing kelas. *Pattern* yang paling dekat ini disebut sebagai *support vector*. Garis tebal pada gambar b menunjukkan *hyperplan*e yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua kelas, sedangkan titik merah dan kuning yang berada dalam lingkaran hitam adalah *support vector.* Data yang tersedia dinotasikan sebagai $x \in R^d$ sedangkan label masing-masing dinotasikan $yi \ni \{-1, +1\}$ untuk $i = 1,2,3, \ldots n$ yang mana n adalah banyaknya data. Diasumsikan

kedua kelas dapat terpisah secara sempurna oleh *hyperlane* berdimensi *d,* yang didefinisikan:

$$\vec{w}.\vec{x} + b = 0 \quad (3.8)$$

Pattern $\overrightarrow{x_i}$ yang termasuk kelas -1 (sampel negatif) dapat dirumuskan sebagai *pattern* yang memenuhi pertidaksamaan:

$$\vec{w}.\vec{x} + b \leq -1 \quad (3.9)$$

Sedangkan *pattern* $\overrightarrow{x_i}$ yang termasuk kelas +1 (sampel positif)

$$\vec{w}.\vec{x} + b \geq +1 \quad (3.10)$$

Keterangan:

w = *vector* bobot

X = nilai masukan atribut

B = bias

Margin terbesar dapat ditemukan dengan memaksimalkan nilai jarak antara jarak dan titik terdekatnya, yaitu $\frac{1}{\|w\|}$. Hal ini dapat dirumuskan sebagai *Quadratic Programming (QP) problem*, yaitu mencari titik minimal persamaan 3.11, dengan memperlihatkan *constraint* persamaan 3.12.

$$min_w = \tau(w) = \frac{1}{2}\|w\|^2 \quad (3.11)$$

$$y_i(x_i.w + b) - 1 \geq 0 \quad (3.12)$$

Masalah ini dapat dipecahkan dengan berbagai teknik komputasi, diantaranya *lagrange multiplier* yang dinyatakan pada persamaan 3.13

$$L(w, b, a) = \frac{1}{2\|w\|^2} - \sum_{i=1}^{l} a_i \left(y_i\left((\vec{x}.\vec{w} + +b) - 1\right)\right) \quad (3.13)$$

dengan $i = 1,2, \ldots, l$

Dimana a adalah *lagrange multiplier,* yang bernilai 0 atau positif $a_i \geq 0$. Nilai optimal dari persamaan 3.13 dapat dihitung dengan meminimalkan L terhadap w dan b, dan memaksimalkan L terhadap a_i, dengan memperhatikan sifat bahwa pada titik optimal *gradient* L = 0 persamaan 3.13 dapat dimodifikasi sebagai maksimalisasi *problem* yang hanya mengandung a_i, sebagaimana terlihat pada persamaan 3.14 dan 3.15 dibawah ini.

$$\sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,i=1}^{l} a_i a_j y_i y_j \overrightarrow{x_i}.\overrightarrow{x_j} \quad (3.14)$$

dimana $$a_i \geq 0(i = 1,2, \dots l)\sum_{i,i=1}^{l} a_i y_i = 0 \quad (3.15)$$

Dengan demikian, maka akan diperoleh a_i yang kebanyakan bernilai positif, Data yang berkorelasi dengan a_i yang positif inilah yang disebut sebagai *support vector*.

3.16.2 SVM Pada Data Tidak Terpisah Secara Linear

Kasus data yang tidak terpisah secara linear diasumsikan bahwa kelas pada *input space* tidak dapat terpisah secara sempurna. Hal ini menyebabkan *constraint* pada persamaan 3.12 tidak dapat terpenuhi, sehingga optimisasi tidak dapat dilakukan, untuk mengatasi masalah ini *SVM* dirumuskan ulang dengan memperkenalkan teknik *softmargin*. Dalam *softmargin* persamaan 3.12 dimodifikasi dengan menggunakan *slack* variabel sehingga terlihat pada persamaan 3.16.

$$y_i(x_i.w + b) \geq 1 - \varepsilon_i \quad (3.16)$$

Dengan demikian, persamaan 3.11 diubah menjadi persamaan 3.17.

$$min_w = \tau(w) = \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{l} \varepsilon_i \quad (3.17)$$

Fitur c digunakan untuk mengontrol *teadeoff* antara *margin* dan kesalahan klasifikasi ε.

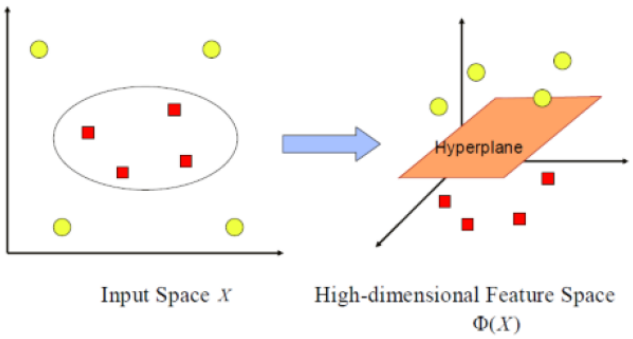
3.16.3 *Kernel Trick* dan *Non-Linear Classification* Pada SVM

Pada umumnya masalah dalam domain dunia nyata (*real world problem*) jarang yang bersifat *linear separable* dan kebanyakan bersifat non linear. Untuk menyelesaikan masalah non linear, *SVM* dimodifikasi dengan memasukkan fungsi *Kernel*. Dalam non linear *SVM*, pertama-tama data x dipetakan oleh fungsi $\Phi(x)$ ke ruang vektor yang berdimensi lebih tinggi. Pada ruang vektor yang baru ini, *hyperplane* yang memisahkan kedua kelas tersebut dapat dikonstruksikan. Hal ini sejalan dengan teori Cover yang menyatakan *"Jika suatu transformasi bersifat*

non linear dan dimensi dari feature space cukup tinggi, maka data pada input space dapat dipetakan ke feature space yang baru, dimana pattern-pattern tersebut pada probabilitas tinggi dapat dipisahkan secara linear".

Ilustrasi dari konsep ini dapat dilihat pada **Gambar 3.3**. Pada **Gambar 3.3** diperlihatkan data pada kelas kuning dan data pada keals merah yang berada pada *input space* berdimensi dua tidak dapat dipisahkan secara linear. Selanjutnya fungsi Φ memetakan tiap data pada *input space* tersebut ke ruang vektor baru yang berdimensi lebih tinggi (dimensi 3), dimana kedua kelas dapat dipisahkan secara linear oleh sebuah *hyperplane*. Notasi matematika dari *mapping* ini adalah sebagai berikut.

$$\Phi : \Re^d \rightarrow \Re^d \;\; d < q \tag{3.18}$$



(Sumber : Ulwan, 2016)

Gambar 3.3 *Fungsi* ***Φ*** *memetakan data ke ruang vector yang berdimensi lebih tinggi, sehingga kedua kelas dapat dipisahkan secara linear oleh sebuah hyperplane*

Selanjutnya proses pembelajaran pada *SVM* dalam menemukan titik-titik *support vector*, hanya bergantung pada *dot product* dari data yang sudah ditransformasikan pada ruang baru yang berdimensi lebih tinggi, yaitu $\Phi(xi).\Phi(xi)$. Karena umumnya transformasi Φ ini tidak diketahui, dan sangat sulit untuk difahami secara mudah, maka perhitungan dot product tersebut sesuai teori Mercer dapat digantikan dengan fungsi *kernel* yang terlihat pada persamaan 3.19.

$$K\left(x_i.x_j\Phi(x_i).\Phi\left(x_j\right)\right) \tag{3.19}$$

Beberapa *kernel* yang umum dipakai pada SVM adalah:

a. *Polynomial*

Kernel tirck polynomial cocok digunakan untuk menyelesaikan masalah klasifikasi, dimana dataset pelatihan sudah normal. *Kernel trick* ini dinyatakan dalam persamaan.

$$K(\overrightarrow{x_i}, \overrightarrow{x_j}) = (\overrightarrow{x_i}, \overrightarrow{x_j} + 1)^p \quad (3.20)$$

b. *Radial Basis Function (RBF)* atau *Gaussian*

Kernel trick radial basis function atau *gaussian* merupakan *kernel* yang paling banyak digunakan untuk menyelesaikan masalah klasifikasi untuk dataset yang tidak terpisah secara linear, dikarenakan akurasi pelatihan dan akurasi prediksi yang sangat baik pada *kernel* ini, dimana *kernel radial basis function* dinyatakan dalam persamaan 3.21.

$$K(\overrightarrow{x_i}, \overrightarrow{x_j}) = exp\left(-\left\|\overrightarrow{x_i}, \overrightarrow{x_j}\right\|^2\right)\gamma \quad (3.21)$$

c. *Sigmoid*

Kernel sigmoid merupakan *kernel trick* SVM yang merupakan pengembangan dari jaringan saraf tiruan, dimana *kernel* ini dinyatakan dengan persamaan 3.22.

$$K(\overrightarrow{x_i}, \overrightarrow{x_j}) = tanh(\overrightarrow{ax_i}.\overrightarrow{x_j} + \beta) \quad (3.22)$$

Kernel trick memberikan beberapa kemudahan, karena dalam proses pembelajaran *SVM*, untuk menentukan *support vector*, pengguna hanya cukup mengetahui fungsi *kernel trick* yang dipakai, tanpa perlu mengetahui wujud dari fungsi non-linier. Dari keseluruhan *kernel trick* tersebut, *kernel trick radial basis function* merupakan *kernel trick* yang memberikan hasil terbaik pada proses klasifikasi khususnya untuk data yang tidak bisa dipisahkan secara linear. Selanjutnya hasil klasifikasi dari data x diperoleh dari persamaan berikut:

$$f(x) = \sum_{i=1,\overrightarrow{x_i},\in SV}^{n} a_i y_i K(\overrightarrow{x_i}, \overrightarrow{x_j}) + b \quad (3.23)$$

3.17 Asosiasi Kata

Istilah korelasi sering digunakan untuk menyatakan hubungan dua atau lebih variabel yang sifatnya kuantitatif, sedangkan istilah asosiasi sering dimaknai keeratan hubungan antara dua atau lebih variabel yang sifatnya kualitatif (Ulwan, 2016). Asosiasi merupakan proses pencarian hubungan antar elemen data. Dalam dunia industri retail, analisis asosiasi biasanya disebut *Market Basket Analysis* (Miner et al, 2012). Penelitian ini menggunakan pendekatan asosiasi untuk menemukan hubungan antar kata pada masing-masing klasifikasi ulasan positif dan ulasan negatif, sehingga mendapatkan informasi yang dapat dijadikan referensi bagi pihak hotel maupun pengunjung untuk mengetahui topik yang paling sering dibicarakan terkait hotel The Phoenix Hotel Yogyakarta.

$$\boldsymbol{r} = \frac{\boldsymbol{N}\sum \boldsymbol{X_i Y_i} - (\sum \boldsymbol{X_i})(\sum \boldsymbol{Y_i})}{\sqrt{\left\{\boldsymbol{N X_i^2} - (\sum \boldsymbol{X_i})^2\right\}\{\boldsymbol{N}\sum \boldsymbol{Y_i^2} - (\sum \boldsymbol{Y_i})^2\}}} \tag{3.23}$$

3.18 Diagram Sebab-Akibat *(Fishbone Diagram)*

Diagram sebab-akibat dikembangkan oleh Dr. Kaoru Ishikawa pada tahun 1943, sehingga sering disebut dengan diagram Ishikawa. Diagram sebab-akibat (cause and effect diagram atau fishbone diagram) adalah sebuah teknik grafis yang digunakan untuk mengurutkan dan menghubungkan interaksi antara faktor-faktor yang berpengaruh dalam suatu proses.

Diagram ini berguna untuk menganalisia dan menemukan faktor-faktor yang berpengaruh atau efek secara signifikan di dalam menentukan karakteristik kualitas output kerja. Efek ini bisa bernilai "baik" dan bisa bernilai "buruk". Jadi dengan diketahui sebab dari efek yang terjadi, diharapkan hasil dari proses produksi bisa diperbaiki dengan mengubah faktor terkontrol dari suatu proses. Diagram ini juga berguna untuk mengidentifikasi akar penyebab potensi dari suatu masalah. Diagram sebab akibat memfokuskan pada penekanan masalah atau gejala yang merapakan akar penyebab masalah. Diagram sebab akibat juga menampilkan penyebab-penyebab masalah dengan cara menghubungkan penyebab-penyebab menjadi satu (Fauziah, 2009).

Analisa tulang ikan dipakai untuk mengkategorikan berbagai sebab potensial dari satu masalah atau pokok persoalan dengan cara yang mudah dimengerti dan rapi. Juga alat ini membantu kita dalam menganalisis apa yang sesungguhnya terjadi dalam proses. Yaitu dengan cara memecah proses menjadi sejumlah kategori yang berkaitan dengan proses, mencakup manusia, material, mesin, prosedur, kebijakan dan sebagainya (Imamoto *et al*., 2008).

Manfaat analisa tulang ikan yaitu:

1. Memperjelas sebab-sebab suatu masalah atau persoalan.
2. Dapat menggunakan kondisi yang sesungguhnya untuk tujuan perbaikan kualitas produk atau jasa, lebih efisien dalam penggunaan sumber daya, dan dapat mengurangi biaya.
3. Dapat mengurangi dan menghilangkan kondisi yang menyebabkan ketidaksesuaian produk atau jasa, dan keluhan pelanggan.
4. Dapat membuat suatu standarisasi operasi yang ada maupun yang direncanakan.
5. Dapat memberikan pendidikan dan pelatihan bagi karyawan dalam kegiatan pembuatan keputusan dan melakukan tindakan perbaikan.

Langkah-langkah dalam analisis *fishbone* adalah:

a. Menyiapkan sesi sebab-akibat.
b. Mengidentifikasi akibat.
c. Mengidentifikasi berbagai kategori.
d. Menemukan sebab-sebab potensial dengan cara sumbang saran.
e. Mengkaji kembali setiap kategori sebab utama.
f. Mencapai kesepakatan atas sebab-sebab yang paling mungkin.

BAB IV
METODOLOGI PENELITIAN

4.1 Populasi dan Sampel Penelitian

Populasi yang digunakan dalam penelitian ini adalah data ulasan/*review* hotel The Phoenix Hotel Yogyakarta - MGallery Collection yang terdapat pada situs *website TripAdvisor*, yang diambil sejak bulan Januari 2011 hingga bulan November 2017. Sedangkan sampel yang digunakan dalam penelitian ini adalah semua data ulasan hotel The Phoenix Hotel Yogyakarta berbahasa Inggris dan berbahasa Indonesia. Data ulasan berbahasa Inggris berjumlah sebanyak 1.943 ulasan, sedangkan ulasan berbahasa Indonesia berjumlah sebanyak 1.233 ulasan, sehingga total semua ulasan yang di ambil adalah sebanyak 3.176 ulasan.

4.2 Teknik Pengumpulan Data

Proses pengambilan data dari situs *TripAdvisor* dilakukan dengan menggunakan cara teknik *web scraping*, yaitu proses pengambilan sebuah dokumen semi-terstruktur dari internet. Proses *scraping* dilakukan dengan menggunakan program *R* versi 3.4.3.

4.3 Variabel dan Definisi Operasional Variabel

Variabel yang digunakan dalam penelitian ini ditampilkan dalam **Tabel 4.1** tentang penjelasan dan definisi operasional penelitian :

Tabel 4.1 *Definisi Operasional Variabel*

Variabel	Definisi Operasional Variabel
Review	Deskripsi isi ulasan pengunjung tentang hotel
Date	Tanggal pengunjung menulis ulasan
Rating	Rating penilaian pengunjung terhadap hotel

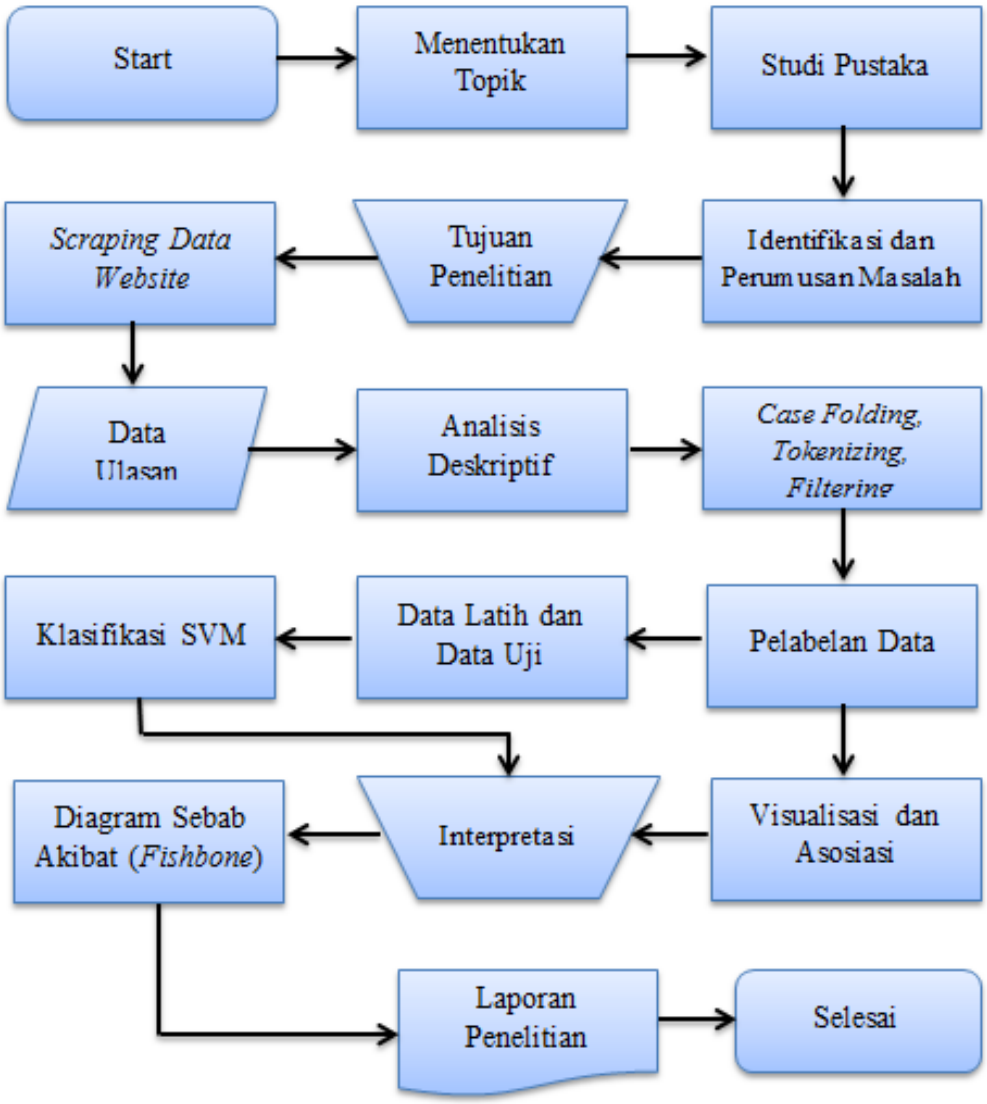
4.4 Metode Analisis Data

Proses analisis dalam penelitian ini menggunakan bantuan *software Microsoft Excel 2016, R 3.4.3, SPSS,* dan *Xmind 7.* Ada beberapa metode yang digunakan dalam penelitian ini, diantaranya sebagai berikut :

1. *Web Scraping*, digunakan untuk mendapatkan/mengumpulkan data ulasan pengunjung hotel The Phoenix Hotel Yogyakarta secara *online* yang bersumber dari situs *TripAdvisor*.
2. Analisis Deskriptif , digunakan untuk memberikan gambaran umum ulasan hotel The Phoenix Hotel Yogyakarta yang ada pada situs *TripAdvisor*.
3. *Analisis Sentimen* berbasis kamus *lexicon*, digunakan untuk melakukan pelabelan data ke dalam kelas sentimen positif dan negatif.
4. Klasifikasi *Machine Learning* dengan algoritma *Support Vector Machine*, yang digunakan untuk mengklasifikasikan ulasan berdasarkan ulasan positif, dan ulasan negatif serta melihat tingkat akurasi dalam melakukan klasifikasi teks ulasan.
5. *Barplot* dan *Wordcloud*, digunakan untuk melakukan visualisasi kata yang paling sering muncul/banyak digunakan dalam ulasan.
6. *Association*, digunakan untuk mengidentifikasi dan membentuk pola kata yang berasosisasi dengan kata lainnya guna mendapatkan informasi yang dianggap penting dan berguna.
7. Diagram sebab-akibat (*Fishbone*), digunakan untuk mengidentifikasi faktor-faktor penyebab yang paling dominan terhadap permasalahan yang didapatkan dari ulasan negatif sehingga bisa dilakukan rencana penaggulangan dari masalah yang dihadapi.

