IMPLEMENTASI TEKNIK WEB SCRAPING DAN KLASIFIKASI SENTIMEN MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER DAN ASOSIASI TEKS

(Studi Kasus : Data Ulasan Penumpang Maskapai Penerbangan Garuda Indonesia Pada Situs TripAdvisor)

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana

Program Studi Statistika

Muhammad Mulajati

13 611 217

JURUSAN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2017

HALAMAN PERSETUJUAN PEMBIMBING

TUGAS AKHIR

Judul	:	Implementasi Teknik Web Scraping dan Klasifikasi Sentimen Menggunakan Metode Naïve Bayes Classifier dan Asosiasi Teks (Studi Kasus : Data Ulasan Penumpang Maskapai Penerbangan Garuda Indonesia Pada Situs TripAdvisor)
Nama Mahasiswa	:	Muhammad Mulajati
Nomor Mahasiswa	:	13 611 217

TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK DIUJIKAN

Yogyakarta, 30 Mei 2017

Pembimbing

(Dr. RB. Fajriya Hakim, S.Si., M.Si.)

HALAMAN PENGESAHAN

TUGAS AKHIR

IMPLEMENTASI TEKNIK WEB SCRAPING DAN KLASIFIKASI SENTIMEN MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER DAN ASOSIASI TEKS

(Studi Kasus : Data Ulasan Penumpang Maskapai Penerbangan Garuda Indonesia Pada Situs TripAdvisor)

Nama Mahasiswa : Muhammad Mulajati

Nomor Mahasiswa : 13 611 217

TUGAS AKHIR INI TELAH DIUJIKAN

PADA TANGGAL 21 JULI 2017

	Nama Penguji	**Tanda tangan**
1	Ir. Ali Parkhan, M.T.	.....
2	Tuti Purwaningsih, S.Stat., M.Si.	.....
3	Dr. RB. Fajriya Hakim, 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 senantiasa saya panjatkan kehadirat Allah SWT yang telah melimpahkan rahmat, hidayah, dan nikmatnya yang tak terhingga, sehingga penulis dapat menyelesaikan tugas akhir yang berjudul **"*Implementasi Teknik Web Scraping dan Klasifikasi Sentimen Menggunakan Metode Naïve Bayes Classifier dan Asosiasi Teks (Studi Kasus : Data Ulasan Penumpang Maskapai Penerbangan Garuda Indonesia Pada Situs TripAdvisor)*"** sebagai salah satu persyaratan yang harus dipenuhi dalam menyelesaikan jenjang strata satu di Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia. Shalawat serta salam semoga selalu tercurah kepada Nabi Muhammad SAW serta para sahabat dan pengikutnya yang senantiasa menjaga keimanan dan keislamannya hingga akhir hayatnya.

Penyelesaian tugas akhir ini tidak terlepas dari dukungan, bantuan, arahan, dan bimbingan dari berbagai pihak. Untuk itu pada kesempatan ini penulis ingin mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Allah SWT yang telah memberikan rahmat, nikmat sehat dan kesempatan sehingga penyusun dapat melaksanakan penelitian hingga menyusun Tugas Akhir dengan baik.
2. Rasulullah Muhammad SAW, atas segala cintanya kepada umat manusia dan merupakan suri tauladan sepanjang zaman, semoga kita mendapatkan syafa'at beliau di akhirat kelak. Aamiin.
3. Kedua orang tua saya yang sangat saya cintai, Dae dan Mama yang selalu memberikan semangat, do'a dan dukungan disetiap langkah saya.
4. Adik-adik saya, Nurul Istiqomah dan Agil Mubarak, serta Keluarga Besar saya yang selalu mendo'akan, mendukung dan memberikan semangat kepada saya.

5. Drs. Allwar. M.Sc, Ph.D. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam.
6. Bapak Dr. RB. Fajriya Hakim, M.Si. selaku Ketua Program Studi Statistika sekaligus sebagai pembimbing saya, yang telah banyak memberikan dukungan dan masukan yang membangun serta selalu bersedia meluangkan waktunya untuk berkonsultasi dan memberikan arahan yang sangat inspiratif.
7. Seluruh staf pengajar Program Studi Statistika Universitas Islam Indonesia yang telah memberikan bekal ilmu kepada penulis, sehingga penulis dapat menyelesaikan tugas akhir ini.
8. Keluarga besar INVISIO STATISTIKA UII, sebagai tim promosi dan publikasi prodi Statistika UII, terimakasih atas kebersamaan, kekeluargaan, kekompakan, keceriaan dan pelajaran berharga lainnya.
9. Ibu/Mbak Asmadhini Handayani R., S.Si., M.M., selaku dosen yang juga sudah saya anggap sebagai kakak saya sendiri, terimakasih atas semua kebaikan, bimbingan, dan semua pelajaran yang telah berikan, semoga segala kebaikannya menjadi amal baik dan mendapatkan balasan dari Allah SWT.
10. Sahabat sepermainan dan seperjuangan (INVISIO Gen I) : Baron Setyo Utomo, Slamet Abtohi, Khair Norrasid, Feby Syafitri, Yenni Tria Paramitha dan Alfianisa Shafira yang selalu menemani, mendukung, memberi semangat dan telah berjuang bersama-sama selama beberapa tahun ini. Semoga ukhuwah kita tetap terjaga dan selalu diridhoi Allah SWT.
11. Teman-teman bimbingan TA : Tohi, Febi, Lusi, Baron, Yeni, Nurul, Iva, Bella, Zara, Ema, Zhazen, Gina, Bang Tama, dan Bang Aznin yang sudah sama-sama berjuang, saling mengingatkan dan memberi motivasi serta dorongan untuk menyelasaikan Tugas Akhir ini.
12. Keluarga "Dapluk" yang selalu setia menemani, memberikan keceriaan dan hiburannya. Semoga kebersamaan kita tetap selalu terjaga.
13. Keluarga Besar Forum Mahasiswa Sila (FORMASI) Yogyakarta, sebagai salah satu wadah perkumpulan mahasiswa Sila - BIMA NTB yang sedang sama-sama berjuang menempuh studi di tanah rantauan (Yogyakarta). Terimakasih atas kebersamaannya sehingga tanah rantauanpun terasa seperti kampung halaman.

14. Teman-teman KKN UII Unit 380 Pucung, Girisubo, Gunungkidul : Alan, Ibnu, Angga, Ulya, Andra, Yuri, dan Sonia. Terimakasih atas pelajaran yang sangat berharga selama satu bulan bersama di lokasi KKN.
15. Mbak Gebri, Mbak Sisca, Kak Yunia, dan Mas Ulwan, selaku senior yang telah banyak membantu, memberi nasehat, arahan, dan bimbingan dalam menyelesaikan skripsi ini.
16. Teman-teman "Akal 06" SMAN 1 BOLO : Ahmaddin, Dayat, Fauzan, Gufran, dan Uswatun, yang selalu bersama dan saling menyemangati.
17. Teman-teman Statistika UII Angkatan 2013 yang bersama-sama menjadi pejuang gelar S.Stat dan Toga UII, terimakasih semangatnya.
18. Pihak-pihak lain yang mungkin penulis belum sebutkan, yang telah membantu dalam penyusunan tugas akhir ini.

Demikian Tugas Akhir ini, penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan bantuan baik moril maupun materil sehingga tugas akhir ini dapat diselesaikan. Penulis menyadari bahwa tugas akhir ini masih jauh dari kata sempurna dan masih banyak kekurangan. Hal tersebut dikarenakan keterbatasan ilmu dan pengetahuan yang dimiliki penulis semata. Oleh karena itu penulis mengharapkan kritik dan saran dari pembaca untuk menyempurnakan penulisan laporan ini. Semoga Tugas Akhir ini dapat memberikan manfaat bagi penulis khususnya dan umumnya bagi semua pihak yang membutuhkan. Akhir kata, semoga Allah SWT senantiasa melimpahkan rahmat serta hidayah-Nya kepada kita semua, Amin amin ya robbal 'alamiin.

Wassalamu'alaikum Warahmatullaahi Wabarakaatuh

Yogyakarta, 30 Mei 2017

Penulis

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, 30 Mei 2017

Muhammad Mulajati

IMPLEMENTASI TEKNIK WEB SCRAPING DAN KLASIFIKASI SENTIMEN MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER DAN ASOSIASI TEKS

(Studi Kasus : Data Ulasan Penumpang Maskapai Penerbangan Garuda Indonesia Pada Situs TripAdvisor)

Muhammad Mulajati

Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Islam Indonesia

INTISARI

Meningkatnya kebutuhan akan informasi mendorong manusia untuk mengembangkan teknologi-teknologi baru agar pengolahan data dan informasi dapat dilakukan dengan mudah dan cepat. Di dunia sekarang ini, hampir semua data yang dibutuhkan sudah tersedia di internet, satu-satunya hal yang membatasi untuk menggunakannya adalah kemampuan untuk mengaksesnya. *Web scraping* merupakan salah satu solusi yang dapat dilakukan untuk memperoleh data atau informasi dari sebuah situs atau *website*. Penelitian ini menggunakan teknik *web scraping* untuk memperoleh data ulasan maskapai penerbangan Garuda Indonesia dari situs *TripAdvisor*. Data yang diperoleh dari situs *TripAdvisor* selanjutnya dilakukan pelabelan dan dianalisis dengan menggunakan metode *Naïve Bayes Classifier* (NBC) untuk mengklasifikasikan ulasan berdasarkan kategori sentimen positif dan negatif. Dari hasil pelabelan yang telah dilakukan kemudian akan dilihat asosiasi teks pada setiap kelas sentimen untuk menemukan sebuah fakta dan informasi yang dianggap penting dan dapat berguna untuk pengambilan keputusan. Hasil pelabelan kelas sentimen pada data ulasan didapatkan bahwa dari total 1143 ulasan, jumlah ulasan positif adalah sebanyak 976 ulasan, dan ulasan negatif adalah sebanyak 167 ulasan. Dari data tersebut selanjutnya dibuat perbandingan data latih dan data uji sebesar 80% : 20% dan diperoleh hasil klasifikasi sentimen dari data uji tersebut dengan menggunakan model *Naïve Bayes Classifier* (NBC) diperoleh tingkat akurasi sebesar 82,02%, artinya dari 228 data ulasan yang diujikan, terdapat 187 ulasan yang benar pengklasifikasiannya. Selanjutnya, pada proses asosiasi teks diperoleh informasi bahwa penumpang maskapai Garuda Indonesia mayoritas membicarakan mengenai *service*, *staff*, *food*, dan *check-in* karena selalu muncul baik pada kelas sentimen positif maupun negatif. Secara umum metode asosiasi teks yang digunakan menunjukkan hasil ekstraksi informasi pada kelas positif diantaranya terkait *service, food, seat, time, staff, entertainment, check-in,* dan *cabin*. Sedangkan pada kelas negatif yang sering dikeluhkan diantaranya *service, staff, seat, food, hour, check-in, luggage,* dan *boarding*.

Kata kunci : *Web Scraping*, Analisis Sentimen, *Naïve Bayes Classifier*, *Text Mining*, Asosiasi Teks, Garuda Indonesia, *TripAdvisor*.

IMPLEMENTATION OF WEB SCRAPING TECHNIQUES AND SENTIMENT CLASSIFICATION USING NAÏVE BAYES CLASSIFIER METHOD AND TEXT ASSOCIATION

(Case Study : Garuda Indonesia Airlines Passengers Reviews on TripAdvisor Site)

Muhammad Mulajati

Department of Statictics, Faculty of Mathematics and Natural Sciences

Islamic University of Indonesia

ABSTRACT

The increased need for information encourages people to develop new technologies for data processing and information can be done easily and quickly. In today's world, almost all of the required data is already available on the internet, the only thing limiting it to using it is the ability to access it. Web scraping is one solution that can be done to obtain data or information from a website. This study uses web scraping techniques to obtain Garuda Indonesia airline reviews data from the TripAdvisor site. Data obtained from the TripAdvisor site is further labeled and analyzed using the Naïve Bayes Classifier (NBC) method to classify reviews based on positive and negative sentiment categories. From the labeling results that have been done then will be seen text associations on each class of sentiment to find a fact and information that is considered important and can be useful for decision-making. The results of the sentiment class labeling on the reviews data obtained that from a total of 1143 reviews, the number of positive reviews is as much 976 reviews, and the negative reviews was 167 reviews. From the data then made comparison of trainer data and test data of 80%: 20% and obtained the results of the classification of sentiments from the test data using the Naïve Bayes Classifier (NBC) model obtained an accuracy of 82.02%, which means that of 228 tested review data, there are 187 reviews that correctly classify it. Furthermore, in the text association process, it is found that passengers that Garuda Indonesia airline mostly discuss about service, staff, food, and check-in as it always appears in both positive and negative sentiment class. In general, the text association method used shows the extraction of information on positive classes such as service, food, seat, time, staff, entertainment, check-in, and cabin. While in the negative class that is often complained of including service, staff, seats, food, hour, check-in, luggage, and boarding.

Keywords : *Web Scraping, Sentiment Analysis, Naïve Bayes Classifier, Text Mining, Text Association, Garuda Indonesia, TripAdvisor.*

BAB I
PENDAHULUAN

1.1. Latar Belakang Masalah

Meningkatnya kebutuhan akan informasi mendorong manusia untuk mengembangkan teknologi-teknologi baru agar pengolahan data dan informasi dapat dilakukan dengan mudah dan cepat. Salah satu teknologi yang sedang berkembang dengan pesat saat ini adalah teknologi informasi/komputer (Abdillah dan Emigawaty, 2009 dikutip dalam Josi, dkk. 2014). Perkembangan pesat teknologi informasi dan komunikasi, khususnya internet telah menjadi bagian penting dalam aktifitas sehari-hari. Berdasarkan data hasil survei dari Asosiasi Penyelenggara Jasa Internet Indonesia (APJII 2016), jumlah pengguna internet di Indonesia pada tahun 2016 telah mencapai 132,7 juta pengguna. Angka ini jauh lebih tinggi dari pada tahun 2014 yaitu sebanyak 88 juta pengguna. Angka tersebut juga menunjukkan jumlah yang melampaui setengah dari total penduduk Indonesia, yaitu sekitar 51,8%.

Sejak munculnya internet membuat informasi yang terkandung dalam *web* berkembang secara pesat. Banyak pihak yang memanfaatkan internet untuk berbagai tujuan. Akibatnya, informasi yang beredar di internet terus meningkat secara eksponensial. Semakin bebasnya ruang komunikasi dalam ranah dunia maya serta munculnya berbagai situs maupun forum internet, menyebabkan beberapa hal, baik yang positif maupun negatif. Salah satunya, berpengaruh dalam dinamika komunikasi pemasaran. Pemasaran di dunia maya yang disebut juga *e-commerce* menjadi bentuk baru dalam komunikasi pemasaran yang cukup terkenal di masa kini. Konsumen di masa sekarang sangat kritis, dan cenderung mencari informasi atau referensi dan mempercayai opini-opini yang terdapat didalam komunitas tentang suatu produk maupun jasa yang berada di sekitarnya. Fenomena ini dalam istilah pemasaran sering disebut dengan *word-of-mouth* (WOM). Tidak bisa dipungkiri bahwa kekuatan *word-of-mouth* berpengaruh besar pada keputusan pembelian konsumen dan dalam pembentukan pola perilaku konsumen. Banyak penelitian telah menunjukkan bahwa WOM memberikan pengaruh yang cukup kuat

dibanding dengan media-media komunikasi tradisional yang umum adanya seperti iklan maupun advertorial. Di dalam dunia internet, komunikasi jenis ini disebut *electronic word-of-mouth* atau e-wom. Semakin banyaknya situs dan forum yang bermunculan di internet, mampu menciptakan kesempatan bagi e-WOM melalui berbagai media seperti forum diskusi, *web-based opinion platform, newsgroup, blogs, review sites, bulletin board systems, review sites* dan *social networking sites*. e-WOM menyebabkan konsumen tidak hanya mendapatkan informasi mengenai suatu produk dan jasa dari orang-orang yang sudah dikenal, melainkan juga dari sekelompok orang dari berbagai belahan dunia yang memiliki pengalaman terhadap produk dan jasa yang dimaksud. Berkembangnya e-WOM tidak hanya terjadi di lingkup sektor perdagangan maupun penjualan saja, tapi juga berkembang di sektor pariwisata atau *Travel*.

Pariwisata merupakan sebuah aktivitas popular dalam kehidupan modern yang telah banyak memberikan kontribusi positif yang signifikan dalam perkembangan ekonomi diseluruh dunia. Dengan adanya persaingan yang kian meningkat diantara para pelaku jasa pariwisata, metode yang digunakan dalam menarik turis untuk datang kesuatu destinasi wisata mendapatkan atensi penuh dari para periset, pemasar, maupun pemerintah. Salah satunya, dengan melalui situs atau forum pariwisata. Sebelum merencanakan suatu perjalan, para turis umumnya melakukan pencarian informasi mengenai destinasi wisata melalui media tersebut. Saat ini dimana kemunculan situs maupun forum pariwisata yang memberikan banyak informasi mengenai tujuan pariwisata, bahkan melingkupi reservasi tiket penerbangan maupun akomodasi sangat menjamur, salah satu situs dan forum pariwisata yang sangat dikenal di seluruh dunia adalah *TripAdvisor.com*.

TripAdvisor adalah salah satu situs wisata terbesar di dunia yang membantu wisatawan mengoptimalkan potensi setiap perjalanan. *TripAdvisor* menawarkan saran 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, restoran, dan juga transportasi penjalanan terbaik. Pada halaman *TripAdvisor* juga menyediakan informasi mengenai ulasan-ulasan para wisatawan tentang suatu hotel, restoran, dan beberapa maskapai penerbangan.

Melalui *TripAdvisor* wisatawan yang sudah menggunakan beberapa layanan tersebut bisa menuliskan pengalamannya sehingga wisatawan potensial yang melihat ulasan tersebut bisa tertarik untuk kembali menggunakan jasa yang sama.

Salah satu yang tidak kalah penting dalam suatu perjalanan wisata adalah mengenai transportasi. Pariwisata tidak bisa putus dari dunia transportasi sebagai salah satu faktor penunjang pariwisata. Aktivitas kepariwisataan banyak bergantung pada transportasi dan komunikasi. Faktor jarak dan waktu sangat mempengaruhi keinginan seseorang untuk melakukan perjalanan wisata. Dewasa ini transportasi menyebabkan pertumbuhan pariwisata yang sangat pesat sekali. Kemajuan fasilitas transportasi mendorong kemajuan kepariwisataan dan sebaliknya ekspansi yang terjadi dalam industri pariwisata dapat menciptakan permintaan akan transportasi yang dapat memenuhi kebutuhan wisatawan. Tidak dapat disangkal lagi bahwa fungsi utama transportasi sangat erat hubungannya dengan "*accessibility*". Maksudnya, frekuensi penggunaannya, kecepatan yang dimilikinya dapat mengakibatkan jarak yang jauh seolah-olah menjadi lebih dekat. Hal ini berarti mempersingkat waktu dan tentunya akan lebih meringankan biaya perjalanan. Dengan demikian transportasi dapat memudahkan orang untuk mengunjungi suatu daerah tertentu, seperti misalnya daerah tujuan wisata. Selain menjadi salah satu situs dan forum pariwisata yang sangat dikenal di seluruh dunia, *TripAdvisor* juga menawarkan jasa revervasi tiket perjalanan khususnya transportasi udara. Tidak hanya itu, pada situs *TripAdvisor* juga dilengkapi dengan ulasan atau komentar wisatawan yang sudah menggunakan beberapa jasa maskapai penerbangan tersebut.

Banyak maskapai penerbangan yang sudah menjalin kerjasama dengan *TripAdvisor* dalam hal reservasi tiket perjalanan salah satunya adalah maskapai Garuda Indonesia. Garuda Indonesia adalah maskapai pertama dan tertua di Indonesia yang dimiliki oleh pemerintah Republik Indonesia. Garuda Indonesia adalah maskapai penerbangan nasional Indonesia yang terbang ke lebih dari 40 tujuan domestik dan 36 tujuan internasional. Garuda Indonesia meraih penghargaan sebagai Maskapai Penerbangan Regional Terbaik di Dunia yang diberikan oleh *Skytrax*. Terbang untuk pertama kalinya di tahun 1949, saat ini Garuda Indonesia

membawa lebih dari 25 juta penumpang setiap tahunnya. Sebagai salah satu maskapai nasional yang dimiliki Indonesaia tentunya Garuda Indonesia harus tetap konsisten dalam memberikan pelayanan yang baik kepada para pelanggannya.

Untuk itu penyusunan kebijakan sangat perlu dilakukan untuk mengetahui dan memahami segala bentuk aspirasi dan keluhan penumpang atau pelanggannya. Pemerintah (BUMN) sebagai pemilik dari maskapai tersebut perlu mengetahui isu yang paling banyak menjadi sorotan pelanggannya serta masalah apa yang sedang terjadi dan belum terselesaikan dengan baik. Saat ini, tingkat kepuasan pelanggan dapat diperoleh secara mudah dengan adanya layanan-layanan *online* berbasis pariwisata, salah satunya adalah situ *TripAdvisor*. Namun dengan banyaknya jumlah data ulasan pengunjung yang masuk ke situs *TripAdvisor* mengakibatkan sulitnya pihak maskapai dalam memperoleh informasi secara keseluruhan dari semua ulasan. Dengan banyaknya data ulasan yang masuk, maka diperlukan teknik khusus untuk melakukan ekstraksi data pada suatu halaman *website* tersebut salah satunya dengan menggunakan teknik *Web scraping*. *Web scraping* menjadi salah satu solusi yang dapat digunakan untuk dilakukan untuk memperoleh data atau informasi dari sebuah situs atau *website* sebelum data tersebut diolah dan dianalisis. Kemudian pada proses analisisnya, penulis mencoba untuk melakukan klasifikasi teks ulasan penumpang guna mengidentifikasi mana ulasan yang berbentuk positif dan negatif. Setalah melakukan klasifikasi, penulis mencoba mengekstrak dan mengeksplorasi seluas-luasnya informasi apa yang ada pada ulasan-ulasan tersebut yang sekiranya dianggap penting untuk digunakan pada berbagai keperluan.

Dalam penelitian ini proses klasifikasi sentimen akan dilakukan dengan menggunakan metode *Naïve Bayes Classifier*. Kemudian untuk proses ektraksi dan eksplorasi penulis menggunakan statistik deskriptif dan asosiasi antar *terms* (kata atau topik yang sering dibicarakan) yang saling berkaitan. Metode *Naïve Bayes Classifier* ini adalah metode klasifikasi biner yang memanfaatkan probabilitas statistika sederhana dengan menerapkan aturan *Bayesian* menggunakan asumsi independen yang kuat. Metode *Bayesian* merupakan metode analisis berdasar informasi sampel dan informasi prior. Gabungan dari informasi sampel dengan informasi prior tersebut dinamakan peluang posterior. Penerapan metode *Naïve*

Bayes Classifier adalah dengan memanfaatkan data *training* untuk menguji data *testing*.

Harapannya dengan penelitian ini mampu mengklasifikasikan teks dengan baik sehingga nantinya informasi yang ada di dalamnya dapat diekstraksi dengan baik serta penyajian informasi dari data yang diamati dapat memberikan informasi yang berguna bagi berbagai pihak yang membutuhkannya.

1.2. Rumusan Masalah

Berdasarkan permasalahan diatas, adapun permasalahan yang akan dikaji dalam penelitian ini adalah sebagai berikut:

1. Bagaimana mengimplementasikan teknik *web scraping* untuk mendapatkan data ulasan maskapai penerbangan Garuda Indonesia dari situs *TripAdvisor*?
2. Bagaimana gambaran umum mengenai presepsi pelanggan maskapai penerbangan Garuda Indonesia pada *website TripAdvisor*?
3. Bagaimana hasil penerapan metode *Naïve Bayes Classifier* dalam mengklasifikasikan data ulasan maskapai Garuda Indonesia menjadi kelas positif dan negatif?
4. Informasi apa yang didapatkan dalam setiap klasifikasi dan asosiasi teks yang telah dilakukan?

1.3. Batasan Masalah

Adapun batasan masalah yang digunakan peneliti agar pembahasan dalam penelitian ini tidak menyimpang dari pokok pembahasan. Maka peneliti memiliki batasan masalah sebagai berikut:

1. Penelitian ini menggunakan data ulasan mengenai maskapai Garuda Indonesia pada *website TripAdvisor* sejak bulan Januari 2016 – Maret 2017.
2. Ulasan yang diambil atau digunakan adalah ulasan yang berbahasa Internasional (Bahasa Inggris).

1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Mengimplementasikan teknik *web scraping* untuk mendapatkan data ulasan maskapai penerbangan Garuda Indonesia dari situs *TripAdvisor*.
2. Mengetahui gambaran umum tentang presepsi pelanggan maskapai penerbangan Garuda Indonesia berdasarkan data ulasan penumpang pada *website TripAdvisor*.
3. Menerapkan metode *Naïve Bayes Classifier* dalam mengklasifikasikan data ulasan maskapai Garuda Indonesia menjadi kelas positif dan negatif.
4. Mendapatkan informasi yang penting dan berguna dalam setiap klasifikasi dan asosiasi teks yang dilakukan.

1.5. Manfaat Penelitian

Adapun maanfaat dari penelitian ini yakni sebagai berikut:

1. Mengetahui penerapan teknik *web scraping* untuk memperoleh data ulasan dari situs *TripAdvisor*.
2. Mengetahui gambaran umum tentang presepsi pelanggan maskapai penerbangan Garuda Indonesia berdasarkan data ulasan penumpang pada *website TripAdvisor*.
3. Pengklasifikasian ulasan penumpang dapat memudahkan pihak yang memiliki kepentingan agar lebih cepat dan efisien dalam mencari apa yang dibutuhkannya.
4. Pihak yang memiliki kepentingan dapat melihat informasi yang tersembunyi dalam kumpulan komentar yang sangat banyak, sehingga dapat dilakukan penanganan dan fokus untuk evaluasi kearah yang lebih baik.

1.6. Sistematika Penulisan

Sistematika penulisan yang dipergunakan dalam penulisan tugas akhir ini dapat diuraikan sebagai berikut:

BAB I PENDAHULUAN

Pada bab ini akan dibahas tentang latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini memaparkan penelitia-penelitian terdahulu yang berhubungan dengan permasalahan yang diteliti dan menjadi acuan konseptual.

BAB III LANDASAN TEORI

Pada bab ini akan dibahas tentang teori-teori dan konsep yang berhubungan dengan penelitian yang dilakukan dan mendukung dalam pemecahan masalahnya. Selain itu, bab ini juga memuat teori-teori dalam pelaksanaan pengumpulan dan pengolahan data serta saat melakukan penganalisaan.

BAB IV METODOLOGI PENELITIAN

Bab ini memaparkan populasi dan sampel, variabel penelitian, jenis dan sumber data, metode analisis data, dan tahapan penelitian.

BAB V ANALISIS DAN PEMBAHASAN

Pada bab ini akan dibahas mengenai analisa yang dilakukan terhadap hasil pengumpulan, pengolahan dan analisa data yang diperoleh dari hasil penelitian.

BAB VI PENUTUP

Pada bab ini akan dibahas mengenai kesimpulan yang diperoleh dari hasil penelitian dan analisa data yang telah dilakukan serta saran-saran yang dapat diterapkan dari hasil pengolahan data yang dapat menjadi masukan yang berguna kedepannya.

BAB II
TINJAUAN PUSTAKA

2.1. Pariwisata

Definisi pariwisata dapat ditinjau dari berbagai sudut pandang dan tidak memiliki batasan-batasan yang pasti. Para ahli pariwisata banyak mengungkapkan definisi pariwisata dari berbagai sudut pandang, namun dari berbagai definisi tersebut memiliki makna yang sama. Menurut Suwantoro (2004), pariwisata adalah suatu proses kepergian sementara dari seseorang atau lebih menuju tempat lain di luar tempat tinggalnya. Dorongan kepergiannya adalah karena kepentingan, baik karena kepentingan ekonomi, sosial, kebudayaan, politik, agama, kesehatan maupun kepentingan lain seperti karena sekedar ingin tahu, menambah pengalaman ataupun belajar.

Adapun definisi pariwisata lain yang diungkapkan oleh Hunzieker dan Kraft (1942) dalam Yoeti (2001) mengemukakan definisi pariwisata dengan batasan yang lebih bersifat teknis yang diterima secara *official* oleh *The Association Experts Scientific Internationale des Experts Scientifique du Tourisme* (AIEST), batasan yang diberikan sebagai berikut : "*Tourism is the sum of the phenomenom and relationships arising from the travel and stay of non resident, in so far as they do not lead to permanent residence and are not connected with any earning activity*" (pariwisata adalah gabungan dari gejala dan hubungan-hubungan yang muncul dari adanya perjalanan dan tinggal sementara dari orang-orang yang bukan penduduk setempat, sejauh mereka tidak menunjukkan keinginan untuk menetap dan sejauh mereka tidak berhubungan dengan kegiatan yang menghasilkan uang).

Sugiama (2011) mengungkapkan bahwa pariwisata adalah rangkaian aktivitas, dan penyediaan layanan baik untuk kebutuhan atraksi wisata, transportasi, akomodasi, dan layanan lain yang ditujukan untuk memenuhi kebutuhan perjalanan seseorang atau sekelompok orang. Perjalanan yang dilakukan hanya untuk sementara waktu saja meninggalkan tempat tinggalnya dengan maksud beristirahat, berbisnis atau untuk maksud lainnya.

Berdasarkan seluruh definisi diatas dapat disimpulkan bahwa pariwisata adalah kegiatan yang dilakukan dengan meninggalkan tempat tinggalnya ke daerah tujuan wisata untuk sementara waktu dan bukan untuk menetap. Kegiatan perjalanannya bertujuan untuk menikmati layanan dan fasilitas yang dibutuhkan selama berada diluar tempat tinggalnya.

2.2. *E-Commerce*

E-commerce menurut Kotler dan Armstrong (2012) merupakan usaha untuk menginformasikan, mengkomunikasikan, mempromosikan dan menjual produk dan jasanya melalui internet. Maka dapat disimpulkan bahwa *e-commerce* merupakan usaha perusahaan untuk mempromosikan, mengkomunikasikan, menginformasikan dan menjual produknya baik berupa barang ataupun jasa dengan menggukanakan aplikasi elektronik salah satunya melalui internet.

2.3. *Word of Mouth*

Hardjana (2003) mendefinisikan komunikasi interpersonal sebagai interaksi tatap muka antara dua atau beberapa orang, sehingga komunikator dapat menyampaikan pesan secara langsung dan komunikan dapat menerima serta menanggapi secara langsung pula. Dalam dunia pemasaran, komunikasi interpersonal antar konsumen dapat berupa penyampaian pesan *Word of Mouth* (WOM). Sen dan Lerman (2007) mendefinisikan WOM sebagai komunikasi informal pribadi ke pribadi antar komunikator (yang dinilai bukan bagian dari pemasar/sumber komersial) dan penerima mengenai produk, merek, pelayanan, dan perusahaan. Komunikasi WOM terjadi ketika konsumen memberikan saran atau pendapat dan berbagai pengalaman kepada konsumen lain tentang sebuah produk, jasa, atau merek (Schiffman dan Kanuk, 2010).

2.4. E-WOM

Schiffman dan Kanuk (2010) mendefinisikan e-WOM sebagai *Word of Mouth* yang dilakukan secara *online*. Definisi lain menyebutkan bahwa komunikasi e-WOM adalah pernyataan positif atau negatif yang dibuat oleh konsumen

potensial, konsumen riil, atau mantan konsumen tentang sebuah produk atau perusahaan yang dapat diakses oleh banyak orang atau institusi melalui internet (Hennig-Thurau, et al., 2004). Harrison-Walker (2001) mengungkapkan e-WOM sebagai komunikasi informal antar individu, terjadi diantara komunikator dan penerima secara non-komersil berkaitan dengan sebuah merek, produk, organisasi, ataupun jasa.

2.5. *Travel Website*

Dalam Suryana dan Koesheryatin (2014) disebutkan bahwa *website* berisi halaman-halaman yang dapat menampilkan teks, gambar, grafik, suara serta elemen-elemen multimedia lainnya.

"A travel website is a website on the world wide web that is dedicated to travel. The site may be focused on travel reviews, the booking of travel, or a combination of both." (Saks, 2006)

Terjemahan, suatu situs perjalanan adalah situs dalam jaringan dunia luas yang didedikasikan pada perjalanan. Situs ini memfokuskan pada ulasan tentang perjalanan, pemesanan perjalanan atau kombinasi keduanya.

Oleh karena itu bisa disimpulkan bahwa *travel website* merupakan situs atau halaman-halaman yang menampilkan berbagai informasi mengenai hal-hal yang berhubungan dengan perjalanan yang dilakukan oleh seseorang ke suatu tempat seperti akomodasi, tiket penerbangan, tempat wisata dan lainnya.

TripAdvisor merupakan salah satu contoh *travel website* yang menggunakan kombinasi antara *online review* atau ulasan mengenai suatu perjalanan dengan pemesanan perjalanan. Menurut Miguens et al (2008), *TripAdvisor* merupakan situs yang menyediakan ulasan bagi wisatawan untuk merencanakan perjalanan dan juga membantu mereka dalam mengambil keputusan.

Dalam *TripAdvisor* orang-orang menuliskan komentar atau ulasan terhadap suatu hotel yang didukung juga dengan foto atau gambar (Miguens et al, 2008). Selain itu, Cunningham et al (2010) menyebutkan *"TripAdvisor reviews also featured prominently in the set of search results for the hotel name"*, yang artinya bahwa ulasan dalam *TripAdvisor* juga menonjol dalam pencarian untuk nama hotel.

2.6. Profil PT. Garuda Indonesia

PT Garuda Indonesia (Persero) atau biasa dikenal dengan Garuda Indonesia merupakan salah satu maskapai penerbangan terkemuka di Indonesia. Maskapai penerbangan ini pertama kali mengudara pada tahun 1940-an dalam era pendudukan Belanda. Pada saat itu maskapai masih bernama *Indonesian Airways* sejak 26 Januari 1949 dengan pesawat pertama-nya yang bernama Seulawah atau Gunung Emas. Pada awalnya Garuda Indonesia merupakan hasil kerjasama antara pemerintah Indonesia dengan Koninklijke Luchtvaart Maatschappij (KLM), yang merupakan maskapai Belanda yang kemudian semua sahamnya dimiliki oleh Indonesia pada tahun 1953. Pada tahun 1953, Garuda Indonesia telah berhasil memiliki 27 pesawat berserta staf-staf profesional.