4.5 Langkah Penelitian

Tahapan atau langkah dalam penelitian ini digambarkan dalam *flowchart* melalui **Gambar 4.1** berikut ini:



Gambar 4.1 *Flowchart Penelitian*

BAB V
ANALISIS DAN PEMBAHASAN

Berdasarkan kajian teori dan hasil-hasil penelitian sebelumnya, maka paparan tentang implementasi *SVM* dan asosiasi untuk sentimen analisis data ulasan The Phoenix Hotel Yogyakarta MGallery Collection – Yogyakarta pada situs *TripAdvisor*, dapat diketengahkan beberapa hal yang merupakan analisis dan pembahasan hasil-hasil penelitian, berkaitan dengan (1) pengumpulan data dengan *web scraping*, (2) analisis deskriptif, (3) *text preprocessing*, (4) pelabelan kelas sentimen, (5) pembuatan data latih dan data uji, (6) klasifikasi dengan *support vector machine*, (7) visualisasi dan asosiasi, dan (8) diagram sebab-akibat (*Fishbone Diagram*).

5.1 Pengumpulan Data dengan *Web Scraping*

Proses *web scraping* dilakukan dengan menggunakan *software R 3.4.3.* Sebelum melakukan proses *scraping*, beberapa atribut yang diperlukan diantaranya sebagai berikut:

1. *Software R* dan *web browser google chrome* yang telah terinstal di perangkat komputer;
2. *Packages* "rvest" yang telah terinstal di dalam *software* R;
3. *Add-extensions* "*selector gadget*" yang di pasang pada *web browser google chrome*. *Selector gadget* berfungsi untuk melakukan seleksi *CSS* untuk mengetahui letak data yang akan di ekstrak pada halaman *website*;
4. Koneksi *internet*.

Pada kasus ini, dilakukan pengambilan data/informasi dari situs *web* www.tripadvisor.com. Data yang akan diambil adalah berupa data ulasan (*review*) pengunjung The Phoenix Hotel Yogyakarta dalam bahasa Inggris dan bahasa Indonesia yang didalamnya memuat konten berupa atribut *id*, *quote*/kutipan, *rating*, tanggal, dan isi ulasan. Data yang tersimpan pada halaman *web* adalah data yang berupa kumpulan-kumpulan kode *HTML*, *javascript*, *CSS*, dan sebagainya,

maka dibutuhkan suatu teknik untuk memisahkan teks dari *tag-tag* kode *HTML* pada halaman *website*, sehingga isi halaman *web* dapat diambil secara spesifik.

Untuk mendapatkan data ulasan The Phoenix Hotel Yogyakarta, terlebih dahulu dilakukan dengan menuliskan *keyword* "The Phoenix Hotel Yogyakarta" pada kolom pencarian situs *TripAdvisor* menggunakan *web browser Google Chrome*, hingga diperoleh halaman ulasan hotel Royal Ambarrukmo seperti terlihat pada gambar berikut:

Gambar 5.1 *Halaman review The Phoenix Hotel pada situs TripAdvisor*

Gambar 5.1 merupakan contoh tampilan halaman ulasan pengunjung The Phoenix Hotel dalam bahasa Inggris. Ulasan tersebut mengandung beberapa informasi berupa atribut diantaranya id penulis ulasan, kutipan, *rating*, tanggal, dan isi ulasan. Untuk melakukan proses *scraping* dibutuhkan koneksi internet untuk menyambungkan *software R* ke situs *TripAdvisor*. Adapun proses *scraping data* dilakukan dengan beberapa langkah sebagai berikut:

1. *Install packege* 'rvest' pada *software R* dengan cara menjalankan *script* `install.packages("rvest")`. Setelah *package* terinstal pada *software R*, kemudian *package* tersebut di panggil dengan cara menjalankan perintah `library(rvest)`. Perintah "`library`" digunakan untuk mengaktifkan *package* yang ada pada *software R*.
2. Identifikasi *URL* halaman *website* ulasan The Phoenix Hotel pada situs *TripAdvisor*. Adapun *URL* halaman ulasan The Phoenix Hotel adalah:

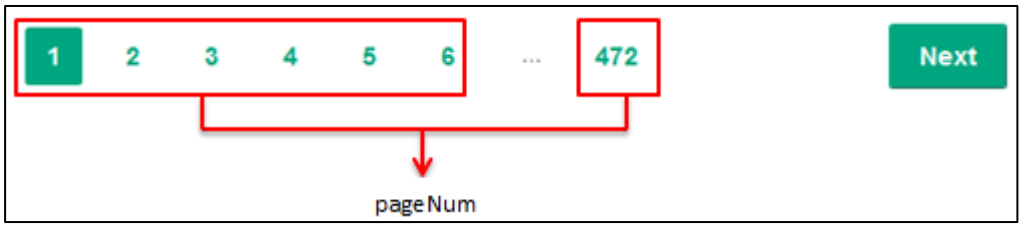
https://www.tripadvisor.com/Hotel_Review-g294230-d446805-Reviews-The_Phoenix_Hotel_Yogyakarta_MGallery_Collection-Yogyakarta_Java.html #REVIEWS

Untuk membaca situs tersebut ke dalam *software R*, dilakukan dengan cara menjalankan *script* berikut:

```
url<-read_html("https://www.tripadvisor.com/Hotel_Review-g294230-
d446805-Reviews-The_Phoenix_Hotel_Yogyakarta_MGallery_Collection-
Yogyakarta_Java.html")
```

Gambar 5.2 *Script R untuk membaca URL website*

3. Setiap halaman ulasan pada situs *TripAdvisor* hanya berisi 5 ulasan pengunjung, sehingga untuk mendapatkan semua data ulasan perlu dilakukan proses *looping* pada setiap halaman. Sebelum melakukan proses *looping*, terlebih dahulu perlu diketahui *nodes* atau *tag* yang mengapit informasi nomor halaman pada *website*. Dalam hal ini digunakan *selector gadget* untuk mendapatkan kode *CSS* nomor halaman. *Selector gadget* adalah *CSS selectors* yang digunakan untuk memudahkan dalam memilih elemen pada halaman *web* (Cantino, 2013). Untuk mendapatkan kode nomor halaman, dapat dilakukan dengan cara mengaktifkan *selector gadget* yang telah terpasang pada *browser google chrome* kemudian dilakukan penyeleksian data dengan cara mengklik indeks halaman hingga diperoleh kode *CSS* untuk letak nomor halaman. Dari hasil penyeleksian diperoleh kode *CSS* ".pageNum" yang menyatakan letak nomor halaman, seperti terlihat pada gambar berikut:



Gambar 5.3 *Kode CSS letak nomor halaman*

4. Mengidentifikasi semua halaman dan letak nomor halaman ke dalam bahasa pemrograman *R* dengan cara menjalankan perintah berikut:

```
npages<-url%>%
  html_nodes(" .pageNum")%>%
  html_attr(name="data-page-number")%>%
  tail(.,1)%>%
  as.numeric()
```

Gambar 5.4 *Script R untuk merecord nomor halaman*

5. Mencari indeks yang menyatakan halaman pada *URL* situs *TripAdvisor*, untuk mengetahui indeks tersebut dapat dilihat dengan cara membandingkan tiga buah *URL* halaman pertama, kedua dan halaman ketiga seperti berikut:

URL halaman pertama
https://www.tripadvisor.com/Hotel_Review-g294230-d446805-Reviews-The_Phoenix_Hotel_Yogyakarta_MGallery_Collection-Yogyakarta_Java.html
URL halaman kedua
https://www.tripadvisor.com/Hotel_Review-g294230-d446805-Reviews-or5-The_Phoenix_Hotel_Yogyakarta_MGallery_Collection-Yogyakarta_Java.html
URL halaman ketiga
https://www.tripadvisor.com/Hotel_Review-g294230-d446805-Reviews-or10-The_Phoenix_Hotel_Yogyakarta_MGallery_Collection-Yogyakarta_Java.html

Gambar 5.5 *Menemukan indeks nomor halaman*

Berdasarkan **Gambar 5.5** dapat diketahui bahwa nomor halaman kedua ditunjukkan dengan nama indeks "-or5-", dan nomor halaman ketiga ditunjukkan dengan nama indeks "-or10-". Indeks tersebut menunjukkan bahwa untuk setiap penambahan satu nomor halaman, indeks angka akan bertambah sebanyak 5, dan setiap indeks halaman memiliki angka yang berada satu tingkat di bawah nomor halaman. Sehingga untuk mengidentifikasi indeks nomor halaman dapat dilakukan dengan cara menjalankan perintah seperti gambar berikut:

```
a<-0:(npages-1)
res<-numeric(length=length(a))
for (i in seq_along(a)) {
res[i]<-a[i]*b
}
```

Gambar 5.6 *Script R untuk mengidentifikasi indeks nomor halaman*

6. Membuat sebuah *data frame* atau tabel sebagai tempat untuk menyusun data yang akan dilakukan *scraping* dengan nama "*tableout*" menggunakan perintah seperti gambar berikut:

```
tableout <- data.frame()
```

Gambar 5.7 *Script R untuk data frame*

7. Melakukan proses *looping* dengan cara merubah indeks angka halaman menjadi "`,i,`" pada url *website*, dengan cara menjalankan perintah berikut:

```
for(i in res){
cat(".")

url <- paste ("https://www.tripadvisor.com/Hotel_Review-
g294230-d446805-Reviews-or",i,"-
The_Phoenix_Hotel_Yogyakarta_MGallery_Collection-
Yogyakarta Java.html#REVIEWS",sep="")
```

Gambar 5.8 *Script R untuk melakukan proses looping pada semua halaman*

8. Setelah dilakukan *looping* pada semua halaman, proses selanjutnya adalah mempelajari dokumen *HTML* dari *website* yang akan diambil informasinya dari *tag HTML* yang mengapit data/informasi yang akan diambil. Untuk dapat mengetahui letak atau *tag* yang mengapit informasi, maka digunakan kembali *selector gadget*. Pada kasus ini *selector gadget* digunakan untuk mengetahui letak masing-masing informasi yakni tentang *id, quote, rating, date* dan *review* yang akan diambil datanya. Untuk mendapatkan data *id, quote, rating, date* dan *review* maka digunakan perintah seperti berikut:

```
reviews <- url %>%
    html() %>%
    html_nodes("#REVIEWS .innerBubble")
id <- reviews %>%
    html_node(".quote a") %>%
    html_attr("id")
quote <- reviews %>%
    html_node(".quote span") %>%
    html_text()
rating <- reviews %>%
    html_node(".rating .ui_bubble_rating") %>%
    html_attrs() %>%
    gsub("ui_bubble_rating bubble_", "", .) %>%
    as.integer() / 10
date <- reviews %>%
    html_node(".innerBubble, .ratingDate") %>%
    html_text()
review <- reviews %>%
    html_node(".entry .partial_entry") %>%
    html_text()
```

Gambar 5.9 *Script R untuk mengambil data id, quote, rating, date, dan review dari website*

Perintah `html_nodes` dan `html_node` digunakan untuk mengekstrak potongan dari dokumen *HTML* menggunakan pemilih *CSS*. Penggunaan perintah `html_nodes` diikuti dengan kode *CSS* yang menunjukkan letak informasi/data yang akan di ekstrak. Sedangkan perintah `html_text` dan `html_attr` digunakan untuk mengekstrak atribut, teks dan *tag* nama dari *HTML* (Wickham, 2016).

9. Data ulasan hasil *scraping* memiliki bentuk dan susunan yang tidak terstruktur dengan baik, maka perlu dilakukan penyusunan data agar diperoleh data dengan struktur yang lebih baik. Proses penyusunan data dilakukan dengan cara menghilangkan simbol `\n` (*enter*) yang dapat merusak susunan data, dan kemudian data disusun ke dalam bentuk *data frame* atau tabel. Proses penyusunan data tersebut dilakukan dengan menjalankan *script* berikut:

```
reviewnospace <- gsub("\n", "", review)
temp.tableout <- data.frame(id, quote, rating, date, reviewnospace)
tableout <- rbind(tableout,temp.tableout)
}
```

Gambar 5.10 *Script R untuk menyusun data scraping ke dalam bentuk tabel*

10. Setelah data diperoleh dalam bentuk *data frame*, selanjutnya data disimpan ke dalam folder penyimpanan komputer dengan format *.csv*, menggunakan perintah berikut:

```
write.csv(tableout, "E:/Bismillah TA/Data
Script/1.dataulasaneng.csv")
```

Gambar 5.11 *Script R untuk menyimpan data dalam format csv*

Berikut adalah contoh tampilan data yang diperoleh dari hasil *web scraping* pada situs *TripAdvisor*.

Tabel 5.1 *Data hasil web scraping berbahasa Inggris*

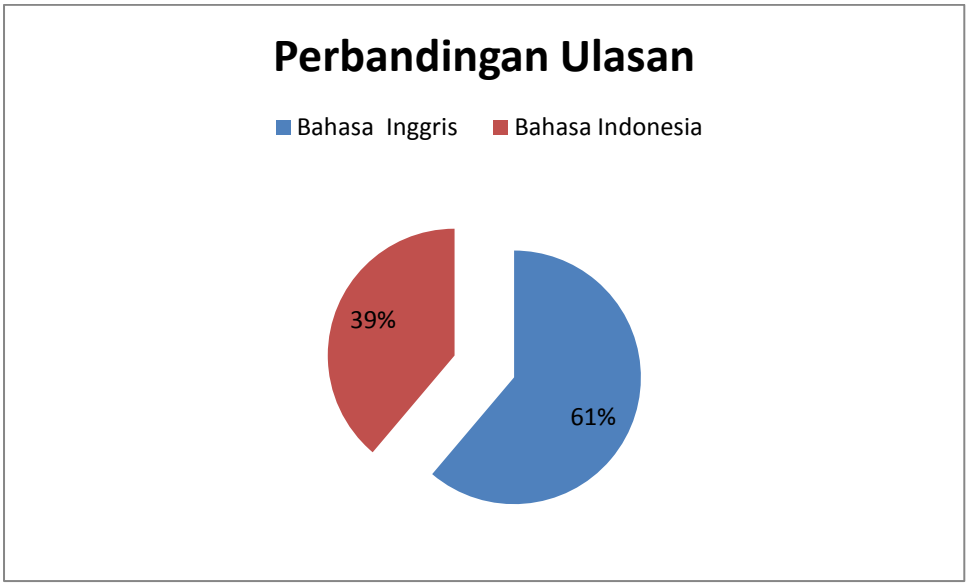
No	Id	Quote	Rating	Date	Reviewonspace
1	rn543516031	Heritage hotel with spacious rooms	4	26 November 2017	Phoenix Hotel has a great location. The staff are really friendly and helpful. From the check in to the restaurants, lounge and house keeping - everyone works to make things better for the guests.The rooms are spacious and have been recently done up.
2	rn543249240	Majestic, Old and quaint	5	25 November 2017	Heritage hotel of the by gone era. Breakfast buffet is outstanding with several nationalities of food. Rooms were recently updated. The pool was adequate. Hotel was within walking distance of shopping and dining
3	rn543243658	Stay in Phoenix Hotel Yogyakarta - MGallery Collection	5	25 November 2017	The hotel is well located in Yogyakarta. Our stay was very comfortable and the overall service, including meals and beverages, was excellent. If we come back again in Yogyakarta, it would be certainly an option to stay in that city.
4	rn543091440	Excellent	4	24 November 2017	The staff is exceptionaly friendly and mindful. The breakfast buffet is unmached. The outdoor pool is beautiful and the gym is well equipped. I you like hotels with historical charm and contemporary services - you will love The Phoenix Hotel Yogyakarta.
5	rn543023303	excellent hotel beyond expectations	5	24 November 2017	We had a great time in Yogjakarta. The hotel was part of that experience. Great food very nice and helpful staff! The hotel is situated close by the touristic areas and we had no problems reaching all the other attractions outside Yogjakarta

Tabel 5.2 *Data hasil web scraping berbahasa Indonesia*

No	Id	Quote	Rating	Date	Reviewonspace
1	rn527483311	Hotel Modern berbalut Klasik	4	26 September 2017	Hotel yang bisa dibilang modern dengan nuansa Klasik Khas Jogja... pelayanan yg baik, fasilitas lengkap (ada kolam renang), Kamar yg cukup luas juga kebersihannya cukup terjaga, letak yg cukup strategis tidak jauh dari Tugu Jogja serta Menu sarapan yg enak dan beragam membuat kita betah.
2	rn458249417	Heritage style	5	9 Februari 2017	Kamar dengan pool akses sangatlah nyaman, ditambah lagi dengan tempat tidur yang nyaman dan kamar yang sangat rapih dan juga bersih. Pilihan terbaik untuk keluarga maupun pasangan. Jaraknya juga dekat ke Tugu ataupun Malioboro.
3	rn523223844	asik klasik berkelas	4	10 September 2017	beberapa hari lalu nginep di hotel ini, well pelayanan bagus hotelnya juga bagus bersih. pokoknya nyaman ngine disini, dengan gaya klasiknya, bagus banget view nya. puas bisa nginep di hotel ini dah pokoknya.
4	rn521577523	Hotel Klasik dengan Perlengkapan Modern	4	6 September 2017	Hotelnya bagus, mudah pula mencarinya untuk yang belum pernah menginap disana. Nuansanya klasik dan unik. Kamarnya lumayan luas dan bersih. kolam renangnya juga nyaman, ada untuk anak2. Pelayanannya baik dan ramah. Proses check in dan check outnya cepat.
5	rn494401137	Tempatnya indah, bersih	4	20 Juni 2017	Lumayan lah nginep disana sehari sangat nyaman tempatnya, ada kolam renangnya juga.. sejuk. dan lagi kamarnya bersih dan rapi semua. pelayanannya juga ramah semua.. Saya sangat senang sekali bisa nginepa disini. besok lagi aku mau nginep juga disini.

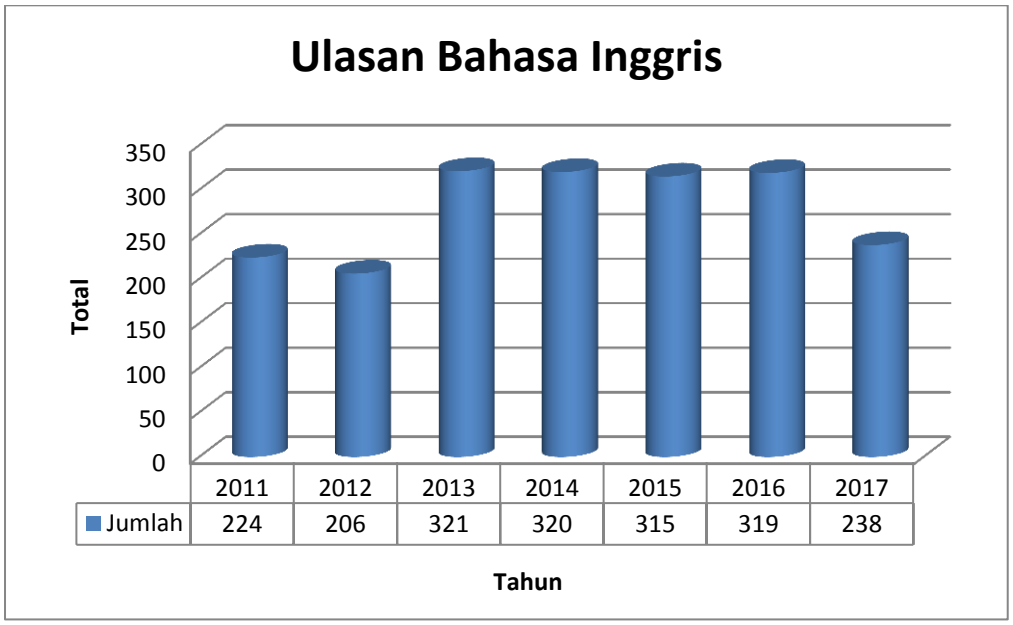
5.2 Analisis Deskriptif

Analisis deskriptif dalam penelitian ini digunakan untuk melihat gambaran secara umum informasi tentang The Phoenix Hotel Yogyakarta berdasarkan data ulasan pengunjung dari situs *TripAdvisor*, yang dilihat dari beberapa aspek diantaranya perbandingan jumlah ulasan antara ulasan berbahasa Inggris dan ulasan berbahasa Indonesia, jumlah ulasan yang masuk berdasarkan urutan waktu, dan rating hotel yang diberikan pengunjung.