Perkembangan penyedia jasa penerbangan Garuda Indonesia semakin meningkat. Pada tahun 1960-an, Garuda Indonesia mendatangkan tiga pesawat turboprop Lockheed L-188C Electra seiring dengan dibuka-nya rute penerbangan baru ke Hong Kong. Beberapa tahun kemudian, Garuda kembali mendatangkan tiga pesawat baru jenis Convair 990A yang merupakan pesawat yang memiliki kecepatan tinggi dengan teknologi canggih. Dengan pesawat baru ini, Garuda kembali membuka rute penerbangan dari Jakarta ke Amsterdam melewati Kolombo, Bombay, Roma, dan Praha. Tak berhenti sampai di sana, pada tahun 1966, Garuda kembali mendatangkan pesawat jet baru, yaitu Douglas DC-8 dan membeli beberapa pesawat turboprop baru, Fokker F27 guna melayani penerbangan domestik.

Pada awal tahun 1970-an Garuda kembali memperkuat armada-nya dengan membeli beberapa jenis narrow-body jet yaitu McDonnell-Douglas DC-9 dan Fokker F28 serta pesawat jenis turboprop Fokker F27 guna mendukung penerbangan domestik. Kemudian pada tahun 1973, guna memenuhi penerbangan internasional, seperti tujuan Eropa, Asia dan Australia, Garuda kembali mengirim pesawat McDonnell Douglas DC-10-30 dan Douglas DC-8. Selanjutnya untuk penerbangan ke Eropa dan Amerika Serikat Garuda mengoperasikan Boeing 747-2U3B baru-nya.

Pada tahun 1990-an terjadi bencana yang menimpa maskapai andalan Indonesia ini. Bencana pertama terjadi pada tanggal 13 Juni 1996 saat pesawat dari Fukuoka, Jepang menuju Jakarta. Awalnya saat pesawat hendak lepas landas, kipas turbin depan mesin pecah dan terpisah dari poros mesin sehingga mengakibatkan pesawat meledak dan terbakar saat kru mencoba menghentikan pesawat. Peristiwa ini menewaskan 3 dari 275 penumpang. Peristiwa lainnya terjadi pada tanggal 26 September 1997 saat pesawat Airbus A300-B4 yang jatuh di Desa Buah Nabar, Kecamatan Sibolangit, Kabupaten Deli Serdang, Sumatera Utara. Dalam peristiwa seluruh penumpang yang berjumlah 222 orang dan 12 awak tewas seketika. Ini merupakan kecelakaan pesawat terbesar dalam sejarah penerbangan Indonesia. Karena dua peristiwa tersebut membuat maskapai kesulitan ekonomi. Hal ini ditambah dengan dampak Krisis Finansial Asia yang sedang dialami Indonesia membuat Garuda sama sekali tidak melakukan penerbangan ke Eropa maupun Amerika. Untungnya, pada pertengahan tahun 2000 Garuda dapat mengatasi masalah keuangan-nya dengan baik.

Pada tahun 2000, Garuda membentuk anak perusahaan yang bernama Citilink yang menawarkan penerbangan dengan biaya murah ke kota-kota di Indonesia. Dengan adanya peristiwa-peristiwa nasional yang terjadi, seperti Serangan 11 September 2001, Bom Bali I dan Bom Bali II, wabah SARS, dan Bencana Tsunami Aceh 26 Desember 2004 serta peristiwa jatuhnya sebuah Boeing 737 di Yogyakarta berdampak masalah keuangan kembali terjadi di pihak Garuda. Hal ini diperparah dengan sanksi Uni Eropa yang melarang semua pesawat maskapai Indonesia menerbangi rute Eropa.

Setelah kembali menata krisis keuangan yang melanda Garuda. Garuda mulai mencatatkan sahamnya di Bursa Efek Indonesia sejak tanggal 11 Februari 2011. Selain itu, Garuda juga menjadi sponsor dalam pagelaran SEA Games 2011 yang digelar di Jakarta dan Palembang. Pada tahun 2012, Garuda Indonesia juga menjalin kerjasama dengan salah satu klub sepak bola Inggris, Liverpool FC sebagai Partner Resmi Liverpool FC dan Partner Maskapai Penerbangan Global Resmi Liverpool FC. Hingga saat ini Garuda Indonesia tetap menjadi pilihan utama konsumen Indonesia dalam penerbangan (Garuda Indonesia, 2016).

2.7. Penelitian Terdahulu

Penelitian terdahulu sebagai kajian bagi penulis sangat penting untuk megetahui hubungan antara penelitian yang dilakukan sebelumnya dengan penelitian yang penulis lakukan saat ini serta dapat menghindari adanya duplikasi. Hal ini bermanfaat untuk menunjukkan bahwa penelitian yang dilakukan mempunyai arti penting sehingga dapat diketahui kontribusi penelitian terhadap ilmu pengetahuan.

Penelitian mengenai *online review* sebelumnya sudah dilakukan oleh Gretzel et al. (2007). Hasil penelitian menunjukkan bahwa sebagian besar pengguna *Tripadvisor.com* yang disurvei sering bepergian untuk hiburan. Mereka tidak suka membuat keputusan spontan dan sebagian besar terlebih dahulu merencanakan perjalanan mereka. Untuk mengurangi resiko perjalanan yang tinggi mengenai pembelian terkait perjalanan, wisatawan atau pelancong harus mengumpulkan banyak informasi. Pada saat yang sama, informasi yang tepat tentang tujuan yang dipilih dapat meningkatkan kepercayaan para pelancong atau wisatawan selama proses pengambilan keputusan, membantu mereka untuk membuat keputusan terbaik mereka, oleh karena itu dapat meningkatkan kualitas perjalanan. Tentang perjalanan, mereka juga menyarankan agar meninjau situs seperti *TripAdvisor.com* yang memiliki sebuah keuntungan yang jelas karena memberikan banyak ulasan dan penilaian konsumen lainnya.

Penelitian tentang maskapai penerbangan Garuda Indonesia sebelumnya pernah dilakukan oleh Ginanjar (2009) dalam Tugas Akhirnya yang berjudul Analisis Kualitas Pelayanan Maskapai Garuda Indonesia Terhadap Kepuasan Penumpang. Dari hasil penelitian diketahui bahwa penilaian seluruh penumpang terhadap kualitas pelayanan maskapai Garuda Indonesia berdasarkan urutan yang paling diinginkan adalah *Pre Flight* (Layanan *call center*, Layanan *tiketing* di bandara, Sikap *staff check-in*, Waktu tunggu *check-in*, Ketepatan jadwal penerbangan) adalah baik. *In Flight* (Penampilan awak kabin, Sikap/keramahan awak kabin, Kebersihan ruang kabin, Kenyamanan tempat duduk, Bahan bacaan edisi terbaru, Penampilan interior kabin, Kualitas makanan dan minuman, Hiburan musik) adalah baik. *Post Flight* (Program *frequent flyer*, Sikap/keramahan *staff*

bagasi, Kecepatan menerima bagasi) adalah sangat baik. Untuk memperbaiki kepuasan penumpang, langkah awal yang harus diambil adalah mengurangi kesenjangan antara apa yang mereka harapkan dan apa yang mereka dapatkan. Oleh karena itu setiap perusahaan penerbang harus tahu siapa penumpangnya lalu mengukur kepuasan pelanggannya dengan berbagai elemen untuk menentukan kualitas pelayanan. Sehingga untuk mengukurnya diperlukan suatu studi penelitian yang komprehensif.

Dalam penelitian ini pengumpulan data dilakukan dengan menggunakan teknik *web scraping*. Penggunaan teknik *web scraping* ini bertujuan untuk memudahkan pengambilan data dalam skala besar dari sebuah *website* secara otomatis dan sudah dikenal secara luas. Beberapa penelitian yang terkait dengan implementasi *web scraping* diantaranya adalah penelitian yang dilakukan oleh Josi et al. (2014) menerapkan teknik *web scraping* pada aplikasi mesin pencari artikel ilmiah menggunakan bahasa pemrograman *PHP*. Penerapan teknik *web scraping* dilakukan pada sejumlah portal gratis diantaranya Portal Garuda, *Indonesian Scientific Journal Database (ISJD),* dan *Google Scholar*. Hasil *scraping* kemudian di tampung ke dalam tabel *MySQL* dan penelitian tersebut berhasil menyimpan data hasil *scraping* ke dalam *database*. Sedangkan penelitian yang dilakukan Dewantoro (2016) dalam Tugas Akhirnya menggunakan teknik *web scraping* pada proses *topic modeling.* Teknik *scraping* digunakan untuk mengambil artikel dari portal berita yang kemudian diolah menggunakan *topic modeling. Topic modeling* yang digunakan adalah *Latent Dirichlet Allocation* (*LDA*). Pada proses *scraping*, perancangan sistem dilakukan dengan identifikasi kelas *tag HTML*. 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 lain telah dilakukan oleh Ma'arif (2016), Ma'arif menggabungkan aplikasi *web scraping* bersama dengan analisis *text mining*, pada penelitian tersebut telah dikembangkan sebuah portal terintegrasi yang bisa secara otomatis mengambil informasi dari sebanyak mungkin halaman kemudian

menyajikannya kepada para calon wisatawan dalam bentuk yang lebih ringkas namun lengkap dan akurat. Penelitian yang telah melakukan Ma'arif meliputi proses pengambilan informasi *website* dengan teknologi *web scraping*, kemudian informasi pariwisata yang berhasil dikumpulkan dikelompokkan kedalam beberapa kategori secara otomatis menggunakan *text mining*.

Terdapat beberapa penelitian yang pernah dilakukan mengenai klasifikasi sentimen diantaranya adalah penelitian yang dilakukan oleh Kurniawan et al. (2012) yang berjudul Klasifikasi Konten Berita dengan Metode *Text Mining*. Kurniawan et al. (2012) menyatakan bahwa pengelompokan artikel berita di media *website* secara manual akan menjadi masalah apabila jumlah artikel berita yang akan dimuat di *website* dalam jumlah besar, maka akan 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*. Berita tersebut diklasifikasikan menjadi 4 kategori yaitu politik, ekonomi, olahraga, *entertainment* dengan masingmasing jumlah data uji dan data latih sebanyak 100 artikel.

Saraswati (2011) melakukan penelitian mengenai *text mining* menggunakan *metode naïve bayes classifier* (NBC) dan *support vector machine* (SVM) untuk sentimen analisis. Proses *text mining* meliputi kategorisasi *text*, *text clustering*, ekstraksi konsep/entitas, produksi taksonomi granular, *sentiment analysis*, penyimpulan dokumen, dan pemodelan relasi entitas. Hasil percobaan menunjukkan bahwa metode SVM mempunyai tingkat akurasi yang lebih baik daripada metode NBC untuk mengklasifikasikan opini berbahasa Inggris dan opini positif berbahasa Indonesia. Sedangkan NBC mempunyai tingkat akurasi yang lebih baik dalam mengklasifikasikan data uji opini negatif berbahasa Indonesia.

Penelitian yang dilakukan oleh Ulwan (2016) dalam Tugas Akhirnya yang berjudul *Pattern Recognition* Pada *Unstructured Data* Teks Menggunakan *Support Vector Machine* Dan *Association* (Studi Kasus: Portal Layanan Aspirasi dan Pengaduan Online Rakyat) dengan Menggunakan *Text Mining* metode *machine learning* yaitu *Support Vector Machine* (SVM), menyebutkan bahwa 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 dengan tingkat akurasi sebesar 96.7%. 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.

Pang et al. (2002) menggunakan *supervised machine learning* untuk mengklasifikasikan *movie reviews* dengan mengekstrak fitur yang berbeda dari *review film* tersebut dan menggunakan algoritma *machine learning Naive Bayes*, *Maximum Entropy* (ME) dan *Support Vector Machine* (SVM) dalam pengklasifikasikannya diperoleh akurasi antara 78,8 % dan 82,9 %.

Go et al. (2009) memanfaatkan *emoticon* dalam mempermudah pelabelan data pada analisis sentimen dari *tweet* berbahasa Inggris di *Twitter*. Go et al. (2009) mengklasifikasikan sentimen *tweet* atas 2 kelas yaitu kelas sentimen positif dan negatif. Akurasi yang diperoleh adalah 81,3 % dengan menggunakan *Naive Bayes* dan 80,5% dengan menggunakan *Maximum Entropy* serta 82,2% dengan menggunakan SVM untuk *unigram*. Go et al. (2009) menggunakan *emoticons* yang terkandung dalam *tweet* untuk kelas sentimen positif dan kelas sentimen negatif.

Hamzah (2012) melakukan penelitian tentang klasifikasi teks menggunakan *naive bayes classifier* untuk pengelompokan teks berita dan abstrak akademis. Penelitian ini mengkaji kinerja NBC untuk kategorisasi teks berita dan teks akademis. Penelitian menggunakan data 1000 dokumen berita dan 450 dokumen abstrak akademik. Hasil penelitian menunjukkan pada dokumen berita akurasi maksimal dicapai 91% sedangkan pada dokumen akademik 82%. Seleksi kata dengan minimal muncul pada 4 atau 5 dokumen memberikan akurasi yang paling tinggi.

Arifin (2016) dalam penelitiannya menyebutkan penentuan prioritas pemasangan internet untuk pelanggan baru dapat diimplementasikan dengan menggunakan algoritma *naive bayes*. Pengujian akurasi model dari sistem yang dikembangkan dengan menggunakan metode 10-*fold cross validation*

menghasilkan nilai akurasi sebesar 90% dengan sampel sebanyak 200 data. Sedangkan hasil pengujian akurasi model dari aplikasi *Rapidminer 7.1* menggunakan algoritma *support vector machine* dengan kernel *radial basis function* (RBF) diperoleh akurasi sebesar 88% dengan sampel sebanyak 200 data. Pengujian dengan data testing menghasilkan akurasi 88.89% dengan sampel sebanyak 18 data sehingga untuk program bantu dapat dilakukan.

Pada penelitian yang dilakukan oleh Ramadhani (2015), data yang telah diklasifikasi kemudian dihitung tingkat akurasi kebenarannya menggunakan metode *support vector machine* dengan menentukan fungsi kernel serta proporsi untuk data *training* dan data *testing* yang sesuai. Dari perbandingan nilai akurasi klasifikasi dengan menggunakan kedua metode diatas, didapatkan nilai akurasi paling tinggi adalah dengan menggunakan metode *naïve bayes classifier* (NBC) dengan akurasi sebesar 62,6295%. Dari 5412 dokumen *training* dan 2701 dokumen *testing* yang telah dilkasifikasi, didapatkan proporsi opini untuk kelas positif sebesar 44,46501% dan opini untuk kelas negatif sebesar 55,53499%.

Tabel 2.1 merupakan tabel rangkuman perbandingan dengan penelitian sebelumnya yang berkaitan *text mining* khususnya metode *naïve bayes*.

Tabel 2.1 *Perbandingan dengan penelitian terdahulu*

No.	Penulis	Judul	Metode	Persamaan	Perbedaan
1.	Kurniawan et. al. (2012)	Klasifikasi Konten Berita Dengan Metode *Text Mining*	*Text Mining* dan *Naïve Bayes Classifier*	Sama-sama menggunakan *supervised machine learning* dengan *Naive Bayes Classifier*	Penelitian terdahulu dilakukan untuk konten berita.
2.	Saraswati (2011)	*Naïve Bayes Classifier* Dan *Support Vector Machines* Untuk *Sentiment Analysis*	*Naïve Bayes Classifier (NBC)* dan *Support Vector Maching (SVM)*	Sama-sama menggunakan *supervised machine learning* dengan *Naive Bayes Classifier*	Penelitian terdahulu juga menggunakan metode *Support Vector Machine* (SVM) dan juga data berbahasa indonesia.
3.	Ulwan (2016)	*Pattern Recognition* Pada *Unstructured Data* Teks Menggunakan *Support Vector Machine* Dan *Association*	*Support Vector Machine* dan *Association*	Sama-sama melakukan klasifikasi data teks dari *website*	Penelitian terdahulu menggunakan metode *Support Vector Machine* (SVM).
4.	Pang et al. (2002)	*Thumbs up? Sentiment Classification using Machine Learning Techniques*	*Naive Bayes, Maximum Entropy* (ME) dan *Support Vector Machine* (SVM)	Sama-sama menggunakan *supervised machine learning* dengan *Naive Bayes*	Penelitian saat ini tidak menggunakan *Maximum Entropy* (ME) dan *Support Vector Machine* (SVM).
5.	Go et al. (2009)	*Twitter Sentiment Classification using Distant Supervision*	*Naive Bayes, Maximum Entropy* dan *Support Vector Machine*	Sama-sama menggunakan *supervised machine learning* dengan *Naive Bayes*	Penelitian saat ini tidak menggunakan *Maximum Entropy* dan *Support Vector Machine*.

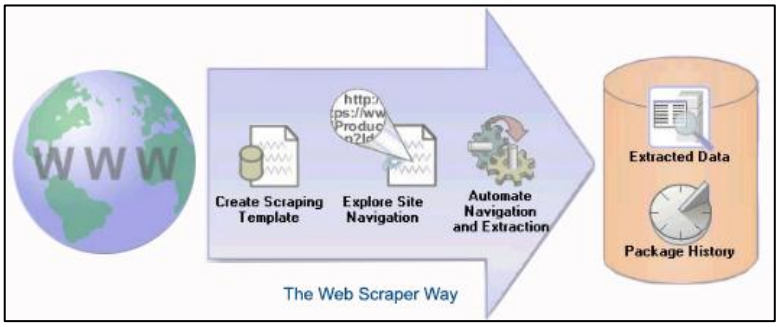
6.	Hamzah (2012)	Klasifikasi teks menggunakan *naive bayes classifier* untuk pengelompokan teks berita dan *abstract* akademis	*Naive Bayes Classifier* (NBC)	Sama-sama menggunakan *supervised machine learning* dengan *Naive Bayes Classifier*	Penelitian terdahulu dilakukan untuk teks berita dan abstrak akademis.
7.	Ramadhani (2015)	Analisis Sentimen Menggunakan Metode *Naïve Bayes Classifier* Dengan Model Dokumen Bernoulli Dan *Support Vector Machine*	*Naïve Bayes Classifier* dan *Support Vector Machine*	Sama-sama menggunakan *supervised machine learning* dengan *Naive Bayes Classifier*	Penelitian saat ini tidak menggunakan *Support Vector Machine*.
8.	Masithoh (2016)	Analisis Klasifikasi Topik Menggunakan Metode *Naïve Bayes Classifier, Naïve Bayes Multinomial Classifier*, Dan *Maximum Entropy* Pada Artikel Berita	*Naïve Bayes Classifier, Naïve Bayes Multinomial Classifier, Maximum Entropy*	Sama-sama menggunakan metode *Naive Bayes Classifier*	Penelitian saat ini tidak menggunakan *Naïve Bayes Multinomial Classifier* dan *Maximum Entropy*. Pada penelitian sebelumnya juga digunakan untuk data artikel berita.
9.	Rianto (2016)	Implementasi Dan Perbandingan Metode Prapemprosesan Pada Analisis Sentimen Gubernur DKI Jakarta Menggunakan Metode *Support Vector Machine* Dan *Naive Bayes*	*Support Vector Machine* Dan *Naive Bayes*	Sama-sama menggunakan metode *Naive Bayes*	Pada penelitian sebelumnya juga menggunakan *Support Vector Machine* dan mengimplementasikannya dengan membuat sebuah aplikasi untuk melakukan analisis klasifikasi.

BAB III
LANDASAN TEORI

3.1 *Web Scraping*

Web Scraping adalah proses pengambilan informasi dari *website* yang ada atau teknik penggalian informasi dari sebuah situs. *Web Scraping* menerapkan pengindeksan dengan cara menelusuri dokumen *HTML* dari *website* yang akan diambil informasinya untuk di *tag HTML* agar bisa mengapit informasi yang diambil untuk ditirukan pada aplikasi *web scraping* yang akan kita buat. Proses *Web scraping* dilakukan dengan cara mengambil sebuah dokumen semi-terstruktur seperti *HTML* atau *XHTML*. Selanjutnya dokumen tersebut di analisis dan kemudian data yang dibutuhkan diambil dari halaman tersebut untuk digunakan bagi kepentingan lain. *Web scraping* bukanlah *data mining* karena *data mining* adalah proses pengambilan informasi untuk memahami pola semantik atau tren dari sejumlah data yang besar *(big data)*. Aplikasi *web scraping* atau *intelligent, automated, or autonomous agent* fokus pada cara memperoleh data melalui pengambilan data (Turland, 2010). *Web scraping* memiliki sejumlah langkah sebagai berikut (Josi, dkk, 2014) :

1) Membuat *template scraping*: Proses ini melakukan observasi terhadap dokumen *HTML website* yang akan diambil informasinya atau dikenai *scraping*. Caranya adalah dengan melakukan *tag HTML* untuk mengapit informasi yang akan diambil.
2) Eksplorasi Navigasi Situs: Proses ini melakukanmenelusuri navigasi pada *website* yang akan diambil informasinya atau dikenai *scraping* untuk ditirukan pada aplikasi *web scraper* yang dibuat.
3) Mengotomatis Navigasi dan mengekstraksi informasi: Berdasarkan informasi yang didapat pada langkat 1 dan 2 di atas, aplikasi *web scraper* dibuat untuk mengotomatisasi pengambilan informasi dari *website* yang ditentukan.
4) Ekstraksi data dan menyimpan *history*: Informasi yang didapat dari langkah 3 disimpan dalam tabel atau *database*. Ilustrasi tahapan *web scraping* dapat dilihat pada **Gambar 3.1** berikut :



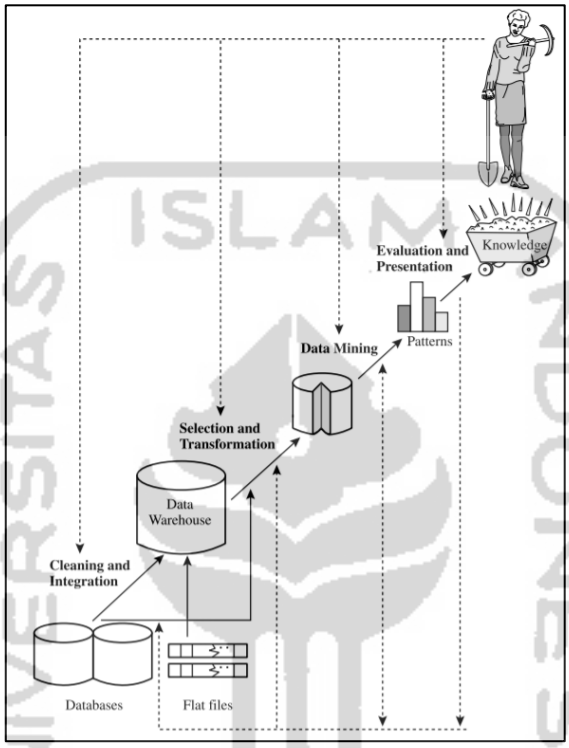
Gambar 3.1 *Ilustrasi cara kerja web scraping*

Pada **Gambar 3.1** dapat dilihat bahwa yang pertama kali perlu dilakukan adalah dengan membuat *template scraping*. Proses tersebut dilakukan dengan cara mempelajari dokumen *HTML* dari *website* yang akan diambil informasinya untuk di *tag HTML*-nya. Tujuan nya adalah untuk mengambil informasi. Setelah itu, proses berikutnya adalah dengan mengeksplorasi navigasi situs yang dikenai *scrapping*. Tujuannya adalah mempelajari teknik navigasi pada *website* yang akan diambil informasinya untuk ditirukan pada aplikasi *web scraping* yang dibuat. Selanjutnya, proses berikutnya adalah melakukan otomatisasi informasi yang didapat dari *website* yang telah ditentukan atau bias disebut juga sebagai proses ekstraksi informasi. Setelah informasi berhasil di ektraksi maka proses berikutnya adalah melakukan penyimpanan informasi ke dalam basis data (Esktraksi Data dan menyimpan *history*) (Juliasari, 2012).

3.2 *Data Mining*

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam *database*. *Data mining* adalah suatu proses yang menggunakan teknik statistik, matematika, kecerdasan tiruan, dan *machine-learning* untuk mengekstraksi serta mengidentifikasi informasi yang bermanfaat untuk pengetahuan yang terkait dari berbagai *database* besar (Turban et al., 2005). Menurut Tan et al. (2006) *data mining* adalah proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. *Data mining* juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan. Istilah *data mining* kadang disebut juga *knowledge discovery*. *Data mining* adalah bagian integral dari penemuan

pengetahuan dalam *database* (KDD) yang merupakan proses keseluruhan mengubah data mentah menjadi pola-pola data menarik yang merupakan informasi yang dibutuhkan oleh pengguna sebagai pengetahuan. Untuk mengetahui proses *knowledge discovery* dalam *database* bisa dilihat pada **Gambar 3.2** berikut :



Gambar 3.2 *Tahapan proses KDD (Han dan Kamber, 2006)*

Han dan Kamber (2006) menyatakan bahwa KDD terdiri dari langkah-langkah sebagai berikut:

1. *Data cleaning* adalah proses menghapus data yang tidak konsisten dan menghilangkan *noise*.
2. *Data integration* adalah proses menggabungkan data apabila memilki sumber data dalam sistem *data mining* tersebut.
3. *Data selection* adalah pengambilan data yang relevan yang akan digunakan dalam proses *data mining*.
4. *Data transformation* adalah proses dimana data ditransformasikan menjadi bentuk-bentuk yang sesuai untuk proses dalam *data mining*.
5. *Data mining* adalah suatu proses yang penting dengan melibatkan metode-metode untuk menghasilkan suatu pola data.

6. *Pattern evaluation* adalah proses untuk menguji kebenaran dari pola data yang mewakili *knowledge* yang ada didalam data itu sendiri.
7. *Knowledge representation* adalah proses visualisasi dan teknik menyajikan *knowledge* digunakan untuk menampilkan *knowledge* hasil *mining* kepada pengguna.

Ada empat tugas utama *data mining* yang terlihat pada **Gambar 3.3** diantaranya sebagai berikut:

1. *Predictive modelling*

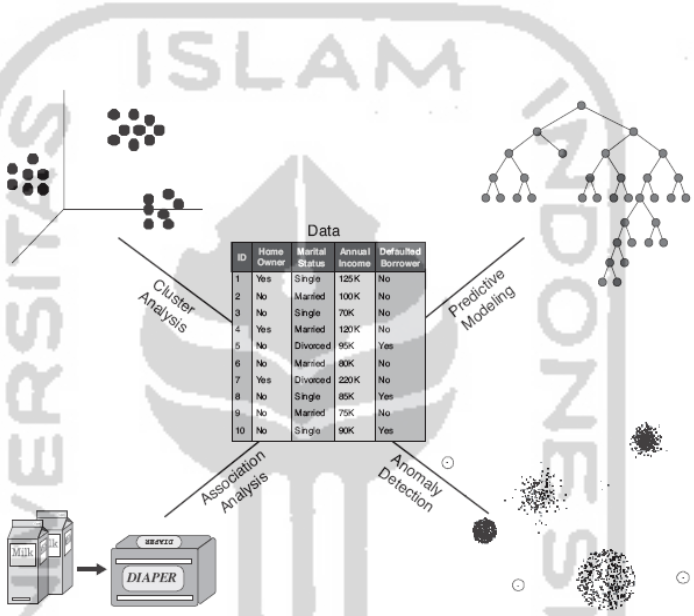
   Predictive modelling digunakan untuk membangun sebuah model untuk target variabel sebagai fungsi dari *explanatory* variabel. *Explanatory* variabel merupakan semua atribut yang digunakan untuk melakukan prediksi, sedangkan variabel target merupakan atribut yang akan diprediksi nilainya. *Predictive modelling* dibagi menjadi dua tipe yaitu *classification* yang digunakan untuk memprediksi nilai dari target variabel yang disktrit dan regresi yang digunakan untuk memprediksi nilai dari target variabel yang kontinu.
2. *Association analysis*

   Association analysis adalah penemuan *association rule* yang menunjukkan pola-pola yang sering muncul dalam data. Terdapat nilai *support* dan *confidence* yang dapat menunjukkan seberapa besar suatu *rule* dapat dipercaya. *Support* adalah ukuran dimana seberapa besar tingkat dominasi suatu *item* atau *itemset* terhadap keseluruhan transaksi. Sedangkan *confidence* adalah ukuran yang menunjukkan hubungan antara dua *item* secara *conditional*.
3. *Cluster analysis*

   Tidak seperti klasifikasi yang menganalisa kelas dan objek yang mengandung label. *Clustering* digunakan untuk menganalisa objek dari data tanpa memeriksa kelas label yang diketahui. Label-label kelas dilibatkan di dalam data *training* karena belum diketahui sebelumnya. *Clustering* merupakan proses mengelompokan sekumpulan objek yang sangat mirip.

4. *Anomaly detection/outlier mining*

   Sebuah *database* dapat mengandung data objek yang tidak sesuai atau menyimpang dari model data. Data objek ini disebut *outlier*. *Outlier mining* juga sering disebut dengan *anomaly detection* yang merupakan metode pendeteksian suatu data dimana tujuannya adalah menemukan objek yang berbeda dari sebagian besar objek lain. *Anomaly* dapat dideteksi dengan menggunakan uji statistik yang menerapkan model distribusi atau probabilitas untuk data.



Gambar 3.3 *Empat tugas utama data mining (Tan et al., 2006)*

3.3 *Machine Learning*

Istilah *machine learning* pertama kali didefinisikan oleh Arthur Samuel ditahun 1959. Menurut Arhtur Samuel, *machine learning* adalah salah satu bidang ilmu komputer yang memberikan kemampuan pembelajaran kepada komputer untuk mengetahui sesuatu tanpa pemrogram yang jelas. Menurut Mohri et al. (2012) *machine learning* dapat didefinisikan sebagai metode komputasi berdasarkan pengalaman untuk meningkatkan performa atau membuat prediksi yang akurat. Definisi pengalaman disini ialah informasi sebelumnya yang telah tersedia dan bisa dijadikan data pembelajar.

Dalam pembelajaran *machine learning*, terdapat beberapa skenario-skenario. Seperti:

1. *Supervised Learning*

   Penggunaan skenario *supervised learning*, pembelajaran menggunakan masukan data pembelajaran yang telah diberi label. Setelah itu membuat prediksi dari data yang telah diberi label.

2. *Unsupervised Learning*

   Penggunaan skenario *Unsupervised Learning*, pembelajaran menggunakan masukan data pembelajaran yang tidak diberi label. Setelah itu mencoba untuk mengelompokan data berdasarkan karakteristik-karakteristik yang ditemui.

3. *Reinforcement Learning*

   Pada skenario *reinforcement learning* fase pembelajaran dan tes saling dicampur. Untuk mengumpulkan informasi pembelajar secara aktif dengan berinteraksi ke lingkungan sehingga untuk mendapatkan balasan untuk setiap aksi dari pembelajar.

Saat ini telah banyak pendekatan *machine learning* yang digunakan untuk deteksi *spam*, *Optical character recognition* (OCR), pengenalan wajah, deteksi penipuan *online*, NER (*Named Entity Recognition*), *Part-of-Speech Tagger*.

3.4 *Natural Language Processing*

Natural Language Processing adalah bidang ilmu komputer dan linguistik berkaitan dengan interaksi antara komputer dan bahasa (alami) manusia (Kumar, 2011 dalam Putranti, 2013). Artinya pada sistem bahasa alami mencakup percakapan informasi dari basis data komputer ke dalam bahasa yang dapat dibaca manusia.

Sebuah *natural language system* harus memperhatikan pengetahuan terhadap bahasa itu sendiri, baik dari segi kata yang digunakan, bagaimana kata-kata tersebut digabung untuk menghasilkan suatu kalimat, apa arti dari sebuah kata, apa fungsi sebuah kata dalam sebuah kalimat dan sebagainya. Menurut Rich dan Knight (2006) dalam Putranti (2013) pengolahan bahasa alami mengenal beberapa tingkat pengolahan, yaitu :

1. Fonetik dan fonologi

   Fonetik dan fonologi berhubungan dengan suara yang menghasilkan kata yang dapat dikenali. Bidang ini menjadi penting dalam proses aplikasi yang memakai metode *speech based system.*

2. Morfologi

   Morfologi merupakan pengetahuan tentang kata dan bentuknya dimanfaatkan untuk membedakan satu kata dengan kata lainnya. Pada tingkat ini juga dapat dipisahkan antara kata dan elemen lain seperti tanda baca. Misalnya kata *going*: *going* (*word*) *go* (*root*) *ing* (*suffix*)

3. Sintaksis

   Sintaksis merupakan pemahaman tentang urutan kata dalam pembentukan kalimat dan hubungan antar kata tersebut dalam proses perubahan bentuk dari kalimat menjadi bentuk yang sistematis. Meliputi proses pengaturan tata letak suatu kata dalam kalimat akan membentuk kalimat yang dapat dikenali.

4. Semantik

   Semantik merupakan pemetaan bentuk struktur sintaks dengan memanfaatkan tiap kata ke dalam bentuk yang lebih mendasar dan tidak tergantung struktur kalimat. Semantik mempelajari arti suatu kata, dan bagaimana arti dari kata-kata tersebut membentuk suatu arti kalimat yang utuh. Dalam tingkatan ini belum tercakup konteks dari kalimat tersebut.

5. Pragmatik

   Pengetahuan pada tingkatan pragmatik berkaitan dengan masing-masing konteks yang berbeda tergantung pada situasi dan tujuan pembuatan sistem.

6. *Discourge Knowledge*

   Discourge knowledge melakukan pengenalan apakah suatu kalimat yang sudah dibaca dan dikenali sebelumya dapat mempengaruhi arti dari kalimat selanjutnya. Informasi ini penting diketahui untuk melakukan pengolahan arti terhadap kata ganti orang dan untuk mengartikan aspek sementara dari informasi.

7. *Word Knowledge*

   Word knowledge mencakup arti sebuah kata secara umum dan apakah ada arti khusus bagi suatu kata dalam suatu percakapan dalam konteks tertentu.

3.5 *Opinion Mining* atau *Sentiment Analysis*

Sentiment analysis disebut juga *opinion mining* adalah bidang studi yang menganalisis sentimen pendapat orang, evaluasi, penilaian, sikap dan emosi terhadap entitas seperti produk, jasa, organisasi, individu, masalah, topik, dan atributnya (Liu, 2012). Aktifitas penelitian di bidang analisis sentimen dan *opinion mining* sangat meningkat hal ini dikarenakan faktor pendukung meliputi (Pang, 2008) :

1. Munculnya metode *machine learning* dalam pengolahan bahasa alami dan pengambilan informasi.
2. Ketersediaan data set untuk algoritma *machine learning* untuk pelatihan pada *world wide web* khususnya, pengembangan *review-agregasi* situs *web*.
3. Realisasi tantangan intelektual menarik dengan aplikasi *intellegence* dan komersial yang ditawarkan.