Gambar 5.12 *Grafik perbandingan jumlah ulasan*

Berdasarkan **Gambar 5.12** dapat diketahui bahwa jumlah ulasan berbahasa Inggris dan Indonesia dalam situs *TripAdvisor* memiliki perbandingan 61% : 39%. Dari 3.176 ulasan, 1.943 data ulasan berbahasa Inggris dan sebanyak 1.233 data berbahasa Indonesia. Hal ini menggambarkan bahwa pengunjung hotel yang berasal dari mancanegara lebih aktif dalam memberikan *feedback* berbentuk ulasan/*review* berdasarkan pengalamannya selama menginap di The Phoenix Hotel Yogyakarta.



Gambar 5.13 *Silinder Bar Chart jumlah ulasan berbahasa Inggris berdasarkan urutan waktu*

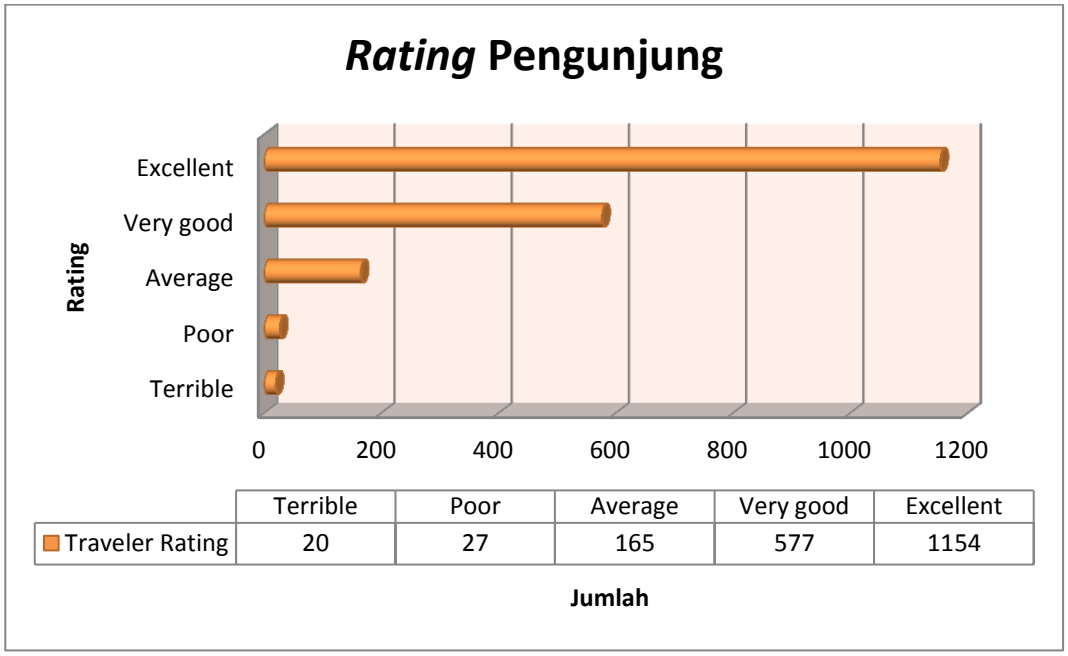
Gambar 5.13 menunjukkan grafik jumlah ulasan berbahasa Inggris yang masuk pada situs *TripAdvisor* berdasarkan urutan bulan dan tahun, yang terhitung sejak tahun 2011 hingga tahun 2017. Data jumlah ulasan pengunjung tersebut dihitung dengan cara menjumlahkan ulasan secara manual setiap setahun selama 7 tahun.

Berdasarkan **Gambar 5.13**, dapat diketahui jumlah ulasan dari tahun ke tahun cenderung mengalami fluktuasi. Dari tahun 2013, dapat dilihat perbedaan yang cukup signifikan bahwa jumlah ulasan yang masuk pada situs *TripAdvisor* memiliki rata-rata lebih tinggi jika dibandingkan dengan tahun sebelumnya yaitu tahun 2011 dan tahun 2012. Secara visual, pada tahun 2013 grafik menunjukkan bahwa kenaikan jumlah ulasan meningkat drastis, dengan jumlah ulasan yang masuk sebanyak 115 ulasan. Akan tetapi pada tahun berikutnya mengalami penurunan yang tidak terlalu signifikan yaitu 1 ulasan pada tahun 2014, 5 ulasan pada tahun 2015, dan 81 ulasan pada tahun 2017. Berdasarkan grafik tersebut, pola data jika dilihat menggunakan garis linier, maka pola data tersebut cenderung naik setiap waktunya walaupun ada beberapa data yang turun atau meningkat setiap pada waktu tertentu.

Gambar 5.14 *Silinder Bar Chart jumlah ulasan berbahasa Indonesia berdasarkan urutan waktu*

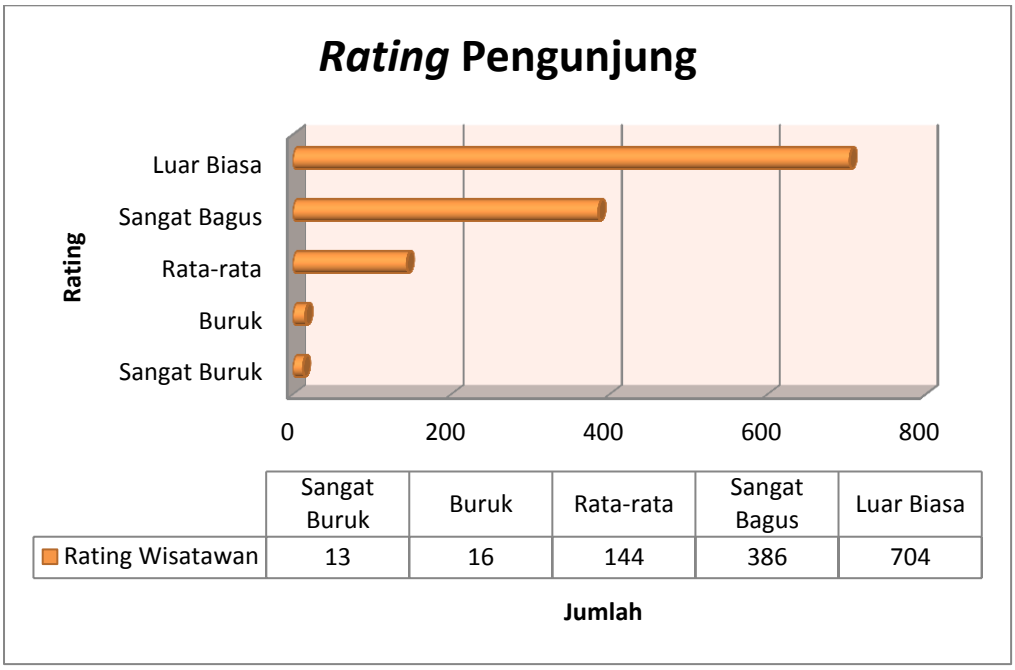
Gambar 5.14 menunjukkan grafik jumlah ulasan berbahasa Indonesia yang masuk pada situs *TripAdvisor* berdasarkan urutan bulan dan tahun, yang terhitung sejak tahun 2011 hingga tahun 2017. Data jumlah ulasan pengunjung tersebut dihitung dengan cara menjumlahkan ulasan secara manual setiap setahun selama 7 tahun.

Berdasarkan **Gambar 5.14**, dapat diketahui jumlah ulasan dari tahun ke tahun cenderung mengalami fluktuasi. Dari tahun 2014, dapat dilihat perbedaan yang cukup signifikan bahwa jumlah ulasan yang masuk pada situs *TripAdvisor* memiliki rata-rata lebih tinggi jika dibandingkan dengan tahun sebelumnya yaitu tahun 2011, 2012, dan 2013. Secara visual, pada tahun berikutnya grafik menunjukkan bahwa kenaikan jumlah ulasan meningkat drastis, dengan jumlah ulasan yang masuk masing-masing sebanyak 193 ulasan tahun 2014, 151 ulasan tahun 2015, dan 13 ulasan tahun 2016. Akan tetapi pada tahun berikutnya mengalami penurunan yang signifikan yaitu 147 ulasan pada tahun 2017. Berdasarkan grafik tersebut, pola data jika dilihat menggunakan garis linier, maka pola data tersebut cenderung naik setiap waktunya walaupun ada beberapa data yang turun atau meningkat setiap pada waktu tertentu.



Gambar 5.15 *Rating The Phoenix Hotel berdasarkan pengunjung situs TripAdvisor berbahasa Inggris*

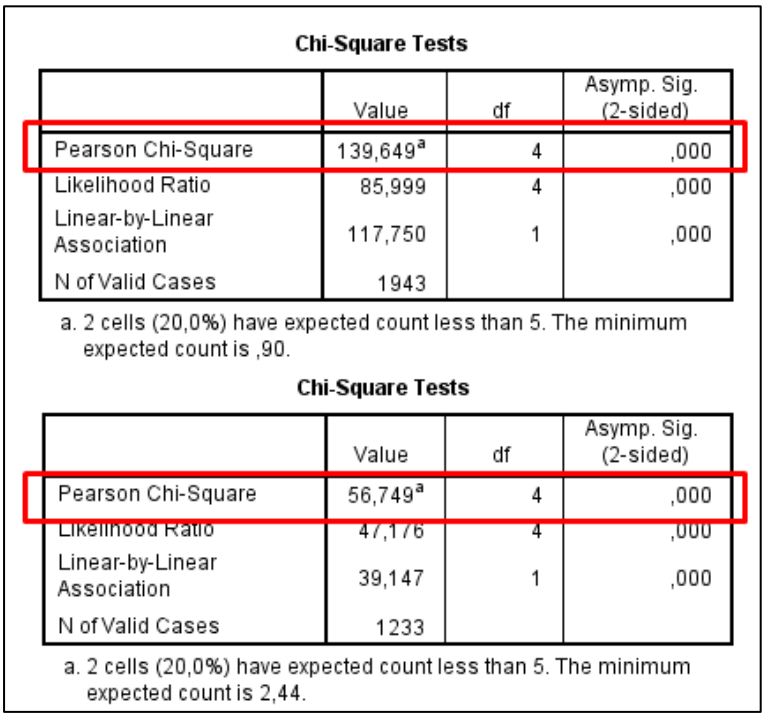
Gambar 5.15 menunjukkan *rating* The Phoenix Hotel Yogyakarta yang diperoleh berdasarkan penilaian pengunjung dari situs *TripAdvisor* dalam ulasan pengunjung yang berbahasa Inggris. *Rating* pada situs *TripAdvisor* mempunyai skala 1-5 yang secara berurutan mempunyai kategori "*terrible*", "*poor*", "*average*", "*very good*", dan "*excellent*". Berdasarkan **Gambar 5.15** dapat diketahui bahwa mayoritas pengunjung The Phoenix Hotel Yogyakarta mempunyai penilaian yang baik terhadap hotel. Hal ini terbukti berdasarkan jumlah penilaian pengunjung bahwa dari 1.943 ulasan, terdapat sebanyak 1.154 pengunjung memberikan penilaian *excellent* (luar biasa), 577 pengunjung memberikan penilaian *very good* (sangat bagus), 165 pengunjung memberikan penilaian *average* (rata-rata), sedangkan penilaian dengan kategori *poor* (buruk) hanya berjumlah 27 dan 20 ulasan dengan kategori *terrible* (sangat buruk).



Gambar 5.16 *Rating The Phoenix Hotel berdasarkan pengunjung situs TripAdvisor berbahasa Indonesia*

Gambar 5.16 menunjukkan *rating* The Phoenix Hotel Yogyakarta berdasarkan penilaian pengunjung dari situs *TripAdvisor* berbahasa Indonesia. Berdasarkan **Gambar 5.16** dapat diketahui bahwa mayoritas pengunjung The Phoenix Hotel Yogyakarta yang memberikan ulasan berbahasa Indonesia mempunyai penilaian yang baik terhadap hotel. Hal itu terbukti berdasarkan jumlah penilaian pengunjung bahwa dari 1.233 ulasan, terdapat sebanyak 704 pengunjung memberikan penilaian luar biasa, 386 pengunjung memberikan penilaian sangat bagus, 144 pengunjung memberikan penilaian rata-rata, sedangkan penilaian dengan kategori buruk hanya berjumlah 16 dan 13 ulasan dengan kategori sangat buruk.

Untuk mengetahui adanya hubungan antara *rating* dengan ulasan, peneliti melakukan uji independensi antara ulasan bahasa Inggris dan bahasa Indonesia sebagai berikut.



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	139,649[a]	4	,000
Likelihood Ratio	85,999	4	,000
Linear-by-Linear Association	117,750	1	,000
N of Valid Cases	1943		

a. 2 cells (20,0%) have expected count less than 5. The minimum expected count is ,90.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	56,749[a]	4	,000
Likelihood Ratio	47,176	4	,000
Linear-by-Linear Association	39,147	1	,000
N of Valid Cases	1233		

a. 2 cells (20,0%) have expected count less than 5. The minimum expected count is 2,44.

Gambar 5.17 *Uji Independensi antara variabel Rating dengan Ulasan Bahasa Inggris dan Indonesia*

Berikut adalah hipotesis untuk metode yang digunakan yaitu uji independensi menggunakan *Chi-Square*.

(i) Hipotesis

H_0 : Tidak ada hubungan antara *rating* dengan ulasan

H_1 : Ada hubungan antara *rating* dengan ulasan

(ii) Tingkat Signifikansi

$\alpha = 5\%$

(iii) Daerah Kritis

Tolak H_0 jika *P-Value* $< \alpha$

(iv) Statistik Uji

P-Value = 0,000

(v) Keputusan

Karena *P-Vaue* (0,000) $< \alpha$ (0,05) maka tolak H_0

(vi) Kesimpulan

Berdasarkan tingkat signifikan 95% maka dapat diambil kesimpulan bahwa terdapat hubungan antara *rating* dengan ulasan bahasa Inggris maupun bahasa Indonesia.

Berdasarkan hasil uji independensi bahwa *rating* dengan nilai 5 (Luar Biasa) cenderung lebih banyak daripada Sangat Buruk, Buruk, Rata-rata, dan Sangat Bagus. Hal ini kontras dengan kelas sentimen positif dari tiap ulasan di mana wisatawan cenderung memberikan ulasan yang positif. Sehingga terdapat hubungan antara *rating* dengan ulasan.

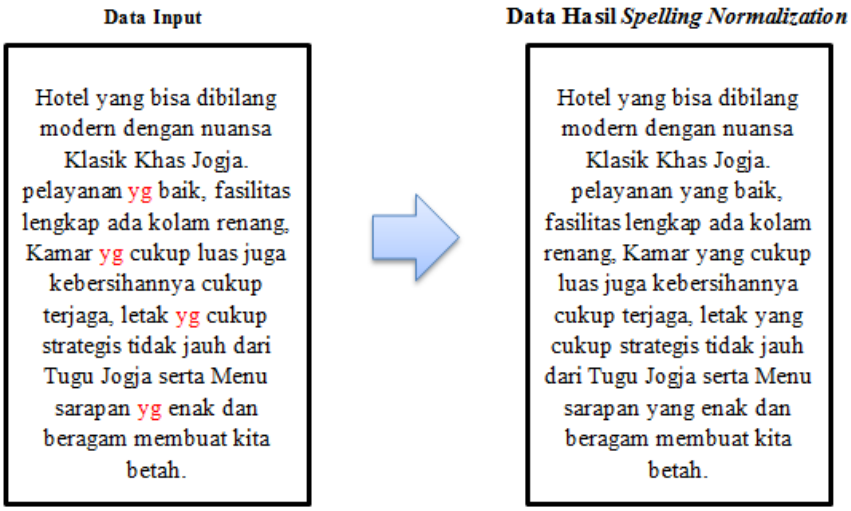
5.3 *Text Preprocessing*

Data ulasan yang diperoleh dari situs *TripAdvisor* merupakan data berupa teks yang memiliki bentuk data tidak struktur (*unstructured data*), karena masih terdapat banyak *noise* pada data dan informasi yang terdapat didalamnya akan sulit diekstrak secara langsung. Oleh sebab itu, data harus di seleksi terlebih dahulu agar lebih terstruktur dan memiliki keseragaman, sehingga akan mempermudah proses analisis dan ekstraksi informasi yang terkandung didalamnya.

Pada tahap *preprocessing*, akan dilakukan pembersihan data menggunakan metode *text mining.* Beberapa tahap yang akan dilakukan diantaranya adalah *spelling normalization, case folding, tokenizing,* dan *filtering* yang akan dijelaskan pada sub bab berikut:

5.3.1 *Spelling Normalization*

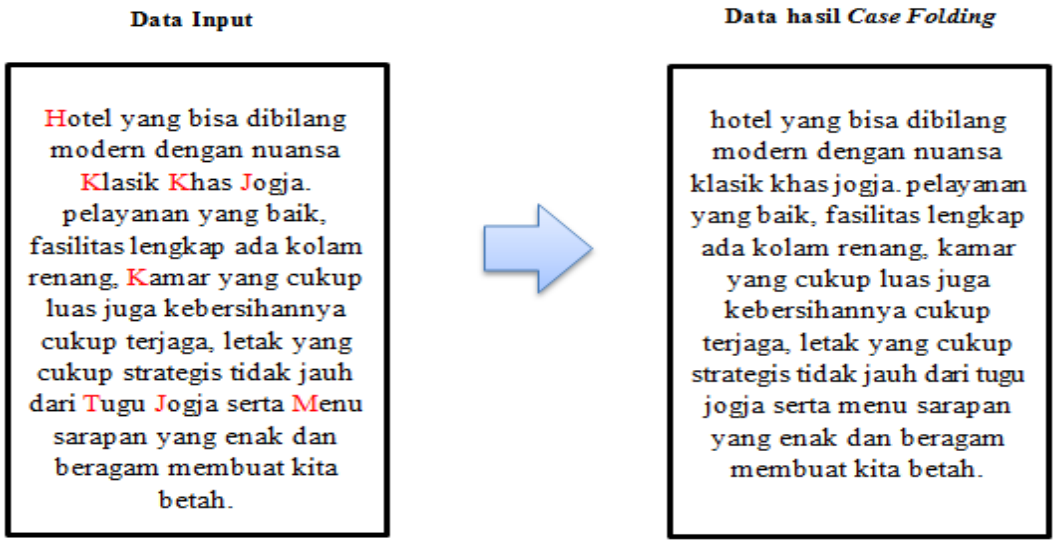
Spelling normalization, merupakan perbaikan kata-kata yang salah eja atau disingkat dengan bentuk tertentu. Misalnya kata “tidak” memiliki banyak bentuk penulisan seperti tdk, gak, nggak, enggak, dan banyak lainnya.