Istilah analisis sentimen pertama kali muncul oleh Nasukawa dan Yi (2003), dan istilah *opinion mining* yang pertama muncul oleh Dave dkk. (2003). Nasukawa dan Yi (2003) pada *paper*-nya menggunakan teknik pembelajaran pengolahan alami pada *classifier sentiment online*. Dave dkk. (2003) dalam *paper*-nya memperkenalkan *opinion mining tool*, yang mengumpulkan pendapat tentang suatu topik tertentu mengklasifikasikan mereka sesuai dengan analisis subjektif. Hal ini dilakukan dengan mengidentifikasi sifat unik dari masalah dan mengembangkan metode yang secara otomatis membedakan antara *review* positif dan negatif.

3.5.1 Model *Opinion Mining*

Secara umum, pendapat dapat dinyatakan pada apa pun, misalnya, produk, layanan, topik, seorang individu, seorang organisasi, atau peristiwa. *Term* obyek istilah umum yang digunakan untuk menunjukkan entitas yang telah dikomentari. Sebuah objek memiliki seperangkat komponen (atau bagian) dan satu set atribut.

Masing-masing komponen mungkin juga memiliki sub-komponen dan set atribut, dan sebagainya. Dengan demikian, objek dapat diurai secara hierarkis berdasarkan pada *part of relationship* (Liu, 2010).

Kata *feature* digunakan untuk mewakili komponen dan atribut. Satu kalimat dapat mengungkapkan pendapat pada lebih dari satu fitur, misalnya adalah, “kualitas gambar dari kamera ini baik, tetapi daya tahan baterai yang singkat". Kualitas gambar dan daya tahan baterai merupakan obyek atau *feature* dari obyek yang diungkapkan pendapatnya pada kalimat diatas (Liu, 2010).

Sebuah dokumen positif pada obyek tidak berarti bahwa yang berpendapat (*opinion holder*) memiliki opini positif yang objektif tentang semua aspek atau *feature* dari objek. Demikian juga, sebuah dokumen yang negatif tidak berarti bahwa *opinion holder* tidak suka segala sesuatu tentang objek (Liu, 2010). Dalam dokumen evaluatif (misalnya *review* produk), *opinion holder* biasanya menulis aspek positif dan negatif dari objek, meskipun sentimen umum pada objek mungkin hanya bersifat positif atau negatif. Berdasarkan model ada tiga tugas utama *opinion mining* adalah:

1. Mengidentifikasi *feature object*: misalnya, dalam kalimat "kualitas gambar dari kamera ini menakjubkan", fitur objek "kualitas gambar".
2. Menentukan orientasi opini: tugas ini menentukan apakah opini yang objektif tentang fitur yang positif, negatif atau netral. Dalam kalimat di atas, pendapat tentang "kualitas gambar" feature ini memiliki orientasi opini positif.
3. Pengelompokan sinonim: sebagai *feature* objek yang sama dapat diungkapkan dengan kata-kata atau frasa yang berbeda.

3.5.2 Klasifikasi Sentimen

Permasalahan klasifikasi sentimen cukup banyak mendapat perhatian dalam riset akademis, dimana klasifikasi sentimen dan subjektifitas dapat diperlakukan sebagai permasalahan klasifikasi. Terdapat dua topik area yaitu (Liu, 2010) :

1. Mengklasifikasikan suatu dokumen opini atas opini positif dan negatif;

2. Mengklasifikasikan suatu kalimat/ klausa dari kalimat atas subjektif atau objektif dan klasifikasi kalimat/ klausa subjektif kalimat/ klausa dari kalimat yang mengekspresikan opini positif, negatif, ataukah netral.

Dalam banyak kasus, ditemukan istilah berbeda untuk dua topik area di atas, namun pada intinya adalah klasifikasi ini ingin menentukan orientasi dari suatu kalimat atau dokumen. Pembagian analisa sentimen menurut Liu (2010) di atas dapat juga dituliskan sebagai berikut :

1. Klasifikasi Sentimen / *coarse-grained sentiment analysis*

   Klasifikasi sentimen adalah suatu tahapan untuk menentukan apakah suatu dokumen memiliki ekspresi positif atau negatif. Klasifikasi sentimen ini juga dikenal sebagai *document level sentiment classification*.

2. Klasifikasi Subjektifitas / *fined-grained sentiment analysis*

   Klasifikasi subjektifitas yang juga dikenal sebagai *sentence level subjectivity classification* adalah suatu tahapan untuk menentukan apakah suatu kalimat bersifat subjektif atau objektif dan asosiasi opininya. Terdapat dua *task* dalam klasifikasi subjektifitas yaitu :

   a. Klasifikasi subjektifitas

      Menentukan apakah suatu kalimat adalah suatu kalimat subjektif ataukah kalimat objektif.

   b. *Sentence level sentiment classification*

      Jika kalimat tersebut adalah kalimat subjektif, ditentukan apakah kalimat tersebut mengekspresikan opini positif ataukah negatif.

Untuk klasifikasi sentimen terdapat beberapa metode. Secara umum, untuk menentukan klasifikasi sentimen ini digunakan *supervised method* dan *unsupervised method*. Dalam penelitian ini menggunakan *supervised method*.

Dalam *supervised method* dibutuhkan data pelatihan dan *testing* data. Dimana dua label kelas ditentukan terlebih dahulu (positif atau negatif). Klasifikasi sentimen dapat diformulasikan sebagai *supervised learning problem*. Berikut adalah beberapa fitur yang digunakan untuk *machine learning*:

1. *Terms* dan frekuensinya

   Cara ini sama dengan *information retrieval*, dengan merepresentasikan teks sebagai suatu vektor dimana *entry* cocok dengan *term* atau *n-gram*.

2. *Part of speech tags*

   Part-of-speech tagging adalah sebuah sistem yang memberikan label kata secara otomatis pada suatu kalimat. Misalkan, ada kalimat saya minum jamu dan ada label PRP=*personal pronoun*, VBT=*verb transitif*, NN=*common noun*. Sistem akan menerima *input* berupa kalimat tersebut, *output*-nya adalah: saya/PRP minum/VBT jamu/NN.

3. *Opinion words* dan *phrase*

   Opinion words adalah kata-kata yang biasanya digunakan untuk mengekspresikan sentimen positif atau negatif. Dalam implementasinya banyak memanfaatkan *POS Tagging* untuk mengekstraknya.

4. *Syntactic*

   Syntactic memainkan peranan yang penting dalam *opinion mining*, dimana *syntactic* digunakan untuk deteksi subjektifitas. Penelitian yang pernah dilakukan untuk menganalisis informasi sentimen seperti hubungan sintatik dan struktur dokumen adalah dengan membangun sebuah koleksi dari kalimat-kalimat yang didapat bersumber pada koleksi masif dari halaman *web* berbahasa Jepang, yang bertujuan untuk membangun sebuah sentimen *lexicon* (Kaji dan Kitsuregawa, 2007).

3.6 *Text Mining*

3.6.1 Pengertian *Text Mining*

Feldman dan Sanger (2007) menyatakan *text mining* adalah sebuah proses pengetahuan intensif dimana pengguna berinteraksi dan bekerja dengan sekumpulan dokumen dengan menggunakan beberapa alat analisis. Tujuan utama dari *text mining* ini adalah untuk mengekstrak informasi dari data berupa bahasa tekstual dengan tujuan tertentu. Jika dibandingkan dengan data yang tersimpan pada suatu basis data, *text* (data berupa data tekstual) merupakan data yang tidak

terstruktur dan sulit menemukan algoritma untuk menyelesaikannya. Namun, saat ini teks merupakan alat yang paling sering digunakan dalam pertukaran informasi.

Pendekatan manual *text mining* secara intensif dalam laboratorium pertama muncul pada pertengahan 1980-an, namun kemajuan teknologi telah memungkinkan ranah tersebut untuk berkembang selama dekade terakhir. *Text mining* adalah bidang *interdisipliner* yang mengacu pada pencarian informasi, *data mining*, *machine learning*, statistik, dan komputasi linguistik. Menurut Clara Bridge (2011) dalam Saraswati (2011) dikarenakan kebanyakan informasi (perkiraan umum mengatakan lebih dari 80%) saat ini disimpan sebagai teks, *text mining* diyakini memiliki potensi nilai komersial tinggi.

Perbedaan antara *text mining* dengan *data mining* terletak pada sumber data yang digunakan. Dalam *text mining* pola–pola yang diekstrak dari data tekstual yang tidak terstruktur bukan berasal dari suatu *database*. Beberapa kesamaannya adalah data yang digunakan merupakan data besar dan data berdimensi tinggi dengan struktur yang terus berubah. Dalam *data mining* data yang diolah adalah data yang terstruktur dari proses *warehousing* sehingga lebih mudah diproses oleh mesin/komputer. Analisis teks lebih sulit karena teks biasanya hanya digunakan sebagai konsumsi manusia secara langsung bukan digunakan untuk mesin/komputer. Ditambah struktur teks yang kompleks, struktur yang tidak lengkap, bahasa yang berbeda, dan arti yang tidak standar. Oleh sebab itu pada umumnya digunakan *Natural Languange Processing* untuk analisis teks yang tidak berstruktur tersebut.

3.6.2 Proses *Text Mining*

Data yang akan diolah dan menjadi input dalam *text mining* adalah berupa kumpulan teks/kumpulan dokumen yang sangat besar dan disebut sebagai *corpus*. Terdapat lima tahap proses pokok pada *text mining* menurut Even dan Zohar (2002) yang akan dijelaskan sebagai berikut:

1. *Text pre-processing*

   Pada tahap ini akan dilakukan proses mencari kebenaran arti dalam suatu *corpus*. Tujuan dari proses ini adalah untuk mempersiapkan data tekstual

yang tersaji agar dapat diolah. Pada tahap ini dilakukan *case folding* untuk membersihkan *corpus*. *Case folding* adalah proses dimana suatu kata diubah menjadi huruf kecil secara keseluruhan (dari a sampai dengan z), karakter lain akan dihapus dari corpus karena dianggap *delimiter*.

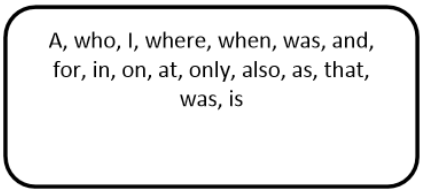
Selanjutnya setelah tahap *case folding*, dilakukan proses tokenisasi dan *stemming*. Tokenisasi adalah memisahkan kata per kata pada sebuah dokumen menjadi kata – kata yang saling independen.

Gambar 3.4 *Contoh proses tokenisasi berbahasa inggris*

Setelah tokeninasi selesai, dokumen akan melanjutkan tahap *stemming* yaitu proses merubah kata yang telah ditokenisasi menjadi kata dasarnya. Namun tahap *stemming* tidak sering digunakan karena mengakibatkan kerancuan dan menjadi tidak spesifik dalam merepresentasikan arti yang sebenarnya dari kata hasil *stemming*.

2. *Features Generation*

Setelah melakukan tahap tokenisasi dan *stemming* pada *Text Preprocessing*, selanjutnya dilakukan pengurangan dimensi kata di dalam *corpus* yang disebut *stopwords*. *Stopwords* merupakan tahap untuk menghilangkan kata – kata yang tidak berpengaruh / tidak informatif namun seringkali muncul dalam teks. Kata-kata *stopwords* akan didata didalam *stoplist*. Setiap bahasa mempunyai *stoplist* masing-masing.



Gambar 3.5 *Contoh stopwords berbahasa inggris*

Setelah melewati proses *stopwords*, tahap ini juga merupakan proses untuk mendapatkan representasi *corpus* yang diharapkan. Pendekatan

representasi *corpus* yang sering digunakan adalah model *bag-of-words*. Model *bag of words* akan merepresentasikan *corpus* menjadi kata perkata lalu menjumlahkan kata yang sama dalam *corpus* tersebut. Dalam *bag-of-words* representasi dari setiap kata diwakili oleh variabel terpisah yang memiliki besaran numerik. Cara menghitung besaran numerik yaitu dengan pembobotan.

Representasi *corpus* diidentifikasikan dalam bentuk matriks atau yang lebih dikenal sebagai *Term Documents Matrix*. *Term Documents Matrix* merepresentasikan kumpulan dokumen yang akan digunakan dalam proses klasifikasi dokumen. Pada *Term Documents Matrix*, sebuah dokumen direpresentasikan sebagai kumpulan fitur dan dapat diilustrasikan sebagai $D^i = [w_{i1} w_{i2} \dots w_{it} \dots w_{nk}]$ dimana D^i merupakan dokumen ke-i dan w_{it} adalah fitur yang digunakan pada kata ke-t yang terdapat dalam dokumen ke-i. Matriks ini akan diisi oleh nilai kemunculan dari suatu kata. Baris pada *Term Documents Matrix* merupakan data dokumen, sedangkan kolom dari *Term Documents Matrix* merupakan fitur yang digunakan. Jika ditulis dalam bentuk matriks :

$$\begin{matrix} dokumen\ ke-1 \\ dokumen\ ke-2 \\ \vdots \\ dokumen\ ke-n \end{matrix} \begin{bmatrix} w_{11} & w_{12} & \cdots & \cdots & w_{1k} \\ w_{21} & \cdots & \cdots & \cdots & w_{2k} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ w_{i1} & \cdots & \cdots & \cdots & w_{nk} \end{bmatrix} \quad \textbf{(3.1)}$$

untuk $i = 1,2,3, \dots, n$ dan $t = 1,2,3, \dots, k$. *Term Documents Matrix* dibuat untuk setiap label kelas yang akan diklasifikasi.

3. *Features Selection*

Tahap ini merupakan tahap lanjutan dari pengurangan dimensi. Walaupun di tahap sebelumnya sudah melakukan penghapusan kata–kata yang tidak deskriptif (*stopwords*), tidak semua kata–kata di dalam dokumen memiliki arti penting. Sehingga untuk mengurangi dimensi, pemilihan hanya dilakukan pada kata–kata yang relevan dan yang benar–benar mempresentasikan isi dari suatu dokumen. Kata-kata yang dinilai penting dilihat dari intensitas kemunculan dan yang paling informatif dari keseluruhan.

4. Analisis Teks

Dalam *text mining* terdapat dua cara yang umum dilakukan yaitu *clustering* dan klasifikasi. Klasifikasi membagi kumpulan dokumen tersebut menjadi dua *data set* yaitu *training data* dan *testing data* terlebih dahulu. Tujuan dari klasifikasi adalah menemukan sebuah model yang tidak terlihat dari *corpus*. *Training data* digunakan untuk membangun model dan *testing data* digunakan untuk validasi dari model tersebut.

3.6.3 Fitur dan Pembobotan

Pembobotan adalah metode yang mengubah *input* data menjadi suatu fitur vektor. Metode ini menggunakan *bag-of-feature* yang umum digunakan. Misalkan terdapat sederetan fitur seperti pada vector $\{f1, f2, \ldots, fn\}$ yang merupakan sekumpulan fitur-fitur sebanyak *n* yang sudah ditentukan sebelumnya. Misalkan kata "*excellent*" maka fitur vektor dari data adalah vektor.

a. Term Presence

Term Presence (TP) ialah metode pembobotan pada suatu dokumen teksyang melihat keberadaan daftar kata-kata (*term*) atau fitur yang ada pada *corpus* terhadap suatu dokumen. Jika suatu fitur yang ada pada daftar fitur acuan terdapat pada dokumen yang sedang diboboti maka nilai fitur tersebut pada *feature vector* akan diberi nilai 1 dan tidak menghiraukan jumlah kemunculan fitur tersebut. Jika fitur tersebut tidak ada pada dokumen maka diberi nilai 0 pada *feature space* (O'Keefe dan Koprinska, 2009). Rumus yang dipakai untuk menghitung *Term Presence* (TP) dari fitur t_i, pada dokumen d_j ditulis dengan notasi 3.2.

$$tp(t_i, d_j) = \begin{cases} 1 & Jika\ terdapat\ t_i \text{ pada } d_j \\ 0 & Jika\ tidak\ terdapat\ t_i \text{ pada } d_j \end{cases} \quad \textbf{(3.2)}$$

b. Term Frequency

Term Frequency (TF) memiliki kesamaan dengan TP yang sudah dijelaskan sebelumnya, tapi yang membedakan adalah TF menghitung jumlah kemunculan fitur acuan pada suatu dokumen bukan hanya keberadaan fitur

tersebut (O'Keefe dan Koprinska, 2009). Rumus TF dapat ditulis dalam persamaan 3.3 dengan $\#(ti, dj)$ mempunyai arti jumlah kemunculan fitur t_i pada dokumen d_j. Misalkan suatu fitur berupa kata "*good*" muncul sebanyak 10 kali maka nilai fitur tersebut pada *feature vector* adalah 10.

$$tf(ti, dj) = \#(ti, dj) \quad \textbf{(3.3)}$$

c. Term Frequency – Inverse Document Frequency

Term Frequency - Inverse Document Frequency (TF-IDF) adalah algoritma pembobotan yang disusun dari dua nilai yang berasal dari dua algoritma pembobotan yang berbeda, yaitu TF dan *Inverse Document Frequency* atau IDF. Rumus 3.4 menunjukkan formula perhitungan IDF pada suatu kumpulan dokumen D dengan |D| menujukkan jumlah dokumen dan $\#d(ti)$ menunjukkan banyaknya dokumen dimana suatu kata (ti).

$$idf\ (ti, dj) = log\frac{|D|}{\#d(ti)} \quad \textbf{(3.4)}$$

$$tfidf\ (ti, dj) = tf(ti, dj) \times idf\ (ti, dj) \quad \textbf{(3.5)}$$

Keluaran dengan fitur/*term* tersebut yang sering muncul pada dokumen akan menghasilkan nilai TF-IDF yang tinggi. Sementara, fitur yang sering muncul pada dokumen akan bernilai rendah. Dengan menggunakan metode ini fitur-fitur penting akan memiliki nilai yang tinggi dan fitur yang kurang penting akan memiliki nilai yang rendah (O'Keefe dan Koprinska, 2009).