Gambar 5.18 *Proses spelling normalization*

5.3.2 *Case Folding*

Case folding merupakan tahapan yang mengubah semua huruf dalam dokumen menjadi huruf kecil, hanya huruf 'a' sampai dengan 'z' yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter (pembatas). Contoh penggunaan *case folding* dapat dilihat contoh pada **Gambar 5.19** berikut:

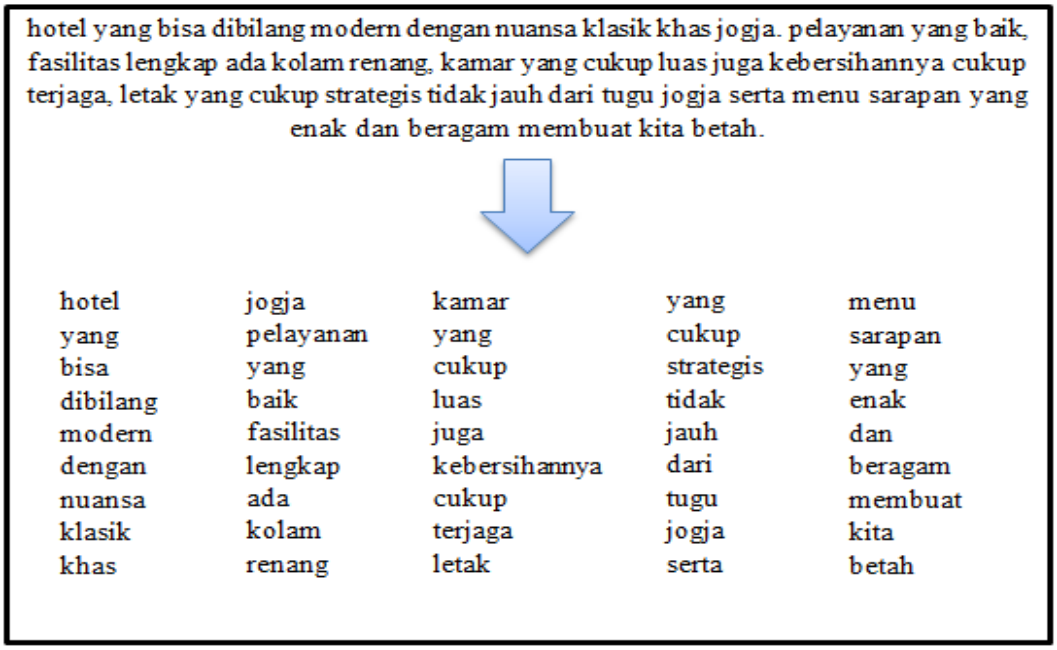


Gambar 5.19 *Proses case folding*

5.3.3 *Tokenizing*

Setelah dilakukan *case folding*, tahap selanjutnya adalah *tokenizing*, yaitu proses pemisahan teks menjadi potongan kata yang disebut token. *Tokenizing*

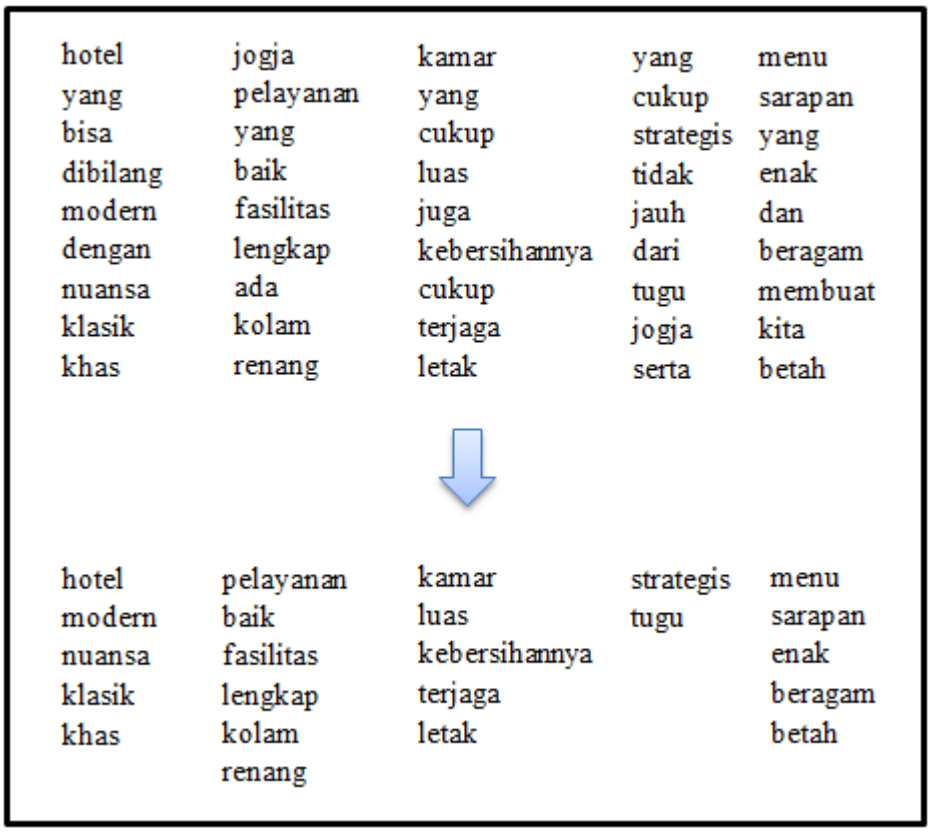
dilakukan untuk mendapatkan token atau potongan kata yang akan menjadi entitas yang memiliki nilai dalam penyusunan matriks dokumen pada proses selanjutnya. Langkah transformasi proses *tokenizing* ditunjukkan pada **Gambar 5.20** berikut:



Gambar 5.20 *Proses tokenizing*

5.3.4 *Filtering*

Tahap *filtering* adalah tahap mengambil kata-kata penting dari hasil *tokenizing*. Proses *filtering* dapat menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata penting). Dalam hal ini, penulis menggunakan *stopword* yang berbahasa Inggris (Bouge, 2011) dan *stopword* berbahasa Indonesia (Diaz, 2016). *Stopword* / *stoplist* adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. Contoh *stopword* dalam bahasa inggris adalah *"the", "and", "to", "i", "you", "was"* dan lain-lain. Selain dengan menggunakan *stopword*, proses *filtering* juga dilakukan secara manual, yakni dengan menghapus kata-kata yang tidak terdapat dalam daftar *stopword* yang dianggap kurang penting dan kurang berpengaruh secara signifikan terhadap hasil analisis. Contoh proses *filtering* dapat dilihat pada **Gambar 5.21** berikut:



Gambar 5.21 *Proses filtering*

5.4 Pelabelan Kelas Sentimen

Setelah melalui tahap *preprocessing*, selanjutnya dilakukan analisis sentimen untuk pelabelan data. Proses pelabelan data dilakukan secara otomatis dengan cara menghitung skor sentimen menggunakan kamus *lexicon.* Pada umumnya, analisis sentimen digunakan untuk melakukan klasifikasi (pelabelan) dokumen teks ke dalam tiga kelas sentimen, yaitu sentimen positif, negatif dan netral. Cara menentukan kelas sentimen adalah dengan menghitung skor jumlah kata positif dikurangi skor jumlah kata negatif dalam setiap kalimat ulasan (Susanti, 2016). Kalimat yang memiliki skor > 0 akan diklasifikasikan ke dalam kelas positif, kalimat yang memiliki skor = 0 akan diklasifikasikan ke dalam kelas netral, sedangkan kalimat yang memiliki skor < 0 diklasifikasikan ke dalam kelas negatif. Berikut ini adalah tahapan melakukan pelabelan dengan menggunakan *Software R.*

Tabel 5.3 *Tahap-tahap pelabelan menggunakan software R*

Script R	**Fungsi**

`library(tm)`
`setwd("E:/Bismillah TA/Data Script/")`
`kalimat2 <-read.csv`
`("datacleaning.csv",header=TRUE)`
	1. Menjalankan *packages "tm"* yang telah terinstal pada program R.
	2. Mengatur direktori kerja dalam program R.
	3. Membuka *file csv* yang akan diberi label .

`positif <- scan("positive-words.txt",`
`what="character",comment.char=";")`
`negatif <- scan("negative-words.txt",`
`what="character",comment.char=";")`
`kata.positif = c(positif, "is near to")`
`kata.negatif = c(negatif, "cant")`
	4. Melakukan *scanning file* daftar kata positif dan kata negatif yang tersimpan dalam format *.txt file*

`score.sentiment = function(kalimat2,`
`kata.positif, kata.negatif,`
`.progress='none')`
`{`
`require(plyr)`
`require(stringr)`
`scores = laply(kalimat2,`
`function(kalimat, kata.positif,`
`kata.negatif) {`
`kalimat = gsub('[[:punct:]]', '',`
`kalimat)`
`kalimat = gsub('[[:cntrl:]]', '',`
`kalimat)`
`kalimat = gsub('\\d+', '', kalimat)`
`kalimat = tolower(kalimat)`
 	
`list.kata = str_split(kalimat, '\\s+')`
`kata2 = unlist(list.kata)`
`positif.matches = match(kata2,`
`kata.positif)`
`negatif.matches = match(kata2,`
`kata.negatif)`
`positif.matches = !is.na(positif.matches)`
`negatif.matches = !is.na(negatif.matches)`
`score = sum(positif.matches) -`
	5. Melakukan proses skoring menggunakan *function* dengan tahapan:
	a. Menjalankan packages *plyr* dan *stringr*
	b. Menggabungkan setiap daftar inisial menjadi sebuah *array*
	c. Menghapus *noise* dan melakukan *case folding*
	d. Merubah kalimat menjadi potongan kata (*tokenizing*) dan menyederhanakan daftar kata
	e. Mengidentifikasi kata positif dan kata negatif pada setiap potongan kata
	f. Mengindikasi kata positif dan kata negatif ke dalam bentuk logika
	g. Menghitung jumlah skor sentimen
	h. Menyimpan skor dan kalimat ke dalam bentuk

Script R	Fungsi
(sum(negatif.matches)) return(score) }, kata.positif, kata.negatif, .progress=.progress) scores.df = data.frame(score=scores, text=kalimat2) return(scores.df) }	tabel
hasil = score.sentiment(kalimat2$text, kata.positif, kata.negatif) hasil$klasifikasi<- ifelse(hasil$score<0, "Negatif","Positif") hasil$klasifikasi data <- hasil[c(3,1,2)] write.csv(data, file = "Hasil_pelabelan.csv")	6. Memanggil *function* hasil skoring yang telah dibuat 7. Melakukan konversi nilai skor ke dalam kelas positif dan negatif 8. Menyimpan file hasil pelabelan ke dalam format *csv*.

Adapun hasil pelabelan kelas sentimen diperoleh perbandingan jumlah data seperti berikut:

Tabel 5.4 *Perbandingan jumlah data pada kelas sentimen*

Sentimen	Bahasa Inggris	Bahasa Indonesia
Positif	1.856	1.002
Negatif	87	231

Dalam penelitian ini, hanya akan digunakan dua pelabelan kelas sentimen, yaitu sentimen positif dan sentimen negatif. Dari total ulasan sebanyak 3.176 data, 1.943 diantaranya berbahasa Inggris dan 1.233 data berbahasa Indonesia. Berdasarkan **Tabel 5.4**, hasil pelabelan kelas sentimen menunjukkan bahwa jumlah ulasan positif memiliki frekuensi yang lebih tinggi dibandingkan dengan jumlah ulasan negatif. Jumlah ulasan positif berbahasa Inggris adalah sebanyak 1.856 ulasan, dan ulasan negatif adalah sebanyak 87 ulasan. Sedangkan untuk ulasan berbahasa Indonesia, diperoleh hasil ulasan positif sebanyak 1.002 ulasan, dan ulasan negatif sebanyak 231 ulasan. Hasil pelabelan data ulasan dapat dilihat pada **Tabel 5.5** berikut:

Tabel 5.5 *Hasil pelabelan kelas sentimen berbasis kamus lexicon dan proses manual*

Ulasan	Klasifikasi	Score	Text
Bahasa Inggris	Positif	7	lovely amazing staff great pool lovely rooms great buffet breakfasts dinners lots fresh fruit juices reliable transfers criticism spa ladies waxed enjoyable stay
	Negatif	-3	good location historic building indifferent front desk staff check making feel poor quality service restaurant incompetent staff unlucky confronted low quality staff
Bahasa Indonesia	Positif	6	hotelnya bagus mudah mencarinya disana nuansanya klasik unik kamarnya lumayan luas bersih kolam renangnya nyaman anak pelayanannya ramah proses check in check outnya cepat
	Negatif	-6	air panas colokan mematikan lampu mengisi ulang elektronik perangkat tidur lampu menyala mengerikan layanan penagihan dll mandi membutuhkan renovasi terburuk mandi berada di membutuhkan renovasi yang terburuk

Berdasarkan teks ulasan *"good location historic building indifferent front desk staff check making feel poor quality service restaurant incompetent staff unlucky confronted low quality staff"*, terdapat 4 kata negatif dan1 kata positifyang terdeteksi pada kamus *lexicon*, yakni *"poor", "incompetent", "unlucky",* dan *"low"* sebagai kata negatif, dan *"good"* sebagai kata positif. Adapun rumus perhitungan skor sentimen yang digunakan dalam proses pelabelan adalah sebagai berikut:

$$Skor = (Jumlah\ kata\ positif) - (Jumlah\ kata\ negatif) \quad (5.1)$$

Tabel 5.6 *Simulasi perhitungan skor sentimen*

Teks Ulasan	Kata Positif	Kata Negatif
good location historic building indifferent front desk staff check making feel poor quality service restaurant incompetent staff unlucky confronted low quality staff	Good	Poor incompetent unlucky low
Jumlah	**1**	**4**

Sehingga dengan demikian diperoleh perhitungan sebagai berikut :

$Skor = (Jumlah\ kata\ positif) - (Jumlah\ kata\ negatif)$

$Skor = (1) - (4)$

$Skor = -3$

Skor akhir yang diperoleh dari simulasi perhitungan bernilai < 0, sehingga hasil klasifikasi ulasan adalah negatif. Banyaknya ulasan positif menunjukkan bahwa pengunjung The Phoenix Hotel Yogyakarta memiliki persepsi yang baik terhadap hotel. Selain itu, hal tersebut juga disebabkan karena kebanyakan pengunjung tidak secara spontan memberikan ulasan negatif, melainkan ulasan negatif diberikan setelah didahului oleh kalimat berupa ulasan positif. Sehingga, pada saat proses pelabelan, kata-kata positif lebih mendominasi bila dibandingkan dengan kata-kata negatif yang hasilnya dapat memberikan skor bernilai positif.

5.5 Pembuatan Data Latih dan Data Uji

Data latih digunakan oleh algoritma klasifikasi untuk membentuk sebuah model *classifier*, model ini merupakan representasi pengetahuan yang akan digunakan untuk prediksi kelas data baru yang belum pernah ada, semakin besar data latih yang digunakan, maka akan semakin baik *machine* dalam memahami pola data. Data uji digunakan untuk mengukur sejauh mana *classifier* berhasil

melakukan klasifikasi dengan benar. Data yang digunakan untuk data latih dan data uji adalah data yang telah memiliki label kelas, dengan jumlah data latih dan data uji digunakan dengan menggunakan rumus *Slovin*. Perbandingan jumlah data latih dan data uji dapat dilihat pada **Tabel 5.7** untuk ulasan berbahasa Inggris dan **Tabel 5.8** untuk ulasan berbahasa Indonesia.

Tabel 5.7 *Perbandingan data latih dan data uji pada ulasan berbahasa Inggris*

Klasifikasi	Jumlah	Data Latih	Data Uji
Positif	1856	1527	329
Negatif	87	71	16
Total	**1943**	**1598**	**345**

Tabel 5.8 *Perbandingan data latih dan data uji pada ulasan berbahasa Indonesia*

Klasifikasi	Jumlah	Data Latih	Data Uji
Positif	1002	716	286
Negatif	231	146	85
Total	**1233**	**862**	**371**

Berdasarkan **Tabel 5.7** dan **Tabel 5.8**, dengan perbandingan data latih dan data uji sebesar 1943 data ulasan berbahasa Inggris, digunakan sebanyak 1598 data sebagai data latih dan 345 data sebagai data uji. Sedangkan pada data ulasan berbahasa Indonesia, dari total data sebanyak 1233 ulasan, digunakan 862 ulasan sebagai data latih dan 371 ulasan sebagai data uji.

Proses pembuatan data latih dan data uji dilakukan secara manual dengan cara mengacak keseluruhan data pada masing-masing kelas, sehingga setiap kalimat mempunyai peluang untuk menjadi data latih dan data uji. Pengacakan tidak dilakukan pada keseluruhan data tanpa memperhatikan kelas data, hal ini dilakukan karena proporsi data yang tidak seimbang, yakni jumlah data pada kelas positif jauh lebih banyak dibandingkan dengan jumlah data pada kelas negatif.

5.6 Klasifikasi dengan *Support Vector Machine*

Proses klasifikasi dilakukan dengan cara mempelajari pola data menggunakan data latih. Data latih yang didalamnya terdapat data latih ulasan positif dan data latih ulasan negatif digunakan oleh algoritma *SVM* dalam mempelajari pola data berdasarkan ciri-ciri data pada masing-masing kelas. Hasil pembelajaran pada algoritma *SVM* kemudian dapat di uji menggunakan data uji, sehingga dapat di ukur tingkat akurasinya dalam memprediksi kelas pada data baru, proses inilah yang selanjutnya disebut sebagai *machine learning*.

Pada penelitian ini dilakukan percobaan menggunakan beberapa metode *kernel* pada *SVM* untuk memperoleh klasifikasi dengan hasil akurasi terbaik, beberapa metode tersebut diantaranya *kernel Linear, Polynomial, Radial Basis Function (RBF),* dan *Sigmoid.* Percobaan klasifikasi dilakukan menggunakan sampel ulasan berbahasa Inggris, adapun hasil perbandingan dapat dilihat pada **Tabel 5.9** berikut:

Tabel 5.9 *Perbandingan penggunaan metode kernel pada klasifikasi SVM*

Kernel	Akurasi
Linear	95,47%
Polynomial	95,17%
RBF	96,07%
Sigmoid	96,07%

Pada umumnya permasalahan kategori teks dapat terpisah secara linier (Kaestner, 2013). Berdasarkan perbandingan menggunakan beberapa metode *kernel* pada **Tabel 5.9** diatas, menunjukkan bahwa *kernel RBF* dan *Sigmoid* memiliki tingkat akurasi yang lebih baik dibandingkan metode *kernel* yang lain, sehingga pada penelitian ini digunakan *kernel Sigmoid* dalam melakukan proses klasifikasi. Proses klasifikasi menggunakan algoritma *SVM* dilakukan dengan menggunakan *software R.* Adapun tahap-tahap melakukan klasifikasi ditampilkan dalam **Tabel 5.10** berikut:

Tabel 5.10 *Tahap melakukan analisis SVM dengan software R*

Script R	**Fungsi**

`setwd("E:/Bismillah TA/Data Script/SVM Sari")`
 	
`positif = readLines("PL.csv")`
`negatif = readLines("NL.csv")`
`sari.tr = c(positif, negatif)`
`positiftes = readLines("PT.csv")`
`negatiftes = readLines("NT.csv")`
`sari.ts = c(positiftes, negatiftes)`
	1. Mengatur direktori kerja pada program R
	2. Membuka *file* data latih dan data uji dalam format *csv*

`sentiment=c(rep("positif", length (positif) ), rep("negatif", length(negatif)))`
`sentiment_test=c(rep("positif", length(positiftes) ), rep ("negatif", length(negatiftes)))`
`sentiment_all=as.factor(c(sentiment, sentiment_test))`
	3. Menggabungkan data latih dan data uji yang telah terdefinisi
	4. Mendefinisikan label kelas pada data latih dan data uji
	5. Menggabungkan label kelas dan mengubah tipe data menjadi tipe data faktor

`library(RTextTools)` `library(e1071)`	6. Menjalankan *packages RTextTools* dan *e1071*
`mat = create_matrix(sari.all, language="english", removeStopwords = FALSE, removeNumbers = TRUE, stemWords = FALSE, tm::weightTfIdf)` `mat = as.matrix(mat)`	7. Membuat objek kelas *DocumentTermMatrix* 8. Mengubah data ke dalam bentuk matrix

`container <- create_container(mat, sentiment_all, trainSize=1:1612,testSize=1613:1943, virgin=FALSE)`
`model <- train_model(container, 'SVM',kernel='linear')`
`results <- classify_model(container, model)`
`table(as.character(sentiment_all [1613:1943]), as.character(results [,"SVM_LABEL"]))`
	9. Membuat wadah untuk proses training dan testing data
	10. Melakukan *training* untuk mendapatkan model dengan algoritma *SVM*
	11. Menggunakan model data training untuk mengklasifikasikan data baru
	12. Membuat tabel *confusion matrix*

`recall_accuracy(sentiment_all [1613:1943], results [,"SVM_LABEL"])`	13. Menghitung nilai akurasi

Penelitian ini menggunakan evaluasi model *confusion matrix* untuk mengetahui hasil akurasi klasifikasi. Untuk melakukan evaluasi model, pada percobaan ini

dilakukan dengan membuat 5 buah *machine learning* untuk menemukan nilai akurasi prediksi terbaik. Adapun hasil masing-masing percobaan *machine learning* menggunakan metode *Support Vector Machine* adalah sebagai berikut:

Tabel 5.11 *Perbandingan nilai akurasi machine learning dengan metode SVM*

Machine Learning	Akurasi Metode	
	Ulasan bahasa Inggris	Ulasan bahasa Indonesia
Machine Learning 1	96,07%	84,77%
Machine Learning 2	96,07%	82,78%
Machine Learning 3	95,77%	82,78%
Machine Learning 4	96,07%	82,45%
Machine Learning 5	96,07%	83,77%

Berdasarkan **Tabel 5.11** diatas, dari 5 percobaan *machine learning* yang dilakukan menggunakan metode *SVM*, pada ulasan berbahasa Inggris, *machine learning 1*, *2*, *4*, dan *5* menghasilkan tingkat akurasi tertinggi yakni sebesar 96,07%. Sedangkan pada ulasan berbahasa Indonesia, *machine learning 1* menghasilkan tingkat akurasi tertinggi sebesar 84,77%. Hasil perhitungan tingkat akurasi diperoleh dari jumlah data uji yang terklasifikasi dengan benar dibandingkan dengan total semua data yang di uji.

Untuk menguji performa *machine* dalam melakukan klasifikasi, maka dilakukan *cross validation* menggunakan *5-fold cross validation* dengan hasil rata-rata akurasi diperoleh sebesar 96,01% untuk ulasan berbahasa Inggris dan 83,31% untuk ulasan berbahasa Indonesia.

Untuk memudahkan proses perhitungan akurasi, maka digunakan *confusion matrix* untuk mengetahui jumlah data uji yang terklasifikasi dengan benar dan jumlah data uji yang salah pengklasifikasiannya. Adapun *confusion matrix* yang diperoleh pada *machine learning 1* untuk teks bahasa Inggris dan teks bahasa Indonesia dapat dilihat pada **Tabel 5.12**.

Tabel 5.12 *Confusion matrix*

Prediksi	Aktual (B.Inggris)		Class Precision	Aktual (B.Indonesia)		Class Precision
	Positif	Negatif		Positif	Negatif	
Positif	315	0	**100%**	211	6	**97%**
Negatif	13	3	**81%**	40	45	**47%**
Class Recall	**96%**	**100%**		**84%**	**12%**	
Akurasi = 96,07%				**Akurasi = 84,77%**		

Berdasarkan **Tabel 5.12**, dengan menggunakan metode *Support Vector Machine* untuk ulasan berbahasa Inggris diperoleh hasil prediksi bahwa pada kelas positif, dari 315 ulasan positif, terdapat 13 kesalahan prediksi, artinya sebanyak 302 ulasan positif dapat terklasifikasi dengan benar sebagai ulasan positif sehingga diperoleh nilai *recall* untuk kelas positif sebesar 96%. Sedangkan pada ulasan negatif, dari 3 ulasan terdapat 3 ulasan yang terklasifikasi dengan benar sebagai ulasan negatif dan tidak terdapat kesalahan prediksi yang masuk ke dalam ulasan positif, sehingga diperoleh nilai *recall* kelas negatif sebesar 100%. Hasil klasifikasi dengan metode *SVM* diperoleh tingkat akurasi sebesar 96,07%, artinya dari 331 data ulasan yang diuji, terdapat 331 ulasan yang benar pengklasifikasiannya oleh model *SVM*.

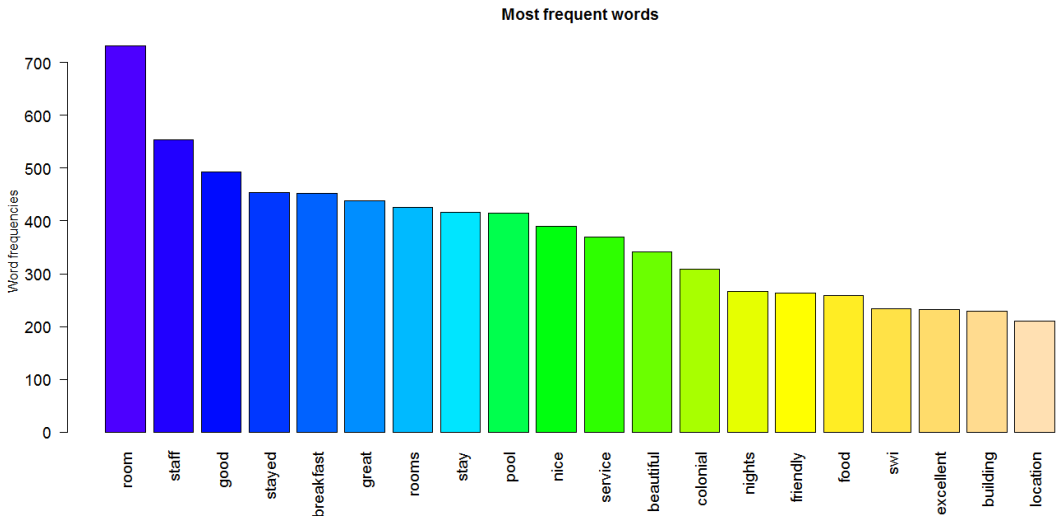
Pada ulasan berbahasa Indonesia, diperoleh hasil prediksi bahwa pada kelas positif, dari 211 ulasan positif, terdapat 40 kesalahan prediksi, artinya sebanyak 171 ulasan positif dapat terklasifikasi dengan benar sebagai ulasan positif sehingga diperoleh nilai *recall* untuk kelas positif sebesar 84%. Sedangkan pada ulasan negatif, dari 51 ulasan terdapat 45 ulasan yang terklasifikasi dengan benar sebagai ulasan negatif dan terdapat kesalahan prediksi sebanyak 6 ulasan yang masuk ke dalam ulasan positif sehingga diperoleh nilai *recall* kelas negatif sebesar 12%. Tingkat akurasi pada ulasan berbahasa Indonesia adalah sebesar 84,77%, artinya dari 302 data ulasan yang diuji, terdapat 296 ulasan yang benar pengklasifikasiannya oleh model *SVM*. Tingkat akurasi bahasa Indonesia lebih rendah dibandingkan dengan bahasa Inggris karena pendeteksian kamus yang tidak seimbang.

5.7 Visualisasi dan Asosiasi

Visualisasi yang dilakukan pada penelitian ini masing-masing diklasifikasikan yaitu berbahasa Inggris dan bahasa Indonesia. Adapun tujuan visualisasi adalah untuk mengekstraksi informasi berupa topik yang paling sering di bicarakan/di ulas oleh pengunjung hotel, sehingga dari sekian banyak teks ulasan yang ada, dapat diambil informasi yang dianggap penting serta dicari asosiasi antar kata yang paling sering muncul secara bersamaan, sehingga mampu memperkuat pencarian informasi tersebut. Berikut akan dijelaskan hasil visualisasi dan asosiasi kata dari setiap klasifikasi ulasan bahasa Inggris dan bahasa Indonesia.

5.7.1 Ulasan Positif

Data ulasan positif yang digunakan adalah hasil pelabelan menggunakan analisis sentimen berbasis *lexicon.* Ekstraksi informasi pada ulasan positif dilakukan secara berulang-ulang hingga mendapatkan informasi tentang ulasan positif pengunjung The Phoenix Hotel Yogyakarta yang paling sering di ulas/ dibicarakan. Ulasan positif tersebut di identifikasi berdasarkan frekuensi kata dalam ulasan, berikut adalah visualisasi hasil ekstraksi informasi yang didapatkan dari ulasan pengunjung dengan klasifikasi ulasan positif.



Gambar 5.22 *Kata yang paling banyak muncul dari kelas positif berbahasa Inggris*

Pada hasil klasifikasi ulasan positif berbahasa Inggris, dari jumlah ulasan positif sebanyak 1.856 ulasan, diperoleh beberapa kata yang paling banyak muncul diantaranya adalah kata *"room"* dengan frekuensi sebanyak 731 kali, *"staff"* sebanyak 553 kali, *"good"* 493 kali, *"stayed"* 454 kali dan seterusnya. Kata-kata yang muncul seperti pada **Gambar 5.22** merupakan kata yang memiliki sentimen positif berbahasa Inggris dan merupakan topik pembicaraan yang paling banyak di ulas oleh pengunjung The Phoenix Hotel Yogyakarta. Kata-kata tersebut selanjutnya digunakan sebagai dasar untuk menemukan asosiasi dengan kata lainnya, sehingga dapat diperoleh informasi yang lebih baik. Kumpulan kata-kata yang sering muncul tersebut dapat ditampilkan dalam bentuk *wordcloud* seperti terlihat pada **Gambar 5.23.**

(a) **(b)**

Gambar 5.23 *Wordcloud ulasan positif berbahasa Inggris*

Pada visualisasi *wordcloud* dapat di lihat dengan lebih jelas topik dan kata-kata positif yang sering digunakan pengunjung dalam memberikan ulasan. Semakin besar ukuran kata pada *wordcloud* menggambarkan semakin tinggi pula frekuensi kata tersebut, artinya semakin sering pengunjung menggunakan kata tersebut sebagai topik pembicaraan atau penilaian positif dalam ulasan. *Wordcloud* pada **Gambar 5.23 (a)** merupakan *wordcloud* pada ulasan positif yang didalamnya mengandung kata *"room"* dan "staff", sedangkan **Gambar 5.23 (b)** merupakan *wordcloud* pada ulasan positif tanpa menggunakan kata *"room"* dan *"staff"*. Selanjutnya, dilakukan pencarian asosiasi antar kata yang sering muncul secara bersamaan dan diperoleh hasil sebagai berikut:

Tabel 5.13 *Asosiasi kata pada klasifikasi positif berbahasa Inggris*

room		**staff**		**Good**		**stayed**	
deluxe	0,24	friendly	0,45	food	0,19	nights	0,44
balcony	0,22	helpful	0,39	location	0,19		
view	0,20			wine	0,16		
superior	0,20						
bed	0,17						
facing	0,16						
size	0,16						

breakfast		**Great**		**rooms**	
buffet	0,38	location	0,2	beds	0,17
variety	0,18			bathrooms	0,15
excellent	0,17				
choices	0,16				
spread	0,16				

Berdasarkan **Tabel 5.13**, diperoleh beberapa asosiasi kata pada klasifikasi kelas positif. Proses ekstraksi informasi dengan asosiasi dilakukan secara berulang-ulang dengan cara menyaring kata-kata yang memiliki hubungan dengan kata lain dan didasarkan pada relevansi kata dengan topik yang di ulas. Dari **Tabel 5.13** diatas, jika dilihat asosiasi kata yang berkaitan dengan kata *"room"*, dapat diperoleh informasi tentang kamar dengan tempat tidur yang memiliki tipe *deluxe* dan *superior* terdapat *view* yang menghadap ke balkon.

Kata-kata yang berasosiasi dengan kata *"staff"* memberikan informasi tentang kinerja staff hotel yang dinilai ramah dan sangat membantu terhadap pengunjung hotel.

Kata-kata yang berasosiasi dengan kata *"good"* memberikan informasi tentang baik atau bagusnya makanan yang disediakan oleh hotel, lokasi hotel yang bagus, dan tersedianya *wine* oleh hotel.

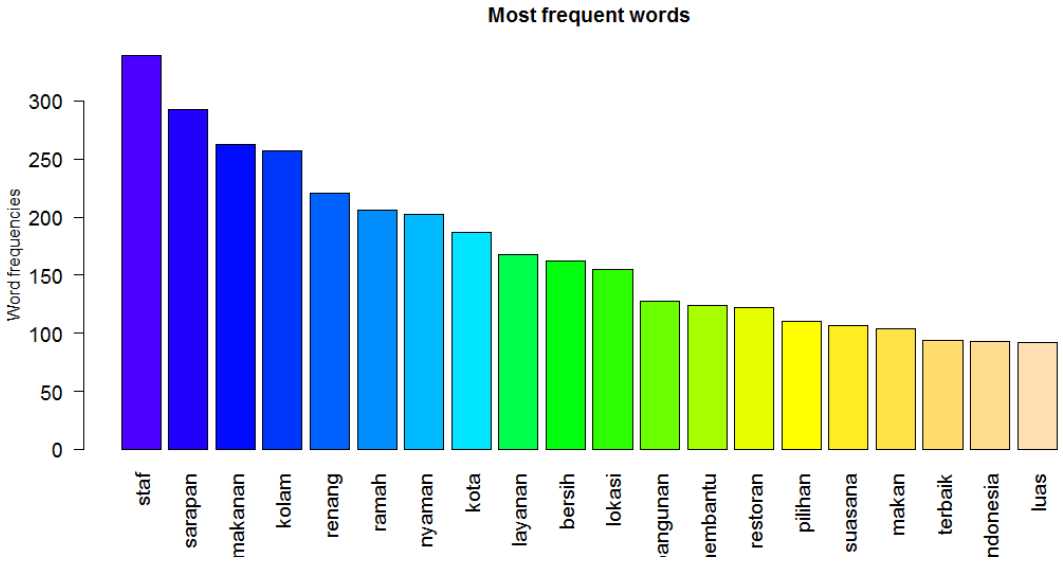
Kata-kata yang berasosiasi dengan *"stayed"* memberikan informasi tentang pengunujung yang menginap di hotel ini sampai bermalam lebih dari semalam.

Kata-kata yang berasosiasi dengan kata *"breakfast"* memberikan informasi tentang menu sarapan yang disajikan secara prasmanan yang banyak variasi

makanan yang dinilai sempurna dan pemilihan sebaran atau tatanan makanan yang tepat.

Kata-kata yang berasosiasi dengan kata *"great"* memberikan informasi tentang lokasi hotel yang strategis.

Kata-kata yang berasosiasi dengan kata *"rooms"* memberikan informasi tentang ruangan yang memiliki temapat tidur dan kamar mandi di dalamnya.



Gambar 5.24 *Kata yang paling banyak muncul dari kelas positif berbahasa Indonesia*

Pada hasil klasifikasi ulasan positif berbahasa Indonesia, diperoleh beberapa kata yang paling banyak muncul diantaranya adalah kata "staf" dengan frekuensi sebanyak 339 kali, "sarapan" sebanyak 293 kali, "makanan" sebanyak 263 kali, "kolam" 257 kali, "renang" 221 kali, dan seterusnya. Kata-kata yang muncul seperti pada **Gambar 5.25** merupakan kata yang memiliki sentimen positif berbahasa Indonesia dan merupakan topik pembicaraan yang paling banyak di ulas oleh pengunjung. Kumpulan kata-kata yang sering muncul tersebut kemudian ditampilkan dalam bentuk *wordcloud* seperti terlihat pada **Gambar 5.25**.

(a) **(b)**

Gambar 5.25 *Wordcloud ulasan positif berbahasa Indonesia*

Pada visualisasi *wordcloud* dapat di lihat dengan lebih jelas topik dan kata-kata positif yang sering digunakan pengunjung dalam memberikan ulasan. Semakin besar ukuran kata pada *wordcloud* menggambarkan semakin tinggi pula frekuensi kata tersebut, artinya semakin sering pengunjung menggunakan kata tersebut sebagai topik pembicaraan atau penilaian positif dalam ulasan. *Wordcloud* pada **Gambar 5.25 (a)** merupakan *wordcloud* pada ulasan positif yang didalamnya mengandung kata "staf" dan "sarapan", sedangkan **Gambar 5.25 (b)** merupakan *wordcloud* pada ulasan positif tanpa menggunakan kata "staf" dan "sarapan". Selanjutnya, dilakukan pencarian asosiasi antar kata yang sering muncul secara bersamaan dan diperoleh hasil sebagai berikut:

Tabel 5.14 *Asosiasi kata pada klasifikasi positif berbahasa Indonesia*

staf		sarapan		Makanan		kolam	
membantu	0,35	prasmanan	0,36	ringan	0,25	menghadap	0,35
sopan	0,21	pilihan	0,31	lokal	0,20	balkon	0,28
perhatian	0,20	jamu	0,17	minuman	0,19	pemandangan	0,20
senior	0,17	booking	0,17	gamelan	0,17	view	0,18
profesional	0,16	usahakan	0,17	javanish	0,17	renangnya	0,16
inggris	0,16	mencintai	0,17	ide	0,17		
				memasak	0,17		

Tabel 5.14 menunjukkan asosiasi antar kata pada ulasan positif berbahasa Indonesia, kata-kata tersebut merupakan topik yang paling sering dibicarakan pengunjung dalam ulasannya. Berdasarkan tabel tersebut dapat diketahui beberapa informasi sebagai berikut:

Kata "staf" memiliki asosiasi dengan kata "membantu", "sopan", "perhatian", "senior", "profesional", dan "inggris". Dari keenam kata tersebut dapat diperoleh informasi tentang staf hotel sangat membantu, sopan, perhatian, terlebih kepada staf yang sudah senior lebih profesional dan bisa berbahsa Inggris sehingga mempermudah pengunjung asing yang datang.

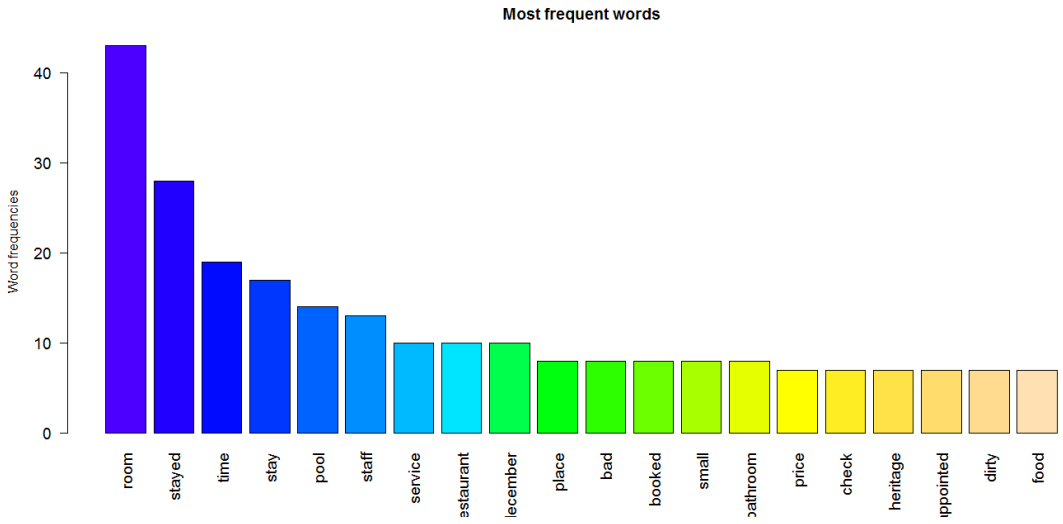
Kata-kata yang berasosiasi dengan kata "sarapan" memberikan informasi tentang sarapan yang disediakan mulai dari prasamanan, memiliki banyak pilihan, terdapat jamu, *booking* sarapan diusahakan dan dengan tepat sehinggan pengunjung merasa dicintai.

Kata-kata yang berasosiasi dengan kata "makanan" memberikan informasi tentang tersedianya makanan ringan, makanan lokal, selain makanan juga disediakan minuman dan terdengar juga musik gamelan javanish sambil makan serta ide memasak yang baik.

5.7.2 Ulasan Negatif

Ekstraksi informasi pada ulasan negatif dilakukan secara berulang-ulang hingga mendapatkan informasi tentang ulasan negatif pengunjung The Phoenix Hotel yang paling sering di ulas/dibicirakan. Berdasarkan hasil pelabelan, ulasan negatif pengunjung terhadap hotel cukup sedikit bila dibandingkan dengan jumlah ulasan positif. Dari total ulasan berbahasa Inggris sebanyak 1.943 ulasan, hanya teridentifikasi sebanyak 87 ulasan negatif. Demikian halnya pada ulasan berbahasa Indonesia, dari total ulasan sebanyak 1.233 ulasan, hanya teridentifikasi sebanyak 231 ulasan negatif. Hal tersebut menunjukkan bahwa mayoritas pengunjung hotel mempunyai persepsi yang baik terhadap hotel. Hasil ekstraksi informasi berupa ulasan negatif diidentifikasi berdasarkan frekuensi kata dalam ulasan, selain itu juga didasarkan pada relevansi kata dengan topik yang mengacu

pada sentimen negatif. Berikut adalah visualisasi hasil ekstraksi informasi yang didapatkan dari ulasan pengunjung dengan klasifikasi ulasan negatif.