3.7 Klasifikasi dengan *Naïve Bayes*

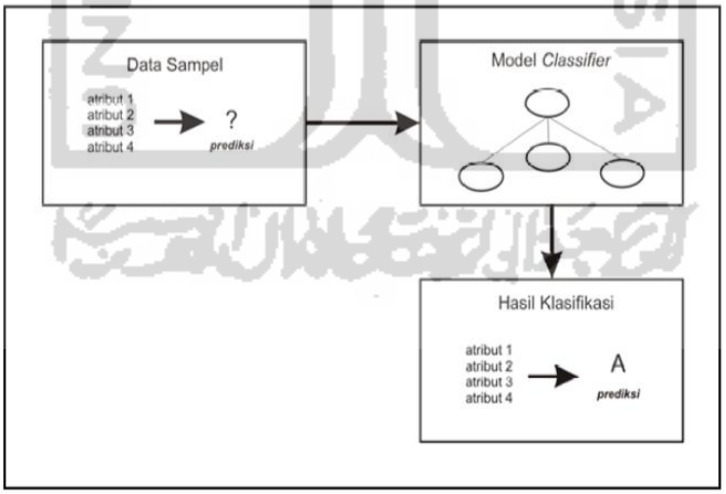
3.7.1 Klasifikasi

Klasifikasi merupakan suatu pekerjaan yang melakukan penilaian terhadap suatu objek data untuk masuk dalam suatu kelas tertentu dari sejumlah kelas yang tersedia (Prasetyo, 2012). Teknik klasifikasi merupakan teknik yang paling cocok untuk memprediksi atau menggambarkan kumpulan data dengan kategori biner atau nominal namun kurang efektif untuk kategori ordinal (misalkan, untuk

mengklasifikasikan seseorang dengan kriteria tinggi, menengah dan rendah untuk tingkat penghasilan) (Tan et al., 2006).

Menurut Han dan Kamber (2006) data *classification* memiliki dua tahap proses. Tahap pertama adalah membangun suatu model yang berdasarkan serangkaian data *class* yang disebut *learned model*. Model tersebut dibangun dengan menganalisa *record database*. Setiap *record* diasumsikan menjadi *predefined class* yang ditentukan oleh suatu atribut yang disebut *class* label atribut. Akibat terdapat *class* label maka tahap ini juga dikenal dengan *supervised learning*. Berbeda dengan *unsupervised learning* atau dikenal dengan *clustering* yang tidak memerlukan *class* label. Tahap pertama ini juga disebut sebagai tahap pembelajaran. Sebuah algoritma klasifikasi akan membangun sebuah model klasifikasi dengan cara menganalisis data *training*. Tahap pembelajaran dapat juga dipandang sebagai tahap pembentukan fungsi atau pemetaan $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ dimana y adalah kelas hasil prediksi dan x adalah *record* yang ingin diprediksi *class*-nya.

Bagan proses klasifikasi data sampel menggunakan model *classifier* untuk mendapatkan hasil prediksi (Han dan Kamber, 2006) dapat dilihat pada **Gambar 3.6** berikut :



Gambar 3.6 *Bagan proses klasifikasi (Han dan Kamber, 2006)*

Beberapa persiapan yang dilakukan untuk mendapatkan hasil klasifikasi yang baik diantaranya adalah (Han dan Kamber, 2006):

1. Pembersihan Data

   Pembersihan data ini dilakukan untuk mengurangi kecacatan data didalam data pelatihan, beberapa metode yang digunakan diataranya dengan teknik *smoothing* untuk menghilangkan noise data, melengkapi data yang hilang dan sebagainya.

2. Analisis Relevansi

   Dari beberapa atribut yang akan digunakan untuk proses klasifikasi mungkin saja terdapat atribut yang sangat berhubungan kuat satu sama lain, kedua atribut ini memiliki kemiripan sehingga menyebabkan proses klasifikasi menjadi tidak optimal, maka salah satu dari atribut ini dapat dibuang.

Hasil klasifikasi dan prediksi dapat dievaluasi menggunakan beberapa kriteria (Han dan Kamber 2006):

1. Akurasi

   Akurasi digunakan untuk mengetahui kemampuan model klasifikasi untuk dapat memberikan ketepatan hasil prediksi.

2. Kecepatan

   Mengetahui kecepatan iterasi untuk mendapatkan model klasifikasi dan iterasi mendapatkan hasil prediksi.

3.7.2 Teorema Bayes

Teorema bayes merupakan teorema yang mengacu pada konsep probabilitas bersyarat (Tan et al., 2006). Metode ini merupakan pendekatan statistik untuk melakukan inferensi induksi pada persoalan klasifikasi. Misalkan A dan B adalah kejadian dalam ruang sampel. Larose (2006) menyatakan probabilitas bersyarat dalam persamaan (3.6).

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \textbf{(3.6)}$$

Dimana P(A∩B) adalah probabilitas interaksi A dan B dan P(B) adalah probabilitas B. Demikian pula P(B|A) = $\frac{P(A \cap B)}{P(B)}$, sehingga nilai P(A∩B) = P (B|A) P(A). Nilai P(A∩B) kemudian disubtitusikan ke dalam persamaan (3.6), maka diperoleh persamaan (3.7).

$$P(A|B) = \frac{\mathrm{P(B|A)P(A)}}{\mathrm{P(B)}} \quad \textbf{(3.7)}$$

Misalkan θ merupakan parameter distribusi yang tidak diketahui. Larose (2006) menyatakan *posterior distribution* dalam persamaan (3.8) .

$$P(\theta|X) = \frac{\mathrm{P(X|\theta)P(\theta)}}{\mathrm{P(X)}} \quad \textbf{(3.8)}$$

Dimana P(X|θ) adalah fungsi *likelihood*, P(θ) merupakan *prior distribution* dan P(X) adalah normalisasi faktor yang disebut *marginal distribution* dari data.

Terminologi HMAP (*Hypotesis maximum a Posteriori probability*) menyatakan hipotesa yang diambil dari nilai probabilitas berdasarkan kondisi prior yang diketahui. HMAP adalah model penyederhanaan dari metode bayes yang disebut *naïve bayes*. Larose (2006) menyatkan persamaan bayes untuk HMAP pada persamaan (3.9) dan (3.10).

$$\theta_{map} = \mathrm{argmax}\ \theta\ \mathrm{P}(\theta|\mathrm{x}) = \mathrm{argmax}\ \frac{\mathrm{P}(\theta|\mathrm{x})\mathrm{P}(\theta)}{\mathrm{P(x)}} \quad \textbf{(3.9)}$$

Sehingga persamaan (3.8) dapat dituliskan menjadi persamaan (3.10) karena persamaan (3.9) merupakan argumen yang memaksimalkan nilai P(θ|X) atas semua θ.

$$\theta_{map} = \mathrm{argmax}\ \theta\ \mathrm{P}(X|\theta)\mathrm{P}(\theta) \quad \textbf{(3.10)}$$

Dalam hal ini :

X	=	data dengan *class* yang belum diketahui
θ	=	hipotesis data X merupakan suatu *class* spesifik
P(θ|X)	=	probabilitas hipotesis θ berdasar kondisi X (*Posteriori probability*)
P(θ)	=	probabilitas hipotesis θ (*Prior probability*)
P(X|θ)	=	probabilitas X berdasar kondisi pada hipotesis θ
P(X)	=	probabilitas dari X

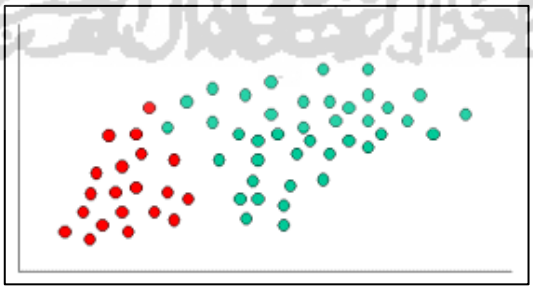
3.7.3 *Naïve Bayes Classifier*

Klasifikasi dibagi menjadi lima kelompok berdasarkan teori yang diadopsi atau teori yang menjadi dasar teknik klasifikasi. Lima pengelompokkan klasifikasi itu adalah *classifier Bayes's theorem*, *distancebased classifier*, *discriminant*

classifier, *neural networks classifier*, dan *decision tree classifier* (Shah, 2013). Pada penelitian ini akan fokus terhadap *classifier Bayes's theorem* atau algoritma klasifikasi yang mengadopsi teorema bayes yaitu *Naïve bayes classifier* .

Naïve bayes classifier merupakan metode klasifikasi menggunakan model probabilitas sederhana berdasarkan penggunaan teorema Bayes dengan asumsi independen yang kuat (*Naïve*). *Naïve Bayes classifier* mengasumsikan ada atau tidaknya suatu fitur tertentu pada sebuah kelas tidak mempengaruhi keberadaan fitur lainnya. Sebagai contoh, sebuah buah disimpulkan sebagai apel jika memiliki warna merah, berbentuk bulat bulat, dan berdiameter 4cm. Meskipun karakteristik tersebut bergantung satu sama lain, *Naïve Bayes classifier* memungkinkan semua karakteristik tersebut secara independen berkontribusi dalam probabilitas bahwa buah tersebut adalah buat apel. Atau dengan kata lain, tidak hanya buah apel yang memiliki karakteristik seperti yang disebutkan di atas.

Contoh sederhana dalam perhitungan *naïve bayes* misalnya terlihat pada **Gambar 3.7** terdapat dua kumpulan data yaitu hijau dan merah (Statsoft, 2015). Data baru akan ditambahkan dan akan ditentukan data baru tersebut merupakan bagian dari kelas yang mana. Karena jumlah data hijau dua kali lebih banyak daripada merah, maka diasumsikan bahwa data yang baru memiliki probabilitas menjadi anggota hijau dua kali lebih besar dari merah. Dalam analisis Bayesian, keyakinan ini dikenal sebagai probabilitas prior. Probabilitas prior didasarkan pada pengalaman sebelumnya, dalam hal ini persentase data hijau dan merah.



Gambar 3.7 *Dua kelompok data Naïve Bayes (Statsoft, 2015)*

$$Probabilitas\ prior\ untuk\ hijau = \frac{Jumlah\ data\ hijau}{Jumlah\ keseluruhan\ data}$$

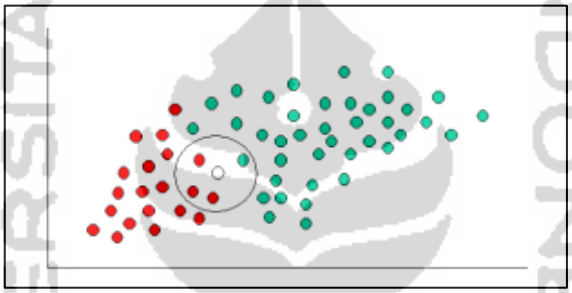
$$Probabilitas\ prior\ untuk\ merah = \frac{Jumlah\ data\ merah}{Jumlah\ keseluruhan\ data}$$

Dengan asumsi jumlah keseluruhan data yaitu 60, dengan 20 merah dan 40 hijau, maka probabilitas prior untuk keanggotaan kelas yaitu:

$$Probabilitas\ prior\ untuk\ hijau = \frac{40}{60}$$

$$Probabilitas\ prior\ untuk\ merah = \frac{20}{60}$$

Setelah menentukan probabilitas prior, objek baru (lingkaran putih) akan ditentukan keanggotaanya seperti terlihat pada **Gambar 3.8**. Diasumsikan bahwa semakin banyak suatu kelompok data tertentu (hijau atau merah) di sekitar objek baru tersebut, maka kemungkinan objek baru mempunyai keanggotaan sesuai kelompok data tersebut semakin besar.



Gambar 3.8 *Penambahan objek baru pada Naïve Bayes (Statsoft, 2015)*

Untuk mengukur kemungkinan tersebut, digambarkan lingkaran di sekitar objek baru X (putih), kemudian jumlah poin dalam lingkaran milik masing-masing label kelas akan dihitung. Dari sini didapatkan:

$$Kemungkinan\ objek\ baru\ adalah\ hijau = \frac{Jumlah\ hijau\ di\ sekitar\ X}{Jumlah\ data\ hijau}$$

$$Kemungkinan\ objek\ baru\ adalah\ merah = \frac{Jumlah\ hijau\ di\ sekitar\ X}{Jumlah\ data\ merah}$$

Sehingga,

$$Kemungkinan\ objek\ baru\ adalah\ hijau = \frac{1}{40}$$

$$Kemungkinan\ objek\ baru\ adalah\ merah = \frac{3}{20}$$

Meskipun probabilitas prior sebelumnya menunjukkan bahwa X mungkin merupakan anggota hijau (dimana jumlah hijau dua kali lebih banyak dibandingkan dengan merah) kemungkinan setelahnya menunjukkan hal yang sebaliknya; bahwa keanggotaan kelas X adalah merah (dimana terdapat lebih banyak merah di sekitar

X daripada hijau). Dalam analisis Bayesian, klasifikasi akhir yang dihasilkan adalah penggabungan kedua sumber informasi tersebut.

$$Probabilitas\ posterior\ untuk\ hijau = \frac{4}{6} \times \frac{1}{40} = \frac{4}{240} = \frac{1}{60}$$

$$Probabilitas\ posterior\ untuk\ merah = \frac{2}{6} \times \frac{3}{20} = \frac{6}{120} = \frac{1}{20}$$

Sehingga, dapat disimpulkan bahwa objek baru tersebut merupakan bagian dari kelompok data merah karena memilik probabilitas posterior merah yang lebih besar.

Berdasarkan sifat dari model probabilitas, *naïve Bayes classifier* dapat diterapkan dengan sangat efisien pada *supervised learning setting*. Secara aplikatif, estimasi parameter untuk model *naïve Bayes classifier* adalah dengan menggunakan metode *maximum likelihood*, dengan kata lain penggunaan model *naïve Bayes* dapat dilakukan dengan memahami terlebih dahulu mengenai aturan Bayesian.

a. Model Probabilitas untuk *Naïve Bayes*

Secara mendasar, model probabilitas untuk sebuah *classifier* adalah model peluang bersyarat.

$$P(C|F_1, \dots, F_n) \quad \textbf{(3.11)}$$

Dengan menggunakan teorema Bayes :

$$P(C|F_1, \dots, F_n) = \frac{P(C)P(F_1,F_2,\dots,F_n|C)}{P(F_1,F_2,\dots,F_n)} \quad \textbf{(3.12)}$$

Dimana Variabel C merepresentasikan kelas, sementara variabel *F1 ... Fn* merepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas C (Posterior) adalah peluang munculnya kelas C (sebelum masuknya sampel tersebut, seringkali disebut prior), dikali dengan peluang kemunculan karakteristik-karakteristik sampel pada kelas C (disebut juga likelihood), dibagi dengan peluang kemunculan karakteristik-karakteristik sampel secara global (disebut juga evidence). Karena itu, rumus di atas dapat pula ditulis secara sederhana sebagai berikut:

$$Posterior = \frac{Prior \times Likelihood}{Evidence} \quad \textbf{(3.13)}$$

Nilai Evidence selalu tetap untuk setiap kelas pada satu sampel. Nilai dari posterior tersebut nantinya akan dibandingkan dengan nilai-nilai posterior kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan.