Gambar 5.26 Kata yang paling banyak muncul dari kelas negatif *berbahasa Inggris*

Pada hasil klasifikasi ulasan negatif berbahasa Inggris, diperoleh beberapa kata yang paling banyak muncul dengan topik yang dianggap relevan sebagai sentimen negatif diantaranya adalah kata *"room"* dengan frekuensi sebanyak 43 kali, *"stayed"* sebanyak 28 kali, *"time"* sebanyak 19 kali, *"stay"* sebanyak 17 kali, *"pool"* sebanyak 14 kali, dan seterusnya. Kata-kata yang muncul seperti pada **Gambar 5.26** merupakan kata yang memiliki sentimen negatif berbahasa Inggris dan merupakan topik pembicaraan yang paling banyak di ulas oleh pengunjung. Kata-kata tersebut selanjutnya digunakan sebagai dasar untuk menemukan asosiasi dengan kata lainnya, sehingga dapat diperoleh informasi berupa sentimen negatif yang lebih akurat. Kumpulan kata-kata yang sering muncul tersebut dapat ditampilkan dalam bentuk *wordcloud* seperti terlihat pada **Gambar 5.27.**

(a) **(b)**

Gambar 5.27 *Wordcloud ulasan negatif berbahasa Inggris*

Visualisasi *wordcloud* pada **Gambar 5.27** memberikan gambaran yang lebih jelas tentang topik dan kata-kata negatif yang sering digunakan pengunjung dalam memberikan ulasan. Beberapa topik yang sering dibahas pengunjung diantaranya adalah tentang *"room", '"stayed", "time", "stay", "pool", "staff"* dan sebagainya. *Wordcloud* pada **Gambar 5.27 (a)** merupakan *wordcloud* pada ulasan negatif yang didalamnya mengandung kata *"room",* sedangkan **Gambar 5.25 (b)** merupakan *wordcloud* pada ulasan negatif tanpa menggunakan kata *"room".* Selanjutnya, dilakukan pencarian asosiasi antar kata yang sering muncul secara bersamaan dan diperoleh hasil sebagai berikut:

Tabel 5.15 *Asosiasi kata pada klasifikasi negatif berbahasa Inggris*

room		**stayed**		**Time**		**stay**	
deluxe	0,36	cheap	0,45	began	0,44	full	0,38
bathroom	0,35	worst	0,45	feels	0,44	spend	0,38
complained	0,34	days	0,35	pick	0,44	reviews	0,37
forget	0,34	bed	0,33	visit	0,39	thought	0,31
comparable	0,34	hostels	0,31			disappoint	0,31
leak	0,34						
renovation	0,34						

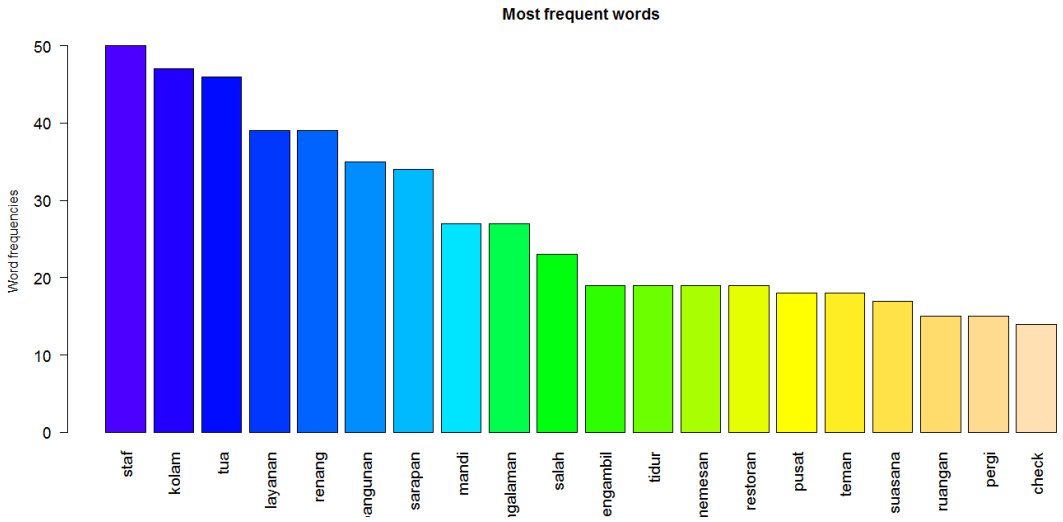
Tabel 5.15 menunjukkan asosiasi antar kata pada ulasan negatif berbahasa Inggris, kata-kata tersebut merupakan topik yang paling sering dibicarakan pengunjung dalam ulasannya. Berdasarkan tabel tersebut dapat diperoleh beberapa informasi berikut.

Kata-kata yang berasosiasi dengan kata *"room"* pada ulasan negatif memberikan informasi pelanggan yang mengeluh terhadap ruang mewah yang kamar mandinya tidak sebanding karena bocor sehingga memerlukan renovasi.

Kata-kata yang berasosiasi dengan kata *"stayed"* pada ulasan negatif memberikan informasi tentang keluhan pelanggan terhadap tempat tinggal yang tidak murah selama berhari-hari dan tempat tidur seperti hostel.

Kata-kata yang berasosiasi dengan kata *"time"* pada ulasan negatif memberikan informasi tentang keluhan pelanggan mulai dari merasa penjemputan terhadap pengunjung yang dianggap masi kurang.

Kata-kata yang berasosiasi dengan kata *"stay"* pada ulasan negatif memberikan informasi tentang keluhan pelanggan terhadap tempat tinggal yang penuh dan habis, ulasan tempat tinggal yang dipikir mengecewakan.



Gambar 5.28 Kata yang paling banyak muncul dari kelas negatif *berbahasa Indonesia*

Minimnya ulasan negatif yang diperoleh dari hasil klasifikasi mempersulit proses ekstraksi informasi. Pengunjung tidak secara spontan memberikan ulasan

negatif terhadap hotel, melainkan kata-kata negatif berupa keluhan, kekecewaan dan ketidakpuasan diungkapkan pengunjung setelah didahului oleh kalimat berupa ulasan positif, sangat jarang ditemukan ulasan yang secara spontan mengulas tentang hal-hal negatif pada hotel. Oleh karena itu, proses ekstraksi informasi berupa ulasan negatif dilakukan dengan cara filtrasi secara berulang-ulang dan menghilangkan beberapa kata yang diidentifikasi sebagai sentimen positif.

Gambar 5.28 menunjukkan hasil klasifikasi ulasan negatif berbahasa Indonesia. Berdasarkan gambar tersebut diperoleh beberapa kata yang paling banyak muncul dengan topik yang dianggap relevan sebagai sentimen negatif diantaranya adalah kata "staf" dengan frekuensi sebanyak 50 kali, "kolam" sebanyak 47 kali, "tua" 46 kali, "layanan" 39 kali, "renang" sebanyak 39 kali dan seterusnya. Kumpulan kata-kata yang sering muncul tersebut kemudian ditampilkan dalam bentuk *wordcloud* seperti terlihat pada **Gambar 5.29.**

Gambar 5.29 *Wordcloud ulasan negatif berbahasa Indonesia*

Visualisasi *wordcloud* pada **Gambar 5.29** memberikan gambaran yang lebih jelas tentang topik dan kata-kata negatif berbahasa Indonesia yang sering digunakan pengunjung dalam memberikan ulasan. Beberapa topik yang sering di bahas pengunjung diantaranya adalah tentang jalan, staf, kolam, tua, dan

sebagainya. Selanjutnya, dilakukan pencarian asosiasi dan diperoleh hasil sebagai berikut:

Tabel 5.16 *Asosiasi kata pada klasifikasi negatif berbahasa Indonesia*

staf		**kolam**		**tua**		**layanan**	
hambatan	0,26	menghadap	0,28	noda	0,43	digolongkan	0,30
pakaian	0,26	balkon	0,26	kecuali	0,37	dikecewakan	0,30
bebas	0,26	membuka	0,25	tanda	0,37	menentu	0,30
keamanan	0,26	menutup	0,25	asli	0,34	menjengkelkan	0,30
negatif	0,23	lantai	0,25	marmer	0,34		
perhatian	0,20	akses	0,25				
kekurangan	0,16	terjauh	0,25				
sisi	0,16						
tingkat	0,16						
perawatan	0,16						

Berdasarkan hasil ekstraksi informasi, pada data ulasan negatif berbahasa Indonesia hanya diperoleh sedikit informasi tentang keluhan pengunjung terhadap hotel. **Tabel 5.16** merupakan hasil ekstraksi informasi ulasan negatif berbahasa Indonesia, berdasarkan tabel tersebut diperoleh beberapa informasi keluhan tentang jalan, staf dan kolam hotel.

Kata-kata yang berasosiasi dengan kata “staf” memberikan informasi tentang hambatan pakaian yang digunakan oleh staf, keamanan yang bebas, perhatian yang negatif, dan kekurangan staf dalam sisi tingkat perawatan.

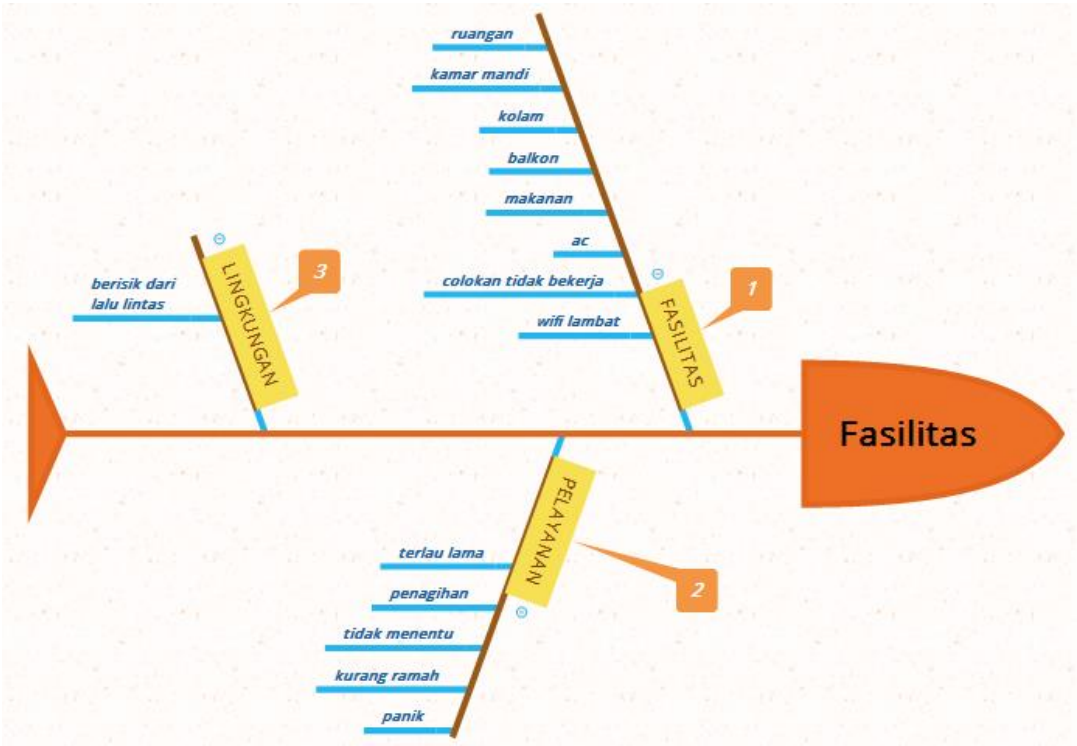
Kata-kata yang berasosiasi dengan kata “kolam” memberikan informasi tentang kolam yang menghadap balkon yang terbuka-tutup dan lantai akses kolam yang jauh.

Kata-kata yang berasosiasi dengan kata “tua” memberikan informasi tentang terdapat noda kecuali tanda asli dari lantai marmer.

Kata-kata yang berasosiasi dengan kata “layanan” memberikan informasi tentang layanan yang digolongkan dikecewakan tentu menjengkelkan.

5.8 Diagram Sebab-Akibat (*Fishbone Diagram*)

Berdasarkan hasil ulasan negatif yang didapatkan dari data, maka dapat diperoleh beberapa masalah yang terjadi terkait dengan ulasan negatif berbahasa Inggris dan Indonesia pada The Phoenix Hotel Yogyakarta berdasarkan **Gambar 5.29**.



Gambar 5.30 *Diagram sebab-akibat ulasan negatif*

Gambar 5.30 dapat diketahui bahwa faktor-faktor yang menyebabkan hotel memiliki ulasan negatif yaitu dari segi fasilitas, lingkungan, pelayanan, dan manusia. Berikut adalah rincian permasalahan dari keempat faktor tersebut:

1. Fasilitas
   a. Ruangan

      Ruangan yang digunakan oleh pelanggan dirasakan kecil/sempit, berjamur, kotor, dan berbau asap rokok.

   b. Kamar mandi

      Fasilitas yang ada di dalam kamar mandi berdasarkan hasil ulasan yaitu kotor, kecil, handuk yang tua, ada ada kerusakan kecil seperti kebocoran.

c. Kolam

Kolam yang disediakan oleh hotel berdasarkan hasil ulasan yaitu bising, kecil dan tidak memiliki kondisi yang baik untuk berjemur.

d. Balkon

Balkon yang disediakan oleh hotel berdasarkan hasil ulasan yaitu ketika hujan turun balkon basah dan tampilan atau *view* nya kurang sesuai.

e. Makanan

Makanan yang disediakan oleh hotel berdasarkan hasil ulasan yaitu mahal, bir mahal, dan ayam yang terdapat tulang.

Selain dari sisi ruangan, kamar mandi dan makan, ulasan yang diberikan oleh pelanggan yang berkaitan dengan fasilitas adalah AC yang berisik, suara kulkas yang terdengar dan balkon yang menghadap dengan *view* yang dirasakan kurang menarik seperti menghadap ke lokasi yang kurang sesuai seperti yang diinginkan.

2. Pelayanan

Faktor pelayanan berdasarkan hasil ulasan didapatkan informasi yaitu lambat dalam hal penagihan, tidak menentu, dan balasan email yang lambat, staf kurang ramah, kurang sopan, dan tidak terlatih (berbicara bahasa Inggris).

3. Lingkungan

Pada faktor lingkungan berdasarkan hasil ulasan didapatkan informasi bahwa pelanggan terganggu dengan kebisingan karena kamar yang didapatkan dekat dengan jalan.

Setelah diketahui faktor penyebab yang paling dominan terhadap permasalahan, langkah selanjutnya yaitu menentukan rencana penanggulanggan untuk memecahkan permasalahan yang ada. Adapun rencana penaggulangan untuk memecahkan permasalahan di The Phoenix Hotel Yogyakarta berdasarakan ulasana berbahasa Inggris dapat dilihat pada **Tabel 5.17** berikut.

Tabel 5.17 *Rencana Penanggulangan Permasalahan*

Faktor yang diamati	Masalah yang terjadi	Rencana penanggulangan
Fasilitas	a. Ruangan yang digunakan oleh pelanggan dirasakan kecil/sempit, berjamur, kotor, dan berbau asap rokok. b. Fasilitas yang ada di dalam kamar mandi berdasarkan hasil ulasan yaitu kotor, kecil, handuk yang tua, ada ada kerusakan kecil seperti kebocoran. c. Makanan yang disediakan oleh hotel berdasarkan hasil ulasan yaitu mahal, bir mahal, dan ayam yang terdapat tulang.	a. Aksesoris ruangan yang memakan ruangan sebaiknya dihindari atau gunakan yang ukuran sesuai dengan ruangan, dan diusahakan terdapat cahaya sinar matahari sehingga ruang tidak berjamur karena lembap yang mengakibatkan kelihatan kotor dan kamar diberi pengharum ruangan agar tidak berbau. b. Kamar mandi dibersihkan lagi setelah dipakai oleh pelanggan sebelumnya, gunakan peralatan atau aksesori ukuran sesuai dengan ruangan, handuk dan kerusakan lainnya sebaiknya diganti lagi. c. Makanan dan bir sebaiknya diberikan *range* harga yang sesuai dengan *budget* pelanggan sehingga pelanggan lainnya bisa membeli sesuai dengan *budget* yang mereka miliki. Untuk ayam yang memiliki tulang sebaiknya dibuatkan dalam bentuk sate.

Faktor yang diamati	Masalah yang terjadi	Rencana penanggulangan
Pelayanan	d. Lambat dalam hal penagihan, tidak menentu, dan balasan email yang lambat, staf kurang ramah, kurang sopan, dan tidak terlatih (berbicara bahasa Inggris).	d. Staf sebaiknya ditambah agar hasil penagihan, tingkat keakuratan dalam penagihan bisa lebih cepat dan bekerja untuk melayani kebutuhan pelanggan lainnya lebih cepat. Untuk staf sebaiknya dilatih agar lebih ramah kepada pelanggan, tidak berteriak di depan pelanggan dan bahasa Inggrisnya lebih di tingkatkan lagi.
Lingkungan	e. Pelanggan terganggu dengan kebisingan karena kamar yang didapatkan dekat dengan jalan.	e. Sebaiknya dinding kamar diberi kedap suara agar pelanggan bisa beristirahat dengan baik.

BAB VI
PENUTUP

6.1 Kesimpulan

Berdasarkan hasil analisis dam pembahasan pada bab sebelumnya, penulis dapat menarik beberapa kesimpulan sebagai berikut:

1. Teknik *web scraping* dapat digunakan sebagai cara alternatif untuk mendapatkan data dari halaman *website. Web scraping* mempermudah dan mempercepat proses pengambilan data dalam skala besar secara otomatis di internet.
2. Jumlah ulasan yang masuk ke situs *TripAdvisor* berdasarkan urutan tahun sejak tahun 2011 hingga 2017 cenderung mengalami fluktuasi, namun secara umum, baik ulasan berbahasa Inggris maupun ulasan berbahasa Indonesia jumlah ulasan selalu mengalami peningkatan secara drastis pada tahun 2013 dan 2014.
3. Berdasarkan *rating*, mayoritas pengunjung The Phoenix Hotel Yogyakarta mempunyai penilaian yang baik terhadap hotel. Pada *rating* penilaian berbahasa Inggris, dari 1.943 ulasan, terdapat sebanyak 1.154 pengunjung memberikan penilaian *excellent* (luar biasa), 577 pengunjung memberikan penilaian *very good* (sangat bagus), 165 pengunjung memberikan penilaian *average* (rata-rata), sedangkan penilaian dengan kategori *poor* (buruk) hanya berjumlah 27 dan 20 ulasan dengan kategori *terrible* (sangat buruk). Sedangkan untuk rating dalam ulasan berbahasa Indonesia, dari 1.233 ulasan, terdapat sebanyak 704 pengunjung memberikan penilaian luar biasa, 386 pengunjung memberikan penilaian sangat bagus, 144 pengunjung memberikan penilaian rata-rata, sedangkan penilaian dengan kategori buruk hanya berjumlah 16 dan 13 ulasan dengan kategori sangat buruk.
4. Pada ulasan berbahasa Inggris, hasil klasifikasi sentimen menggunakan metode *Support Vector Machine* menghasilkan tingkat akurasi sebesar 96,07%, artinya dari 331 data ulasan yang diuji, terdapat 331 ulasan yang

benar pengklasifikasiannya oleh model *SVM*. Sedangkan pada ulasan berbahasa Indonesia, hasil klasifikasi sentimen menggunakan metode *Support Vector Machine* menghasilkan tingkat akurasi sebesar 84,77%, artinya dari 302 data ulasan yang diuji, terdapat 296 ulasan yang benar pengklasifikasiannya oleh model *SVM*.

5. Klasifikasi pada data ulasan menghasilkan nilai *recall negative*, yakni sebesar 100% untuk ulasan berbahasa Inggris, dan 12% untuk ulasan berbahasa Indonesia.
6. Secara umum, pengunjung The Phoenix Hotel Yogyakarta mayoritas memiliki persepsi yang baik terhadap hotel. Pengunjung banyak memberikan penilaian positif diantaranya tentang penilaian kamar hotel yang mewah dan terdapat balkon, staf yang ramah dan membantu, menu sarapan yang disajikan secara prasmanan yang banyak variasi makanan yang dinilai sempurna dan pemilihan sebaran atau tatanan makanan yang tepat. Sedangkan beberapa penilaian negatif pengunjung daiantaranya adalah layanan yang digolongkan dikecewakan tentu menjengkelkan dan kekurangan staf dalam sisi tingkat perawatan.
7. Berdasarkan diagram *fishbone*, urutan faktor yang harus ditingkatkan dalam pemecahan masalah hasil dari ulasan negatif adalah faktor fasilitas, pelayanan dan lingkungan.

6.2 Saran

Berdasarkan hasil analisis dan kesimpulan, dapat diberikan saran sebagai berikut:

6.2.1 Untuk Peneliti Selanjutnya

1. Dalam mendapatkan data *scraping* sebaiknya memiliki koneksi internet yang cepat sehingga data yang didapatkan semakin banyak.
2. Sistem pelabelan kelas sentimen yang digunakan dalam penelitian ini hanya sebatas pada pendeteksian sentimen antar kata menggunakan kamus *lexicon*, sehingga kata-kata negasi belum dapat teridentifikasi dengan baik, untuk penelitian selanjutnya sebaiknya dapat menggunakan

sistem pelabelan yang memiliki tingkatan lebih tinggi, yakni mampu mendeteksi sentimen pada frasa dan kalimat.

3. Dataset yang digunakan dalam penelitian ini memiliki perbandingan kelas yang tidak seimbang, sehingga hasil yang diperoleh kurang maksimal, untuk penelitian selanjutnya disarankan untuk menggunakan dataset yang memiliki perbandingan kelas yang seimbang.
4. Jika ditemukan kasus ketidakseimbangan data (*imbalanced dataset*), sebaiknya dilakukan penanganan khusus dengan menggunakan metode atau cara tertentu agar dapat menghasilkan hasil klasifikasi yang optimal.

6.2.2 Untuk Pihak Hotel

1. Secara keseluruhan ulasan negatif hasil dari pelabelan berbahasa Inggris maupun berbahasa Indonesia sebaiknya pihak hotel lebih memperhatikan dan meningkatkan lagi faktor fasilitas terutama di bagian sub dalam ruangan kamar maupun luar kamar, kamar mandi, kolam, balkon, dan makanan dengan mengacu saran peneliti disajikan dalam tabel Rencana Penanggulangan Permasalahan Ulasan.

DAFTAR PUSTAKA

Abtohi, Slamet. 2017. *Implementasi Teknik Web Scraping dan Klasifikasi Sentimen Menggunakan Metode Support Vector Machine dan Asosiasi*. Tugas Akhir. Universitas Islam Indonesia. Yogyakarta: Fakultas Matematika dan Ilmu Pengetahuan Alam.

Agung, I. N. 2000. *Analisis Statistik Sederhana untuk Pengambilan Keputusan*. Jurnal UGM 11(2), ISSN: 0853-0262.

Antarakalbar. 2016. *TripAdvisor Rayakan 15 Tahun Jadi Situs Rujukan Wisatawan*. http://www.antarakalbar.com. Akses, 5 Januari 2018.

Ardiyanto, F. 2017. *Hotel Phoenix*. https://kebudayaan.kemdikbud.go.id/bpcbyogyakarta. Akses, 30 Desember 2017.

Bouge, Kevin. 2011. *Download Stop Words.* https://sites.google.com/site/kevinbouge/stopwords-lists. Akses, 31 Januari 2018.

Cantino, Andrew. 2013. *SelectorGadget: Point and Click CSS Selectors.* http://selectorgadget.com. Akses 20 November 2017.

Clayton R, Fink. 2011. *Coarse-and Fine-Grained Sentiment Analysis of Social Media Text*. Johns hopkins apl technical digest,volume 30, No 1.

Davies, dan Paul Beynon. 2004. *Database Systems Third Edition,* Palgrave Macmillan, New York.

detiktravel. 2013. *10 Situs Booking Travel Paling Populer di Dunia*.https://m.detik.com. Akses, 5 januari 2018.

Dewantoro, P. 2016. *Implementasi Teknik Web Scraping pada Proses Topic Modelling Porta Berita*. Skripsi. Universitas Gajah Mada. Yogyakarta: Fakultas Pendidikan dan Ilmu Pengetahuan Alam.

Diaz, Gene. 2016. *Indonesian Stopwords Collection.* https://github.com/stopwords-iso/stopwords-id. Akses, 31 Januari 2018.

Elango, V., dan Narayanan, G. 2014. *Sentiment Analysis for Hotel Reviews*. http://stanford.edu.

Ericson. 2017. *10 Negara Pengguna Internet Terbesar di Dunia.* http://www.ilmupengetahuan.com. Akses, 10 Januari 2018.

Evrard, Thomas. 2016. *Sensasi Menginap di Hotel Phoenix Yogyakarta*. http://id.beritasatu.com. Akses, 14 Januari 2018.

Fauziah, Naily. 2009. *Aplikasi Fishbone Analysis dalam Meningkatkan Kualitas Produksi Teh Pada PT Rumpun Sari Kemuning, Kabupaten Karanganyar*. Tugas Akhir. Universitas Sebelas Maret. Surakarta: Fakultas Pertanian.

Fauzie, R. 2010. *Pengenalan Citra Wajah Menggunakan Algoritma VFI5 dengan Praproses Principal Component Analysis.* Bogor. Institut Pertanian Bogor.

Fawcett, T. 2006. *An introduction to ROC analysis.* Pattern Recognition Letters 27.8, pp. 861–874.

Hadi, D.A. 2016. *Ebook Belajar HTML & CSS Dasar*. http://malasngoding.com. Akses, 8 Desember 2017.

Han, J. dan Kamber, M. 2006. *Data Mining Concepts and Techniques Second Edition*. Morgan Kauffman, San Francisco.

Hasan, Iqbal. 2004. *Analisa Data Penelitian dengan Statistik*. Jakarta: PT Bumi Aksara.

Hasan, Iqbal. 2001. *Pokok-Pokok Materi Statistik 1 (Statistik Deskriptif)*. Jakarta: PT BumiAksara.

Hilwah, N., Kudus, A., & Sunendiari, S. 2017. *Klasifikasi Text Mining untuk Terjemahan Ayat-ayat Al-Qur'an menggunakan Metode Klasifikasi Naive Bayes.* Prosiding Statistika. Bandung: Universitas Islam Bandung.

Ilmar, Maulana. 2008. *Analisis dan Perancangan Sistem Informasi Berbasis Website pada SMA Negeri 1 Pemalang*. Skripsi Sekolah Tinggi Manajemen Informatika dan Komputer Yogyakarta.

Imamoto, T. et al. 2008. Perivesical abscess caused by migration of a fish bone from the intestinal tract. *International Journal of Urology*. Vol. 9 (405-409).

Ismarani, Dian. 2017. *Data Penggunaan Internet Tahun 2017 dan Apa Kesimpulan Yang Bisa Diambil Dari Data Tersebut*. http://www.youthmanual.com. Akses, 10 Januari 2018.

Indrayuni, Elly. 2016. *Analisis Sentimen Review Hotel Menggunakan Algoritma Support Vector Machine Berbasis Particle Swarm Optimization*. Journal Evolusi Volume 4 Nomor 2-2016.

Josi, A., Abdillah, L.A., Suryayusra. 2014. *Penerapan Teknik Web Scraping Pada Mesin Pencari Artikel Ilmiah.* Jurnal Sistem Informasi, Volume 5, Nomor 2, September 2014, hal. 159-164.

Kaestner, Celso.2013. *Support Vector Machines and Kernel Functions for Text Processing.* RITA Volume 20 Number 3 2013.

Kurniawan, Bambang, et.al. 2012. *Klasifikasi Konten Berita dengan Metode Text Mining*. Jurnal Dunia Teknologi Informasi vol. 1, (2012) 14-19.

Mandala, Rila dan Setiawan, Hendra. 2002. *Peningkatan Performansi Sistem Temu-Kembali Informasi dengan Perluasan Query Secara Otomatis*, Bandung : Departemen Teknik Informatika Institut Teknologi Bandung.

Megawati, Chyntia. 2015. *Analisis Aspirasi Dan Pengaduan Di Situs Lapor! Dengan Menggunakan Text Mining.* Skripsi. Program Studi Teknik Industri Fakultas Teknik Universitas Indonesia Depok.

Miner, G., et al. (2012). *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications.* Oxford: Elsevier.

Nugeraha, F., A. 2015. *Proses Klasifikasi Teks Pornografi Berbahasa Indonesia Berbasis Machine Learning.* Skripsi. Universitas Gajah Mada. Yogyakarta: Program Studi Teknologi Informasi, Fakultas Teknik.

Nur, M.,Y. dan Santika, D. 2011. *Analisis Sentimen pada Dokumen Berbahasa Indonesia dengan Pendekatan Support Vector Machine*. Jurnal Ilmiah pada Konferensi Nasional Sistem dan Informatika. Bali November 12, 2011 KNS&III-002.

Pramudiono, I. 2007. *Pengantar Data Mining: Menambang Permata Pengetahuan di Gunung Data.* http://www.ilmukomputer.org. Akses 8 Desember 2017.

Prasetyo, Eko. 2012. *Data Mining: Konsep dan Aplikasi Menggunakan MATLAB.* Yogyakarta: Andi Offset.

Prasetyo, Himawan. 2018. *Kebudayaan Yogyakarta.*http://kebudayaan.kemendikbud.go.id. Akses, 15 Januari 2018.

Pratiwi. 2010. *Pengembangan Model Pengenalan Wajah Dengan Jarak Euclid Pada Ruang Eigen Dengan 2DPCA.* Bogor. Program Pascasarjana, Institut Pertanian Bogor.

Rianto, Bagus. 2016. *Implementasi dan Perbandingan Metode Prapemprosesan pada Analisis Sentimen Gubernur DKI Jakarta Menggunakan Metode Support Vector Machine dan Naïve Bayes.* Skripsi. Universitas Gajah Mada Yogyakarta: Program Studi Ilmu Komputer FMIPA.

Riska, S. Y., Cahyani, L., & Rosadi, M. I. (2015). Klasifikasi Jenis Tanaman Mangga Gadung dan Mangga Madu Berdasarkan Tulang Daun. *Jurnal Buana Informatika*, 41-50.

Rusyanto, Edo. 2018. *Sensasi Menginap di Hotel Phoenix Yogyakarta*. http://id.beritasatu.com. Akses, 14 Januari 2018.

…… 2017. *Sistem Informasi Statistik Dinas Pariwisata Daerah Istimewa Yogyakarta*. http://statistikhotel.visitingjogja.com. Akses, 15 Januari 2018.

Santosa, Budi. 2007. *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis.* Yogyakarta: Graha Ilmu.

Sugiyono. 2009. *Metode Penelitian Kuantitatif, Kualitatif dan R&D*, Bandung: Alfabeta.

Suhartono, Bambang. 2014. *Fungsi dan Manfaat Internet dalam Bidang Bisnis dan Perdangan*. s://bambangsuhartono. wordpress.com. Akses, 10 Januari 2018.

Sulastiyono, Agus. 2011. *Manajemen Penyelenggara Hotel*. Alfabeta. Bandung.

Susanti, A.R.2016. *Analisis Klasifikasi Sentimen Twitter Terhadap Kinerja Layanan Provider Telekomunikasi Menggunakan Varian Naïve Bayes.* Tesis. Institut Pertanian Bogor.

Susilowati, E., et.al. 2015. *Implementasi MetodeSupport Vector Machine untuk Melakukan Klasifikasi Kemacetan Lalu Lintas pada Twitter*. e-proceeding of Enginerering: Vol. 2, No 1 April 2015| Page 1478.

Tanaka, M.; Okutomi, M. A novel inference of a restricted boltzmann machine. Pattern Recognition (ICPR), 2014 22nd International Conference on. 2014; pp 1526–1531.

Tribunjogja. 2016. *Kunjungan Wisman dan Wisnus ke DIY Setiap Tahunnya Terus Meningkat*. http://jogja.tribunnews. com. Akses, 10 Desember 2017.

Turland, M. 2010. *Pph | Architect's Guide to Web Scraping with PHP*. Toronto. Canada: Marco Tabini & Associates, Inc.

Ulwan, N.,M. 2016. *Pattern Recognition pada Unstructured Data Teks Menggunakan Support Vector Machine dan Association*. Tugas Akhir. Universitas Gajah Mada. Yogyakarta: Program Studi Statistik FMIPA.

Wati, N., T., P. 2015. *Analisis Pengaruh Online Review dalam TripAdvisor Terhadap Keputusan Menginap Wisatawan di Hotel Centra Taun Seminyak Badung Bali*. Laporan Akhir. Program Studi Diploma IV Pariwisata. Fakultas Pariwisata. Bali: Universitas Udayana.

Wickham, H., dan R.Studio. 2016. *Easily Harvest (Scrape) Web Pages*. https://cran.r-project.org/web/packages/rvest/index.html. Akses tanggal 20 November 2017.

Wieringa, J. E. 2016. *Unstructured data Can Its Power Be Unleashed?*. Faculty of Economics and Business University of Groningen.

Witten, I. H dan Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques Second Edition*. Morgan Kauffman: San Francisco.

Yunus. 2014. *Pengaruh Kualitas Pelayanan dan Fasilitas Terhadap Kepuasan Pelanggan*. Jurnal Ilmu & Riset Manajemen, Vol. 3 No. 12.

Zafikri, Atika. 2008. *Implementasi Metode Term Frequency Inverse Document Frequency (TF-IDF) Pada Sistem Temu Kembali Informasi*. Skripsi. Program Studi S-1 Ilmu Komputer FMIPA USU.

LAMPIRAN

Lampiran 1 *Script R Web Scraping*

```
#Install.packages('rvest')
library(rvest)
url<-read_html("https://www.tripadvisor.com/Hotel_Review-g294230-d446805-
Reviews-The_Phoenix_Hotel_Yogyakarta_MGallery_Collection-
Yogyakarta_Java.html")

#Find the lnumber of the last page listed in the bottom of the main page
npages<-url%>%
  html_nodes(" .pageNum")%>%
  html_attr(name="data-page-number")%>%
  tail(.,1)%>%
  as.numeric()

#Find index page
a<-0:(npages-1)
b<-5
res<-numeric(length=length(a))
for (i in seq_along(a)) {
  res[i]<-a[i]*b
}
tableout <- data.frame()
for(i in res){
  cat(".")

#Change URL address here depending on attraction for review
  url <- paste ("https://www.tripadvisor.com/Hotel_Review-g294230-
d446805-Reviews-or",i,"-The_Phoenix_Hotel_Yogyakarta_MGallery_Collection-
Yogyakarta_Java.html#REVIEWS",sep="")
  reviews <- url %>%
    html() %>%
    html_nodes("#REVIEWS .innerBubble")
  id <- reviews %>%
    html_node(".quote a") %>%
    html_attr("id")
  quote <- reviews %>%
    html_node(".quote span") %>%
    html_text()
  rating <- reviews %>%
    html_node(".rating .ui_bubble_rating") %>%
    html_attrs() %>%
    gsub("ui_bubble_rating bubble_", "", .) %>%
    as.integer() / 10
  date <- reviews %>%
    html_node(".innerBubble, .ratingDate") %>%
    html_text()
  review <- reviews %>%
    html_node(".entry .partial_entry") %>%
    html_text()
```

```
  #Get rid of \n in reviews as this stands for 'enter' and is confusing
dataframe layout
  reviewnospace <- gsub("\n", "", review)
  temp.tableout <- data.frame(id, quote, rating, date, reviewnospace)
  tableout <- rbind(tableout,temp.tableout)
}

#Change output file name depending on attraction for review
write.csv(tableout, "E:/Bismillah TA/Data Script/dataulasan.csv")
```

Lampiran 2 *Script R Preprocessing Data dengan Text Mining*

```
#Install
install.packages("tm")           #for text mining
install.packages("SnowballC")    #for text stemming
install.packages("wordcloud")    #word-cloud generator
install.packages("RColorBrewer") #color palettes

#Load
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
library(stringr)
setwd("E:/Bismillah TA/Data Script/")
docs<-readLines("dataulasan.csv")

#Load the data as a corpus
docs <- Corpus(VectorSource(docs))

#Inspect the content of the document
inspect(docs)

#Replacing "/", "@" and "|" with space:
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ",
x))
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")

#Cleaning the text
# Convert the text to lower case
docs <- tm_map(docs, content_transformer(tolower))

#Remove punctuation
docs <- tm_map(docs, toSpace, "[[:punct:]]")

#Remove numbers
docs <- tm_map(docs, toSpace, "[[:digit:]]")

#Add two extra stop words: "available" and "via"
myStopwords = readLines("stopword_en.csv")

#Remove stopwords from corpus
docs <- tm_map(docs, removeWords, myStopwords)

#Remove your own stop word
#Specify your stopwords as a character vector
docs <- tm_map(docs, removeWords,
c("but","you","also","hotel","Phoenix","phoenix","absolutely","yogyakarta
","jogja"))

#Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)
```

```
#Remove URL
removeURL <- function(x) gsub("http[[:alnum:]]*", " ", x)
docs <- tm_map(docs, removeURL)

#Replace words
docs <- tm_map(docs, gsub, pattern="helful", replacement="helpful")
docs <- tm_map(docs, gsub, pattern="colonnial", replacement="colonial")
docs <- tm_map(docs, gsub, pattern="helpfull", replacement="helpful")
docs <- tm_map(docs, gsub, pattern="Borodubur", replacement="Borobudur")
docs <- tm_map(docs, gsub, pattern="impecable", replacement="impeccable")
docs <- tm_map(docs, gsub, pattern="builiding", replacement="building")
docs <- tm_map(docs, gsub, pattern="Phonix", replacement="Phoenix")
docs <- tm_map(docs, gsub, pattern="Pheonix", replacement="Phoenix")
docs <- tm_map(docs, gsub, pattern="Maliboro", replacement="Malioboro")
docs <- tm_map(docs, gsub, pattern="boropudur", replacement="Borobudur")
docs <- tm_map(docs, gsub, pattern="comort", replacement="comfort")
docs <- tm_map(docs, gsub, pattern="Excutive", replacement="Executive")
docs <- tm_map(docs, gsub, pattern="equiped", replacement="equipped")
docs <- tm_map(docs, gsub, pattern="Borodudur", replacement="Borobudur")
docs <- tm_map(docs, gsub, pattern="matresses", replacement="mattresses")
docs <- tm_map(docs, gsub, pattern="kolonial", replacement="colonial")
docs <- tm_map(docs, gsub, pattern="wonderfull", replacement="wonderful")
docs <- tm_map(docs, gsub, pattern="couteous", replacement="courteous")
docs <- tm_map(docs, gsub, pattern="sumptous", replacement="sumptuous")
docs <- tm_map(docs, gsub, pattern="personel", replacement="personnel")
docs <- tm_map(docs, gsub, pattern="excelent", replacement="excellent")
docs <- tm_map(docs, gsub, pattern="atmospher", replacement="atmosphere")
docs <- tm_map(docs, gsub, pattern="smal", replacement="small")

#Build a term-document matrix
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 50)
dataframe<-data.frame(text=unlist(sapply(docs, `[`)), stringsAsFactors=F)
write.csv(dataframe, "E:/Bismillah TA/Data Script/datacleaning.csv")
```

Lampiran 3 *Script R Pelabelan Kelas Sentimen*

```
library(tm)
setwd("E:/Bismillah TA/Data Script/")
kalimat2<-read.csv("datacleaning.csv",header=TRUE)

#Scoring
positif <- scan("positive-words.txt",what="character",comment.char=";")
negatif <- scan("negative-words.txt",what="character",comment.char=";")
kata.positif = c(positif, "is near to")
kata.negatif = c(negatif, "cant")
score.sentiment = function(kalimat2, kata.positif, kata.negatif,
.progress='none')
{
  require(plyr)
  require(stringr)
  scores = laply(kalimat2, function(kalimat, kata.positif, kata.negatif)
{
    kalimat = gsub('[[:punct:]]', '', kalimat)
    kalimat = gsub('[[:cntrl:]]', '', kalimat)
    kalimat = gsub('\\d+', '', kalimat)
    kalimat = tolower(kalimat)
    list.kata = str_split(kalimat, '\\s+')
    kata2 = unlist(list.kata)
    positif.matches = match(kata2, kata.positif)
    negatif.matches = match(kata2, kata.negatif)
    positif.matches = !is.na(positif.matches)
    negatif.matches = !is.na(negatif.matches)
    score = sum(positif.matches) - (sum(negatif.matches))
    return(score)
  }, kata.positif, kata.negatif, .progress=.progress )
  scores.df = data.frame(score=scores, text=kalimat2)
  return(scores.df)
}
hasil = score.sentiment(kalimat2$text, kata.positif, kata.negatif)
View(hasil)

#CONVERT SCORE TO SENTIMENT
hasil$klasifikasi<- ifelse(hasil$score<0, "Negatif","Positif")
hasil$klasifikasi
View(hasil)

#EXCHANGE ROW SEQUENCE
data <- hasil[c(3,1,2)]
View(data)
write.csv(data, file = "hasil_pelabelan.csv")
```