Dalam prakteknya kita hanya tertarik pada pembilang dari persamaan tersebut, karena penyebut tidak tergantung pada C dan nilai-nilai fitur Fi diberikan, sehingga penyebut secara efektif bernilai konstan. Pembilang pada kasus di atas ekuivalen dengan model probabilitas bersama

$$P(C|F_1, \dots, F_n) \quad \textbf{(3.14)}$$

Persamaan (3.14) menurut Bishop (2005) dapat ditulis berdasarkan definisi peluang bersyarat :

$$\begin{aligned}
&P(C|F_1, \dots, F_n) \\
&= P(C)P(F_1, \dots, F_n|C) \\
&= P(C)P(F_1|C)P(F_2, \dots, F_n|C, F_1) \\
&= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3, \dots, F_n|C, F_1, F_2) \\
&= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2)P(F_4, \dots, F_n|C, F_1, F_2, F_3) \\
&= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2) \dots P(F_n|C, F_1, F_2, F_3, \dots, F_{n-1}) \quad \textbf{(3.15)}
\end{aligned}$$

Kondisi *naïve* mengasumsikan bahwa F_i independen terhadap F_j untuk setiap $j \neq i$. Mengartikan bahwa :

$$\begin{aligned}
P(F_i, F_j|C) &= \frac{P(F_i, F_j|C)}{P(C)} \\
&= \frac{P(F_i, C)}{P(C)} \cdot \frac{P(F_j, C)}{P(C)} \\
&= \frac{P(F_i|C)P(C)}{P(C)} \cdot \frac{P(F_j|C)P(C)}{P(C)} \\
&= P(F_i|C).P(F_j|C) \quad \textbf{(3.16)}
\end{aligned}$$

Untuk setiap $j \neq i$, sehingga untuk model peluang bersama dapat dibentuk menjadi :

$$P(C|F_1, \dots, F_n) = P(C)P(F_1|C)P(F_2|C)P(F_3|C) \dots P(F_n|C)$$
$$= P(C) \prod_{i=1}^{n} P(F_i|C) \quad \textbf{(3.17)}$$

dengan asumsi independensi, probabilitas bersyarat berdasarkan kelas C dengan mensubstitusikan persamaan (3.12) dan (3.17) menjadi :

$$P(C|F_1, \dots, F_n) = \frac{1}{z} P(C) \prod_{i=1}^{n} P(F_i|C) \quad \textbf{(3.18)}$$

dimana Z skala faktor dependen yang hanya terdapat pada $F_1, \dots, F_n$, merupakan sebuah konstan jika variabel di atas diketahui.

Bentuk dari model di atas sangat mudah untuk dikembangkan, P(C) dikatakan sebagai probabilitas prior dan $P = (F_i|C)$ sebagai probabilitas yang independen.

b. Klasifikasi Dokumen

Klasifikasi dokumen pada umumnya direpresentasikan sebagai *bag of words*. Secara sederhana menyebutkan kata apa yang ada pada sebuah dokumen dan seberapa sering kata tersebut keluar dalam satu dokumen. Menentukan sebuah dokumen D masuk pada kelas C adalah dengan melihat probabilitas posterior tertinggi, atau dapat ditulis berdasarkan definisi peluang bersyarat:

$$P(C|D) = \frac{P(D|C)P(C)}{P(D)} \propto P(D|C)P(C) \quad \textbf{(3.19)}$$

model di atas merepresentasikan dokumen menggunakan vektor fitur yang berisikan kosa kata yang terdapat dalam dokumen. Jika terdapat V kosa kata dan mengandung |V| kata, maka dimensi vektor tersebut adalah d = |V|.

a) Model Naïve Bayes

Model dokumen *Bernoulli* direpresentasikan dengan *vector biner*, dimana hanya memperhitungkan ada atau tidaknya sebuah kata pada suatu dokumen. Jika terdapat V kosa kata yang terdiri dari |V| kata, maka dimensi ke-t *vector* dokumen bergantung pada kata w_t yang terdapat dalam kosa kata pada dokumen. Diberikan variabel b_i sebagai fitur dari *vector* dokumen ke-*i* atau D^i, elemen ke-*t* dari b_i dinotasikan sebagai b_{it} bernilai biner dimana nilai

1 diberikan untuk kata yang terdapat dalam dokumen, sedangkan angka 0 diberikan untuk kata yang tidak terdapat dalam dokumen.

$P(W_t|C)$ dinotasikan sebagai peluang kata W_t pada sebuah dokumen berada di kelas C, sedangkan untuk kata yang tidak tergolong ke dalam kelas C memiliki probabilitas $(1 - P(W_t|C))$. Pada prakteknya nilai peluang dari suatu kata W_t masuk ke dalam kelas C dapat diestimasikan dari *training set* dengan menggunakan metode *Maximum Likelihood Estimation* sebagai berikut:

Probability Density Function untuk distribusi *bernoulli* adalah

$$P(x) = p^x(1-p)^{1-x} \quad \textbf{(3.20)}$$

Pada kasus *Bernoulli* untuk dokumen, *p* dinotasikan sebagai peluang kata w_t pada sebuah dokumen berada di kelas C atau ditulis sebagai $P(W_t|C)$ dan *x* sebagai b_i. Sehingga PDF untuk persamaan (3.20) ditulis sebagai :

$$P\left(D^i\middle|C\right) \sim P(b_i|C) = (P(w_t|C)^{b_{it}}(1 - P(w_t|C))^{1-b_{it}} \quad \textbf{(3.21)}$$

Model *Likelihood* untuk persamaan (3.21) adalah :

$$\prod_{t=1}^{|V|}(P(w_t|C)^{b_{it}}\,(1 - P(w_t|C))^{1-b_{it}} \quad \textbf{(3.22)}$$

dengan nilai peluang setiap kata dalam suatu dokumen adalah sama, maka didapatkan *log Likelihood* dari persamaan (3.22) adalah :

$$L\big(P(b_i|C)\big) = log \prod_{t=1}^{|V|}(P(w_t|C)^{b_{it}}\left(1 - P(w_t|C)\right)^{1-b_{it}}$$

$$= \log P(w_t|C)^{\sum_{t=1}^{|V|} b_{it}}\left(1 - P(w_t|C)\right)^{\sum_{t=1}^{|V|} 1-b_{it}}$$

$$= \left(\log P(w_t|C)^{\sum_{t=1}^{|V|} b_{it}} + \log\left(1 - P(w_t|C)\right)^{\sum_{t=1}^{|V|} 1-b_{it}}\right)$$

$$= \Big(\sum_{t=1}^{|V|} b_{it}\,(\log\big(P(w_t|C)\big) + \big(|V| - \sum_{t=1}^{|V|} b_{it}\big)(\log(1 - P(w_t|C))\Big) \quad \textbf{(3.23)}$$

Untuk memaksimumkan nilai pada persamaan (3.23) harus terpenuhi syarat kondisi pertama :

$$\frac{\partial L\left(P(b_i|C)\right)}{\partial P(w_i|C)} = 0$$

$$\frac{\partial}{\partial P(w_i|C)}\left(\sum_{t=1}^{|V|} b_{it}\,(\log\left(P(w_t|C)\right) + \left(|V| - \sum_{t=1}^{|V|} b_{it}\right)(\log(1 - P(w_t|C))\right) = 0$$

$$\frac{\sum_{t=1}^{|V|} b_{it}}{(P(w_i|C)} + \frac{|V| - \sum_{t=1}^{|V|} b_{it}}{(1 - P(w_t|C)}(-1) = 0$$

$$\sum_{t=1}^{|V|} b_{it}\,(1 - (P(w_t|C)) - \left(|V| - \sum_{t=1}^{|V|} b_{it}\right)P(w_t|C)) = 0$$

$$\sum_{t=1}^{|V|} b_{it} - \left(\sum_{t=1}^{|V|} b_{it}\right)(P(w_t|C) - |V|(P(w_t|C)) + \left(\sum_{t=1}^{|V|} b_{it}\right)P(w_t|C)) = 0$$

$$\sum_{t=1}^{|V|} b_{it} - |V|(P(w_t|C)) = 0$$

$$|V|(P(w_t|C)) = \sum_{t=1}^{|V|} b_{it}$$

$$P(w_t|C) = \frac{\sum_{t=1}^{|V|} b_{it}}{|V|} \qquad \textbf{(3.24)}$$

hasil (3.24) diduga sebagai MLE dari $P(w_t|C)$, untuk membuktikannya, maka harus terpenuhi kondisi kedua :

$$\frac{\partial L^2\left(P(b_i|C)\right)}{\partial P(w_t|C)^2} < 0$$

$$\frac{\partial L^2\left(P(b_i|C)\right)}{\partial P(w_t|C)^2} = \frac{\partial}{\partial\left(P(b_i|C)\right)}\left(\frac{\sum_{t=1}^{|V|} b_{it}}{P(w_t|C)} + \frac{|V| - \sum_{t=1}^{|V|} b_{it}}{\left(1 - P(w_t|C)\right)}(-1)\right)$$

$$= -\frac{\sum_{t=1}^{|V|} b_{it}}{P(w_t|C)^2} - \frac{|V| - \sum_{t=1}^{|V|} b_{it}}{\left(1 - P(w_t|C)\right)^2}(-1)$$

$$= -\frac{\sum_{t=1}^{|V|}\sum_{i=1}^{k} b_{it}}{P(w_t|C)^2} - \frac{|V| - \sum_{t=1}^{|V|} b_{it}}{\left(1 - P(w_t|C)\right)^2}(-1)\Bigg|_{P(w_t|C) = \frac{\sum_{t=1}^{|V|} b_{it}}{|V|}}$$

$$-\frac{\sum_{t=1}^{|V|} b_{it}}{\left(\frac{\sum_{t=1}^{|V|} b_{it}}{|V|}\right)^2} + \frac{|V| - \sum_{t=1}^{|V|} b_{it}}{\left(1 - \left(\frac{\sum_{t=1}^{|V|} b_{it}}{|V|}\right)\right)^2} < 0 \qquad \textbf{(3.25)}$$

dari persamaan (3.25) terbukti bahwa $\frac{\sum_{t=1}^{|V|} b_{it}}{|V|}$ merupakan MLE dari $P(w_t|C)$,

Model ini juga memiliki probabilitas prior pada masing – masing kelas atau $P(C = k)$. Parameter dapat diestimasi dengan menggunakan *training set* yang telah dikelompokkan sebelumya masuk ke dalam suatu kelas kelas $C = k$ dengan formula :

$$P(w_t|C = k) = \frac{n_k(w_t)}{N_k} \quad \textbf{(3.26)}$$

dimana $n_k(w_t)$ dinotasikan sebagai jumlah kata w_t pada sebuah dokumen termasuk dalam kelas $C = k$, dan N_k merupakan jumlah dokumen dalam suatu kelas.

Persamaan (3.26) memiliki permasalahan dimana jika $n_k(w_t)$ tidak pernah terjadi dalam dokumen latih maka kemunculan *term* tersebut bernilai nol dan menyebabkan peluang kejadian bersyarat akan menjadi nol pula. Hal ini dapat menyebabkan probabilitas *term* bernilai 0. Oleh karena itu diperlukan *smoothing* atau penambahan angka 1 pada pembilang dan jumlah *vocabulary* atau kosakata pada penyebut seperti pada persamaan 3.27. Variabel |V| atau kosakata merupakan jumlah kata unik yang terdapat pada seluruh dokumen.

$$P(w_t|C = k) = \frac{n_k(w_t)+1}{N_k+|V|} \quad \textbf{(3.27)}$$

Jika terdapat N dokumen pada total *training set* probabilitas prior kelas $C = k$ dapat diestimasi dengan melihat frekuensi *relative* dari dokumen pada kelas $C = k$:

$$P(C = k) = \frac{n_k}{N} \quad \textbf{(3.28)}$$

Untuk menentukan kelas yang sesuai suatu dokumen, probabilitas kata masuk ke dalam kelas C dinotasikan sebagai $P(w_t|C)$, dalam hal ini diasumsikan kata berdistribusi *random* dalam sebuah dokumen, kata tidak bergantung pada panjang dokumen dalam *training set*, posisi dokumen satu sama lain, dan konteks dokumen. Bishop (2005) mengatakan bahwa probabilitas dokumen D yang memuat kata w_t termasuk ke dalam kelas C didefinisikan sebagai :

$$P(D|C) = \prod_{t=1}^{|V|} P(w_t|C) \quad \textbf{(3.29)}$$

untuk mengetahui probabilitas dokumen D masuk ke dalam kelas C adalah:

$$P(C|D) = \frac{P(D \cap C)}{P(D)} \quad \textbf{(3.30)}$$

dengan demikian didapatkan formula untuk mencari probabilitas posterior untuk menentukan kelas dokumen adalah:

$$P(C|D) = P(C)P(D|C)$$

$$P(C|D) = P(C = k) \prod_{t=1}^{|V|} P(w_t|C = k) \quad \textbf{(3.31)}$$

Dengan menggunakan metode *Maximum a Posteriori Estimation* maka dokumen D akan tergolong ke dalam kelas C dengan melihat nilai :

$$C_{NB} = arg \max_p P(C|D)$$

$$C_{NB} = arg \max_p P(C = k) \prod_{t=1}^{|V|} P(w_t|C = k) \quad \textbf{(3.32)}$$

Algoritma klasifikasi sentimen menggunakan *Naïve Bayes Classifier* :

1. Menentukan jumlah kosa kata yang unik pada setiap dokumen.
2. Untuk setiap dokumen ditentukan nilai probabilitas prior $P(C = k) = \frac{n_k}{N}$ untuk masing – masing kelas yang telah diberi label sebelumnya.
3. Untuk setiap kata yang terdapat dalam *corpus* dicari nilai probabilitas $P(w_t|C = k) = \frac{n_k(w_t)+1}{N_k+|V|}$
4. Klasifikasi dokumen data *testing* dilakukan dengan melihat nilai C_{NB} pada persamaan (3.32) yang paling besar pada suatu kelas.

$$C_{NB} = arg \max_p P(C = k) \prod_{t=1}^{|V|} P(w_t|C = k)$$

5. Akurasi metode *naïve Bayes classifier* dihitung dengan membandingkan nilai pada tabel *confussion matrix*.

3.7.4 Simulasi *Naive Bayes Classifier*

Tahapan klasifikasi menggunakan *Naive Bayes Classifier* diawali dengan pembentukan bobot tiap *term* dari data ulasan menggunakan teknik *tf-idf*. Misalkan, terdapat 4 buah ulasan yang sudah melewati *text processing*.

1. *Good Sevice*
2. *Amazing flight service*
3. *Terrible flight experience*
4. *Terrible food*

Dari data ulasan tersebut kemudian dipilih kata-kata unik dari semua ulasan tersebut sehingga matriks yang tertuang pada tabel. 3.1.

Tabel 3.1 *Matriks term ulasan*

Ulasan	Good	Service	Amazing	Flight	Terrible	Experience	Food
1	1	1					
2		1	1	1			
3				1	1	1	
4					1		1

Dari matriks tersebut kemudian dihitung dengan menggunakan persamaan (3.5) sehingga setiap kemunculan kata dalam ulasan tersebut diubah menjadi nilai *tf-idf*. Berikut adalah contoh perhitungan *tf-idf* untuk kata "*good*" pada ulasan1 :

Jumlah dokumen = 4, *tf* ('*good*' pada ulasan1) = 1, *idf* (*good*) = 1

$$tfidf(\text{ ulasan1, '}good\text{' }) = 1 \times log(\frac{4}{1}) \quad = 0{,}602$$

Apabila proses perhitungan tersebut dilakukan untuk semua dokumen dan semua kata maka akan dihasilkan matriks perhitungan *tf-idf* seperti tabel 3.2 berikut:

Tabel 3.2 *Matriks tf-idf*

Ulasan	Good	Service	Amazing	Flight	Terrible	Experience	Food
1	0,602	0,301					
2		0,301	0,602	0,301			
3				0,301	0,301	0,602	
4					0,301		0,602

Kemudian hasil perhitungan *tf-idf* ini akan diolah dengan menggunakan algoritma *NBC* sehingga menghasilkan model probabilitas. Misalkan ulasan 1 dan 2 merupakan ulasan dengan kelas positif sedangkan ulasan 3 dan 4 merupakan ulasan dengan kelas negatif. Sehingga matriks *tf-idf* tersebut akan berubah menjadi seperti pada tabel 3.3.

Tabel 3.3 *Matriks tf-idf berdasarkan kelas*

	Good	Service	Amazing	Flight	Terrible	Experience	Food	Kelas
U1	0,602	0,301						Positif
U2		0,301	0,602	0,301				Positif
U3				0,301	0,301	0,602		Negatif
U4					0,301		0,602	Negatif

Berdasarkan tabel 3.3 diketahui kelas positif memiliki 2 data ulasan dengan jumlah kata sebanyak 5 kata dari 7 kosakata, sedangkan kelas negatif memiliki 2 data ulasan dengan jumlah kata sebanyak 5 kata dari 7 kosakata. Dari data tersebut, maka proses membangun probabilitasnya adalah sebagai berikut:

1. Kelas positif

   Tahap ini dilakukan dengan mengumpulkan seluruh kosakata yang muncul pada semua dokumen, kemudian dihitung nilai probabilitas dari masing masing kosakata menggunakan persamaan (3.27). Berikut merupakan contoh perhitungan nilai probabilitas kata "*good*" pada kelas positif :

$$P(w_t|C = k) = \frac{n_k(w_t) + 1}{N_k + |V|}$$

$$P(good\ |positif) = \frac{1+1}{5+7} = 0{,}167$$

   Keterangan :

   n_i = Jumlah "*good*" pada kelas positif adalah 1 buah

   n = Jumlah kata pada ulasan positif adalah 5 buah

   Kosakata = Jumlah kosakata adalah 7 buah

   Berdasarkan perhitungan tersebut, diperoleh nilai probabilitas kata "*good*" pada kelas positif adalah sebesar 0.167. Hasil perhitungan kata-kata lain pada kelas positif disajikan pada tabel 3.4.

Tabel 3.4 *Probabilitas kata dalam kelas positif*

Kosakata	Nilai
Good	0.167
Service	0.250
Amazing	0.167
Flight	0.167

Terrible	0.083
Experience	0.083
Food	0.083

2. Kelas negatif

   Dengan menggunakan teknik yang sama seperti pehitungan kelas positif diatas, maka hasil dari perhitungan probabilitas kata-kata kelas negatif disajikan tabel 3.5 berikut:

Tabel 3.5 *Probabilitas kata dalam kelas negatif*

Kosakata	Nilai
Good	0.083
Service	0.083
Amazing	0.083
Flight	0.167
Terrible	0.250
Experience	0.167
Food	0.167

Dengan demikian, data *training* tersebut membentuk model probabilitas yang dihasilkan oleh algoritma *NBC* yang dapat dilihat pada tabel 3.6.