Lampiran 4 *Script R Klasifikasi dengan Machine Learning menggunakan SVM*

```
setwd("E:/Bismillah TA/Data Script/SVM Sari")
positif = readLines("PL.csv")
negatif = readLines("NL.csv")
sari.tr = c(positif, negatif)

positiftes = readLines("PT.csv")
negatiftes = readLines("NT.csv")
sari.ts = c(positiftes, negatiftes)

sari.all = c(sari.tr,sari.ts)
length(sari.all)
str(sari.all)

#Memberi label sepanjang variabel positif dan negatif
sentiment = c(rep("positif", length(positif) ),
              rep("negatif", length(negatif)))
sentiment_test = c(rep("positif", length(positiftes) ),
                   rep("negatif", length(negatiftes)))

sentiment_all = as.factor(c(sentiment, sentiment_test))
str(sentiment_all)
length(sentiment_all)

library(RTextTools)
library(e1071)

mat = create_matrix(sari.all, language = "english", removeStopwords =
FALSE,
    removeNumbers = TRUE, stemWords = FALSE, tm::weightTfIdf)
mat = as.matrix(mat)

#SVM
container <- create_container(mat, sentiment_all,
trainSize=1:1612,testSize=1613:1943, virgin=FALSE)

model <- train_model(container, 'SVM',kernel='sigmoid')
model
results <- classify_model(container, model)
results
table(as.character(sentiment_all[1613:1943]),
as.character(results[,"SVM_LABEL"]))
recall_accuracy(sentiment_all[1613:1943], results[,"SVM_LABEL"])
```

Lampiran 5 *Script R Visualisasi dan Asosiasi Kata*

```
#Install
install.packages("tm")            #for text mining
install.packages("SnowballC")     #for text stemming
install.packages("wordcloud")     #word-cloud generator
install.packages("RColorBrewer")  #color palettes

#Load
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
library(stringr)

setwd("E:/Bismillah TA/Data Script")
docs<-readLines("positifinggris.csv")

#Load the data as a corpus
docs <- Corpus(VectorSource(docs))

#Remove your own stop word
#specify your stopwords as a character vector
docs <- tm_map(docs, removeWords,
c("hotel","yogyakarta","jogja","yogya"))

#Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)

#Replace words
docs <- tm_map(docs, gsub, pattern="helful", replacement="helpful")
docs <- tm_map(docs, gsub, pattern="colonnial", replacement="colonial")
docs <- tm_map(docs, gsub, pattern="helpfull", replacement="helpful")
docs <- tm_map(docs, gsub, pattern="Borodubur", replacement="Borobudur")
docs <- tm_map(docs, gsub, pattern="impecable", replacement="impeccable")
docs <- tm_map(docs, gsub, pattern="builiding", replacement="building")
docs <- tm_map(docs, gsub, pattern="Phonix", replacement="Phoenix")
docs <- tm_map(docs, gsub, pattern="Pheonix", replacement="Phoenix")
docs <- tm_map(docs, gsub, pattern="Maliboro", replacement="Malioboro")
docs <- tm_map(docs, gsub, pattern="boropudur", replacement="Borobudur")
docs <- tm_map(docs, gsub, pattern="comort", replacement="comfort")
docs <- tm_map(docs, gsub, pattern="Excutive", replacement="Executive")
docs <- tm_map(docs, gsub, pattern="equiped", replacement="equipped")
docs <- tm_map(docs, gsub, pattern="Borodudur", replacement="Borobudur")
docs <- tm_map(docs, gsub, pattern="matresses", replacement="mattresses")
docs <- tm_map(docs, gsub, pattern="kolonial", replacement="colonial")
docs <- tm_map(docs, gsub, pattern="wonderfull", replacement="wonderful")
docs <- tm_map(docs, gsub, pattern="couteous", replacement="courteous")
docs <- tm_map(docs, gsub, pattern="sumptous", replacement="sumptuous")
docs <- tm_map(docs, gsub, pattern="personel", replacement="personnel")
docs <- tm_map(docs, gsub, pattern="excelent", replacement="excellent")
docs <- tm_map(docs, gsub, pattern="atmospher", replacement="atmosphere")
docs <- tm_map(docs, gsub, pattern="smal", replacement="small")
```

```
#Build a term-document matrix
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 25)

#Generate the Word cloud
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
       max.words=50, random.order=FALSE, rot.per=0.35,
       colors=brewer.pal(8, "Dark2"))

#Explore frequent terms and their associations
findFreqTerms(dtm, lowfreq = 4)

#Asosiasi kata
v<-as.list(findAssocs(dtm, terms =c("room","staff","good","stayed",
"breakfast","great","rooms","stay","pool","nice","service","beautiful"),
corlimit =
c(0.15,0.15,0.15,0.15,0.15,0.15,0.15,0.15,0.15,0.15,0.15,0.15)))
v

#Barplot
k<-barplot(d[1:20,]$freq, las = 2, names.arg =
d[1:20,]$word,cex.axis=1.2,cex.names=1.2,
main ="Most frequent words",
ylab = "Word frequencies",col =topo.colors(20))

termFrequency <- rowSums(as.matrix(dtm))
termFrequency <- subset(termFrequency, termFrequency>=5)

text(k,sort(termFrequency, decreasing = T)-
1,labels=sort(termFrequency, decreasing = T),pch = 6, cex = 1)
```

Lampiran 6 *Stopwords Berbahasa Inggris*

a	appear	c	doesn't	follows
a's	appreciate	c'mon	doing	for
able	appropriate	c's	don't	former
about	Are	came	done	formerly
above	aren't	can	down	forth
according	around	can't	downwards	four
accordingly	As	cannot	during	from
across	Aside	cant	e	further
actually	Ask	cause	each	furthermore
after	asking	causes	edu	g
afterwards	associated	certain	eg	get
again	At	certainly	eight	gets
against	available	changes	either	getting
ain't	Away	clearly	else	given
all	awfully	co	elsewhere	gives
allow	B	com	enough	go
allows	Be	come	entirely	goes
almost	became	comes	especially	going
alone	because	concerning	et	gone
along	become	consequently	etc	got
already	becomes	consider	even	gotten
also	becoming	considering	ever	greetings
although	Been	contain	every	h
always	before	containing	everybody	had
am	beforehand	contains	everyone	hadn't
among	behind	corresponding	everything	happens
amongst	being	could	everywhere	hardly
an	believe	couldn't	ex	has
and	below	course	exactly	hasn't
another	beside	currently	example	have
any	besides	d	except	haven't
anybody	Best	definitely	f	having
anyhow	better	described	far	he
anyone	between	despite	few	he's
anything	beyond	did	fifth	hello
anyway	Both	didn't	first	help
anyways	Brief	different	five	hence
anywhere	But	do	followed	her
apart	By	does	following	here
here's	it'd	meanwhile	of	quite
hereafter	it'll	merely	off	qv
hereby	it's	might	often	r
herein	Its	more	oh	rather
hereupon	Itself	moreover	ok	rd

hers
herself
hi
him
himself
his
hither
hopefully
how
howbeit
however
i
i'd
i'll
i'm
i've
ie
if
ignored
immediate
in
inasmuch
inc
indeed
indicate
indicated
indicates
inner
insofar
instead
into
inward
is
isn't
it
should
shouldn't
since
six
so
some
somebody
somehow
someone
something
J
Just
K
Keep
keeps
Kept
Know
knows
known
L
Last
lately
Later
Latter
latterly
Least
Less
Lest
Let
let's
Like
Liked
likely
Little
Look
looking
Looks
Ltd
M
mainly
many
May
maybe
Me
Mean
theirs
Them
themselves
Then
thence
There
there's
thereafter
thereby
therefore
most
mostly
much
must
my
myself
n
name
namely
nd
near
nearly
necessary
need
needs
neither
never
nevertheless
new
next
nine
no
nobody
non
none
noone
nor
normally
not
nothing
novel
now
nowhere
o
obviously
try
trying
twice
two
u
un
under
unfortunately
unless
unlikely
okay
old
on
once
one
ones
only
onto
or
other
others
otherwise
ought
our
ours
ourselves
out
outside
over
overall
own
p
particular
particularly
per
perhaps
placed
please
plus
possible
presumably
probably
provides
q
que
welcome
well
went
were
weren't
what
what's
whatever
when
whence
re
really
reasonably
regarding
regardless
regards
relatively
respectively
right
s
said
same
saw
say
saying
says
second
secondly
see
seeing
seem
seemed
seeming
seems
seen
self
selves
sensible
sent
serious
seriously
seven
several
shall
she
wouldn't
x
y
yes
yet
you
you'd
you'll
you're
you've

sometime
sometimes
somewhat
somewhere
soon
sorry
specified
specify
specifying
still
sub
such
sup
sure
t
t's
take
taken
tell
tends
th
than
thank
thanks
thanx
that
that's
thats
the
their
therein
theres
thereupon
These
They
they'd
they'll
they're
they've
Think
Third
This
thorough
thoroughly
Those
though
Three
through
throughout
Thru
Thus
To
together
Too
Took
toward
towards
Tried
Tries
Truly
until
unto
up
upon
us
use
used
useful
uses
using
usually
uucp
v
value
various
very
via
viz
vs
w
want
wants
was
wasn't
way
we
we'd
we'll
we're
we've
whenever
where
where's
whereafter
whereas
whereby
wherein
whereupon
wherever
whether
which
while
whither
who
who's
whoever
whole
whom
whose
why
will
willing
wish
with
within
without
won't
wonder
would
would
your
yours
yourself
yourselves
z
zero

Lampiran 7 *Stopwords Berbahasa Indonesia*

Ada	bagai	berakhir	berturut-turut	dekat
Adalah	bagaikan	berakhirlah	bertutur	demi
Adanya	bagaimana	berakhirnya	berujar	demikian
Adapun	bagaimanakah	berapa	berupa	demikianlah
Agak	bagaimanapun	berapakah	besar	dengan
Agaknya	bagi	berapalah	betul	depan
Agar	bagian	berapapun	betulkah	di
Akan	bahkan	berarti	biasa	dia
Akankah	bahwa	berawal	biasanya	diakhiri
Akhir	bahwasanya	berbagai	bila	diakhirinya
Akhiri	baik	berdatangan	bilakah	dialah
Akhirnya	bakal	beri	bisa	diantara
aku	bakalan	berikan	bisakah	diantaranya
akulah	balik	berikut	boleh	diberi
amat	banyak	berikutnya	bolehkah	diberikan
amatlah	bapak	berjumlah	bolehlah	diberikannya
anda	baru	berkali-kali	buat	dibuat
andalah	bawah	berkata	bukan	dibuatnya
antar	beberapa	berkehendak	bukankah	didapat
antara	begini	berkeinginan	bukanlah	didatangkan
antaranya	beginian	berkenaan	bukannya	digunakan
apa	beginikah	berlainan	bulan	diibaratkan
apaan	beginilah	berlalu	bung	diibaratkannya
apabila	begitu	berlangsung	cara	diingat
apakah	begitukah	berlebihan	caranya	diingatkan
apalagi	begitulah	bermacam	cukup	diinginkan
apatah	begitupun	bermacam-macam	cukupkah	dijawab
artinya	bekerja	bermaksud	cukuplah	dijelaskan
asal	belakang	bermula	cuma	dijelaskannya
asalkan	belakangan	bersama	dahulu	dikarenakan
atas	belum	bersama-sama	dalam	dikatakan
atau	belumlah	bersiap	dan	dikatakannya
ataukah	benar	bersiap-siap	dapat	dikerjakan
ataupun	benarkah	bertanya	dari	diketahui
awal	benarlah	bertanya-tanya	daripada	diketahuinya
awalnya	berada	berturut	datang	dikira
dilakukan	ditanyakan	ibaratkan	kalaupun	kinilah
dilalui	ditegaskan	ibaratnya	kalian	kira
dilihat	ditujukan	ibu	kami	kira-kira
dimaksud	ditunjuk	ikut	kamilah	kiranya

dimaksudkan
dimaksudkannya
dimaksudnya
diminta
dimintai
dimisalkan
dimulai
dimulailah
dimulainya
dimungkinkan
dini
dipastikan
diperbuat
diperbuatnya
dipergunakan
diperkirakan
diperlihatkan
diperlukan
diperlukannya
dipersoalkan
dipertanyakan
dipunyai
diri
dirinya
disampaikan
disebut
disebutkan
disebutkannya
disini
disinilah
ditambahkan
ditandaskan
ditanya
ditanyai
masing
masing-masing
mau
maupun
melainkan
melakukan
melalui
melihat

ditunjuki
ditunjukkan
ditunjukkannya
ditunjuknya
dituturkan
dituturkannya
diucapkan
diucapkannya
diungkapkan
dong
dua
dulu
empat
enggak
enggaknya
entah
entahlah
guna
gunakan
hal
hampir
hanya
hanyalah
hari
harus
haruslah
harusnya
hendak
hendaklah
hendaknya
hingga
ia
ialah
ibarat
mendapat
mendapatkan
mendatang
mendatangi
mendatangkan
menegaskan
mengakhiri
mengapa

ingat
ingat-ingat
ingin
inginkah
inginkan
ini
inikah
inilah
itu
itukah
itulah
jadi
jadilah
jadinya
jangan
jangankan
janganlah
jauh
jawab
jawaban
jawabnya
jelas
jelaskan
jelaslah
jelasnya
jika
jikalau
juga
jumlah
jumlahnya
justru
kala
kalau
kalaulah
menyeluruh
menyiapkan
merasa
mereka
merekalah
merupakan
meski
meskipun

kamu
kamulah
kan
kapan
kapankah
kapanpun
karena
karenanya
kasus
kata
katakan
katakanlah
katanya
ke
keadaan
kebetulan
kecil
kedua
keduanya
keinginan
kelamaan
kelihatan
kelihatannya
kelima
keluar
kembali
kemudian
kesampaian
keseluruhan
keseluruhannya
keterlaluan
ketika
khususnya
kini
pasti
pastilah
penting
pentingnya
per
percuma
perlu
perlukah

kita
kitalah
kok
kurang
lagi
lagian
lah
lain
lainnya
lalu
lama
lamanya
lanjut
lanjutnya
lebih
lewat
lima
luar
macam
maka
makanya
makin
malah
malahan
mampu
mampukah
mana
manakala
manalagi
masa
masalah
masalahnya
masih
masihkah
sangatlah
satu
saya
sayalah
se
sebab
sebabnya
sebagai

melihatnya
memang
memastikan
memberi
memberikan
membuat
memerlukan
memihak
meminta
memintakan
memisalkan
memperbuat
mempergunakan
memperkirakan
memperlihatkan
mempersiapkan
mempersoalkan
mempertanyakan
mempunyai
memulai
memungkinkan
menaiki
menambahkan
menandaskan
menanti
menanti-nanti
menantikan
menanya
menanyai
menanyakan
mengatakan
mengatakannya
mengenai
mengerjakan
mengetahui
menggunakan
menghendaki
mengibaratkan
mengibaratkannya
mengingat
mengingatkan
menginginkan
mengira
mengucapkan
mengucapkannya
mengungkapkan
menjadi
menjawab
menjelaskan
menuju
menunjuk
menunjuki
menunjukkan
menunjuknya
menurut
menuturkan
menyampaikan
menyangkut
menyatakan
menyebutkan
semampu
semampunya
semasa
semasih
semata
semata-mata
semaunya
sementara
semisal
semisalnya
sempat
semua
meyakini
meyakinkan
minta
mirip
misal
misalkan
misalnya
mula
mulai
mulailah
mulanya
mungkin
mungkinkah
nah
naik
namun
nanti
nantinya
nyaris
nyatanya
oleh
olehnya
pada
padahal
padanya
pak
paling
panjang
pantas
para
sesuatu
sesuatunya
sesudah
sesudahnya
setelah
setempat
setengah
seterusnya
setiap
setiba
setibanya
setidak-tidaknya
perlunya
pernah
persoalan
pertama
pertama-tama
pertanyaan
pertanyakan
pihak
pihaknya
pukul
pula
pun
punya
rasa
rasanya
rata
rupanya
saat
saatnya
saja
sajalah
saling
sama
sama-sama
sambil
sampai
sampai-sampai
sampaikan
sana
sangat
tampaknya
tandas
tandasnya
tanpa
tanya
tanyakan
tanyanya
tapi
tegas
tegasnya
telah
tempat
sebagaimana
sebagainya
sebagian
sebaik
sebaik-baiknya
sebaiknya
sebaliknya
sebanyak
sebegini
sebegitu
sebelum
sebelumnya
sebenarnya
seberapa
sebesar
sebetulnya
sebisanya
sebuah
sebut
sebutlah
sebutnya
secara
secukupnya
sedang
sedangkan
sedemikian
sedikit
sedikitnya
seenaknya
segala
tersampaikan
tersebut
tersebutlah
tertentu
tertuju
terus
terutama
tetap
tetapi
tiap
tiba
tiba-tiba

sekali-kali
sekalian
sekaligus
sekalipun
sekarang
sekarang
sekecil
seketika
sekiranya
sekitar
sekitarnya
sekurang-kurangnya
sekurangnya
sela
selain
selaku
selalu
selama
selama-lamanya
selamanya
selanjutnya
seluruh
seluruhnya
semacam
semakin
semuanya
semula
sendiri
sendirian
sendirinya
seolah
seolah-olah
seorang
sepanjang
sepantasnya
sepantasnyalah
seperlunya
seperti
sepertinya
sepihak
sering
seringnya
serta
serupa
sesaat
sesama
sesampai
sesegera
sesekali
seseorang
setidaknya
setinggi
seusai
sewaktu
siap
siapa
siapakah
siapapun
sini
sinilah
soal
soalnya
suatu
sudah
sudahkah
sudahlah
supaya
tadi
tadinya
tahu
tahun
tak
tambah
tambahnya
tampak
tengah
tentang
tentu
tentulah
tentunya
tepat
terakhir
terasa
terbanyak
terdahulu
terdapat
terdiri
terhadap
terhadapnya
teringat
teringat-ingat
terjadi
terjadilah
terjadinya
terkira
terlalu
terlebih
terlihat
termasuk
ternyata
tidak
tidakkah
tidaklah
tiga
tinggi
toh
tunjuk
turut
tutur
tuturnya
ucap
ucapnya
ujar
ujarnya
umum
umumnya
ungkap
ungkapnya
untuk
usah
usai
waduh
wah
wahai
waktu
waktunya
walau
walaupun
wong
yaitu
yakin
yakni
yang
yg
ya
x
yagn
you
utk
u
sy

Lampiran 8 *Output SVM* Berbahasa Inggris dan Berbahasa Indonesia

Bahasa Inggris

```
> library(RTextTools)
> library(e1071)
>
> mat = create_matrix(sari.all, language = "english", removeStopwords = FALSE,
+     removeNumbers = TRUE, stemWords = FALSE, tm::weightTfIdf)
> mat = as.matrix(mat)
>
> #SVM
> container <- create_container(mat, sentiment_all, trainSize=1:1612,testSize=1613:1943, virgin=FALSE)
>
> model <- train_model(container, 'SVM',kernel='sigmoid')
> model

Call:
svm.default(x = container@training_matrix, y = container@training_codes,
    kernel = kernel, cost = cost, cross = cross, probability = TRUE,
    method = method)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  sigmoid
       cost:  100
      gamma:  0.0002350729
     coef.0:  0

Number of Support Vectors:  369

> results <- classify_model(container, model)
> table(as.character(sentiment_all[1613:1943]), as.character(results[,"SVM_LABEL"]))

          negatif positif
  negatif       3      13
  positif       0     315
> recall_accuracy(sentiment_all[1613:1943], results[,"SVM_LABEL"])
[1] 0.9607251
```

Bahasa Indonesia

```
> library(e1071)
>
> mat = create_matrix(sari.all, language = "indonesia", removeStopwords = FALSE,
+     removeNumbers = TRUE, stemWords = FALSE, tm::weightTfIdf)
> mat = as.matrix(mat)
>
> #SVM
> container <- create_container(mat, sentiment_all, trainSize=1:931,testSize=932:1233, virgin=FALSE)
>
> model <- train_model(container, 'SVM',kernel='sigmoid')
> model

Call:
svm.default(x = container@training_matrix, y = container@training_codes,
    kernel = kernel, cost = cost, cross = cross, probability = TRUE,
    method = method)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  sigmoid
       cost:  100
      gamma:  0.0003263708
     coef.0:  0

Number of Support Vectors:  428

> results <- classify_model(container, model)
> table(as.character(sentiment_all[932:1233]), as.character(results[,"SVM_LABEL"]))

          negatif positif
  negatif      45      40
  positif       6     211
> recall_accuracy(sentiment_all[932:1233], results[,"SVM_LABEL"])
[1] 0.8476821
```