Tabel 3.6 *Probabilitas kata berdasarkan kelas*

Kosakata	Positif	Negatif
Good	0.167	0.083
Service	0.250	0.083
Amazing	0.167	0.083
Flight	0.167	0.167
Terrible	0.083	0.250
Experience	0.083	0.167
Food	0.083	0.167

Setelah model probabilitas yang ada pada tabel 3.6 terbentuk, kemudian model tersebut disimpan dalam *database*. Setelah itu model tersebut akan diuji akurasinya dengan menggunakan data baru yang tidak diketahui kelasnya. Sebagai contoh, akan menguji ulasan baru yang isinya "*good sevice amazing experience*".

Untuk menentukan kelas ulasan baru tersebut, tahap yang pertama adalah memecah ulasan tersebut menjadi kata per kata, kemudian menggunakan persamaan (3.32) untuk menghitung probabilitas pada masing-masing kata untuk kelas tertentu. Kemudian, membandingkan hasil probabilitas tersebut dan probabilitas kelas. Probabilitas yang tertinggi merupakan prediksi kelas data yang baru tersebut. Untuk lebih jelasnya contoh "*good sevice amazing experience*" akan dicari prediksi kelasnya.

Tabel 3.7 *Model probabilitas data training*

Kelas	good	service	amazing	experience	Nilai probabilitas
Positif (P=0.5)	0.167	0.250	0.167	0.083	0.000289
Negatif (P=0.5)	0.083	0.083	0.083	0.167	0.000048

Pada tabel 3.7 terdapat kolom kelas dengan masing-masing P = 0,5. Nilai 0,5 merupakan peluang masing-masing kelas yang dihitung dengan persamaan (3.28). Sementara itu, pada kolom nilai probabilitas merupakan nilai probabilitas ulasan tersebut terhadap masing-masing kelas dimana nilai yang terbesar dari data tersebut yang merupakan hasil prediksinya. Dengan memperhatikan tabel 3.7, maka dapat disimpulkan bahwa ulasan "*good sevice amazing experience*" merupakan ulasan yang termasuk kedalam kelas positif karena nilai probabilitas kelas positif lebih tinggi dibandingkan dengan kelas negatif.

3.8 Metode Evaluasi Model Klasifikasi

Evaluasi digunakan untuk mengukur kinerja suatu sistem, khusus dalam penelitian ini digunakan untuk mengukur keakuratan metode klasifikasi dokumen teks. Salah satu teknik evaluasi model yang biasa digunakan untuk mengukur keakuratan metode klasifikasi dokumen teks adalah *confusion matrix*. Penelitian ini menggunakan metode *confusion matrix* dalam proses evaluasi dimana *confusion matrix* merupakan salah satu *tools* penting untuk menganalisis kinerja pengklasifikasi yang digunakan pada *mechine learning* yang biasanya memuat dua kategori atau lebih (Manning, dkk, 2009). Setiap unsur matriks menunjukkan

jumlah contoh data uji untuk kelas sebenarnya yang digambarkan dalam bentuk baris sedangkan kolom menggambarkan kelas yang diprediksi. 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*}. (Rianto, 2016).

Menurut Fawcett (2006), 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*. Hasil klasifikasi biner pada suatu dataset dapat direpresentasikan dengan matriks 2 x 2 yang disebut dengan *confusion matrix*. **Tabel 3.8** menunjukkan *confusion matrix* untuk dua kelas *classifier*.

Tabel 3.8 *Confusion matrix*

Predicted	***Actual***	
	Positive	*Negative*
Positive	*True Positive (TP)*	*False Positive (FP)*
Negative	*False Negative (FN)*	*True Negative (TN)*

Terdapat beberapa rumus umum yang dapat digunakan untuk menghitung performa klasifikasi. Hasil dari nilai akurasi, presisi, *recall* dan *f-measure* biasa ditampilkan dalam persentase.

a. *Accuracy*

Akurasi adalah jumlah proporsi prediksi yang benar. Adapun rumus penghitungan akurasi dapat dilihat pada persaman 3.33. (Lim dkk., 2006)

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad \textbf{(3.33)}$$

b. *Precision*

Precision adalah proporsi jumlah dokumen teks yang relevan terkenali diantara semua dokumen teks yang terpilih oleh sistem. Rumus *precision* dapat dilihat pada persamaan 3.34. (Lim dkk., 2006)

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad \textbf{(3.34)}$$

c. *Recall*

Recall adalah proporsi jumlah dokumen teks yang relevan terkenali diantara semua dokumen teks relevan yang ada pada koleksi. Rumus *recall* dapat dilihat pada persamaan 3.35. (Lim dkk., 2006)

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad \textbf{(3.35)}$$

d. *F-Measure*

F-Measure adalah nilai yang mewakili seluruh kinerja sistem yang merupakan rata-rata dari nilai *precision* dan *recall*. Rumus *f-measure* dapat dilihat pada persamaan 3.36. (Lim dkk., 2006)

$$f - measure = 2\frac{precision \times recall}{precision + recall} \quad \textbf{(3.36)}$$

e. *K-Fold Cross Validation*

Cross Validation adalah metode statistika untuk melakukan evaluasi dan membandingkan algoritma pembelajaran dengan cara membagi data menjadi dua bagian, yang pertama digunakan untuk model *training* dan yang lainnya digunakan untuk memvalidasi model. (Refaeilzadeh, dkk. 2008)

Cross Validation adalah metode umum digunakan untuk mengevaluasi kinerja *classifier*. Dalam pendekatan *cross validation*, setiap *record* digunakan beberapa kali dalam jumlah yang sama untuk *training* dan tepat sekali untuk *testing*. Metode ini mempartisi data ke dalam dua subset data yang berukuran sama. Pilih salah satu sebagai data *training* dan satu lagi untuk *testing*, kemudian dilakukan pertukaran fungsi dari subset sedemikian sehingga subset yang sebelumnya sebagai *training set* menjadi *test set* demikian sebelumnya. Pendekatan ini dinamakan *two-fold-cross-validation.* Metode *k-fold cross-*

validation menggeneralisasi pendekatan ini dengan mensegmentasi data ke dalam k partisi berukuran sama. Selama proses, salah satu dari partisi dipilih untuk *training*, sedangkan sisanya untuk percobaan *(testing)*. Prosedur ini diulangi k kali sedemikian sehingga setiap partisi digunakan untuk *testing* tepat satu kali (Tan et al., 2005, dikutip dalam Rianto, 2016).

K-fold cross validation merupakan salah satu metode yang digunakan untuk mengetahui rata-rata keberhasilan dari suatu sistem dengan cara melakukan perulangan dengan mengacak atribut masukan sehingga sistem tersebut teruji untuk beberapa atribut *input* yang acak. *K-fold cross validation* diawali dengan membagi data sejumlah *n-fold* yang diinginkan. Dalam proses *cross validation* data akan dibagi dalam n buah partisi dengan ukuran yang sama $d1,d2,\ldots,dn$, dan selanjutnya proses percobaan dan pembelajaran dilakukan sebanyak *n* kali. Dalam iterasi ke-i partisi $d(i)$ akan menjadi data percobaan dan sisanya akan menjadi data pembelajaran (Kohavi, 1995).

3.9 Asosiasi Teks

Dalam (Ulwan, 2016) disebutkan bahwa 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. Menurut Fadlisyah (2014) tujuan analisis korelasi adalah untuk mencari hubungan variabel bebas (X) dengan variabel terikat (Y), dengan ketentuan data memiliki syarat-syarat tertentu. Persamaan (3.37) adalah persamaan untuk menentukan nilai korelasi dan persamaan (3.38) adalah persamaan untuk mendapatkan *R Square*.

Penelitian ini menggunakan pendekatan asosiasi untuk menemukan hubungan antar komentar penumpang sehingga mendapatkan informasi yang dapat dijadikan bahan rujukan dalam pengembangan atau peningkatan kualitas perusahaan.

$$r = \frac{n.(\sum XY) - (\sum X).(\sum Y)}{\sqrt{(n.\sum X^2 - (\sum X)^2)(n.\sum Y^2 - (\sum Y)^2)}} \quad \textbf{(3.37)}$$

$$R = r^2 \quad \textbf{(3.38)}$$

3.9.1. Simulasi Perhitungan Asosiasi Teks

Proses perhitungan asosiasi yang digunakan adalah pendekatan nilai korelasi dengan terlebih dahulu mentransformasi data teks ke dalam bentuk *document term matriks*. Pada contoh kali ini yang digunakan adalah 5 kumpulan kata yaitu (Ulwan, 2016) :

```
kata1
kata1 kata2
kata1 kata2 kata3
kata1 kata2 kata3 kata4
kata1 kata2 kata3 kata4 kata5
```

Selanjutnya kumpulan kata tersebut dibuat dalam bentuk *document term matriks*.

Docs	kata1	kata2	kata3	kata4	kata5
1	1	0	0	0	0
2	1	1	0	0	0
3	1	1	1	0	0
4	1	1	1	1	0
5	1	1	1	1	1

Setelah terbentuk *document term matriks* selanjutnya dihitung dengan menggunakan rumus korelasi guna mendapatkan asosiasi antar kata, pada contoh kali ini akan dicoba untuk menghitung asosiasi kata2 dengan kata5.

Docs	kata2	kata5	kata2^2	kata5^2	kata2*kata5
1	0	0	0	0	0
2	1	0	1	0	0
3	1	0	1	0	0
4	1	0	1	0	0
5	1	1	1	1	1
Total	4	1	4	1	1

$$r = \frac{N\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{\{N\sum X_i^2 - (\sum X_i)^2\}\{N\sum Y_i^2 - (\sum Y_i)^2\}}}$$

$$r = \frac{(5*1) - (4*1)}{\sqrt{\{(5*4) - (4^2)\}\{(5*1) - (1^2)\}}}$$

$$r = \frac{1}{\sqrt{16}} = \frac{1}{4} = 0{,}25.$$

Jadi, didapatkan asosiasi kata2 dengan kata5 sebesar 0.25.

BAB IV
METODOLOGI PENELITIAN

4.1 Populasi dan Sampel

Populasi dalam penelitian ini adalah *database website TripAdvisor*, yaitu semua data ulasan tentang maskapai penerbangan Garuda Indonesia sejak bulan Januari 2016 sampai dengan bulan Maret 2017. Sedangkan sampel yang digunakan dalam penelitian ini adalah ulasan tentang maskapai penerbangan Garuda Indonesia yang berbahasa Inggris terhitung sejak bulan Januari 2016 sampai dengan bulan Maret 2017 yaitu total sebanyak 1143 ulasan.

4.2 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
Rating	Tingkat kepuasan pengunjung
Date	Tanggal dibuatnya komentar
Review	Isi komentar pengunjung
Flights	Rute penerbangan yang digunakan

4.3 Jenis dan Sumber Data

Jenis data yang digunakan dalam penelitian ini adalah data primer. Data tersebut diperoleh dengan teknik *web scraping* dari halaman situs *web TripAdvisor*® (www.tripadvisor.com). Data yang diperoleh merupakan data dari *database website TripAdvisor*, yaitu data yang berupa ulasan penumpang tentang maskapai Garuda Indonesia yang diambil sejak bulan Januari 2016 sampai dengan bulan Maret 2017 sebanyak 1143 ulasan.

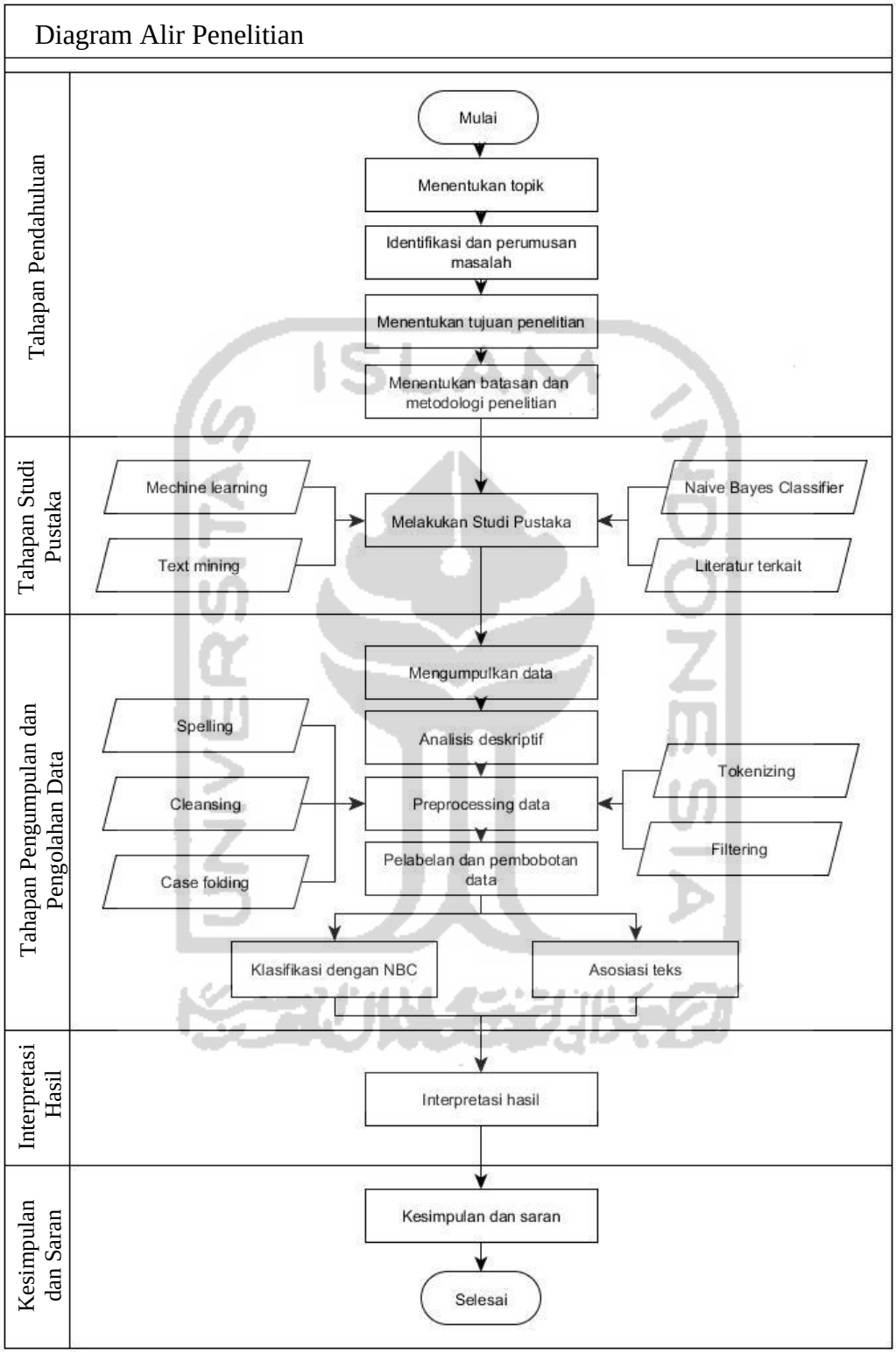
4.4 Metode Analisis Data

Software yang digunakan dalam penelitian ini adalah *Microsoft Excel 2016, R 3.2.2 64 bit* dan *R Studio*. Ada beberapa metode analisis data yang digunakan dalam penelitian ini, antara lain:

1. Analisis Deskriptif, digunakan untuk menggambarkan dan memetakan komentar yang terdapat dalam halaman *web TripAdvisor*.
2. Metode *Maching Learning* yaitu *Naïve Bayes Classifier*, digunakan untuk mengklasifikasikan ulasan yang berbentuk positif dan negatif.
3. *Association*, digunakan untuk mengidentifikasi dan membentuk pola kata yang berasosisasi dengan kata lainnya guna mendapatkan informasi yang dianggap penting dan berguna.

4.5 Tahapan Penelitian

Langkah atau tahapan yang dilakukan pada penelitian ini digambarkan melalui **Gambar 4.1** berikut ini:



Gambar 4.1 *Diagram alir penelitian*

BAB V
ANALISIS DAN PEMBAHASAN

5.1. Pengumpulan data dengan meggunakan teknik *Web Scraping*

Data dan informasi di *web* berkembang secara eksponensial. Hampir semua pengguna internet akan menggunakan sebuah mesin pencari bernama *Google* sebagai sumber pengetahuan pertama, entah itu untuk menemukan ulasan tentang sebuah tempat, untuk memahami sebuah istilah baru dan lain sebagainya. Mesin pencari ini yang nantinya akan mengarahkan ke beberapa situs atau *website* yang memuat informasi yang diinginkan. Semua informasi sudah tersedia dan dapat diakses melalui *web*.

Dengan jumlah data yang tersedia melalui *web*, ini akan membuka cakrawala baru bagi Ilmu Data. Salah satu teknik yang harus dikuasai oleh ilmuan data adalah teknik *web scraping*. Di dunia sekarang ini, hampir semua data yang dibutuhkan sudah tersedia di internet, satu-satunya hal yang membatasi untuk menggunakannya adalah kemampuan untuk mengaksesnya. Untuk itu didalam penelitian ini akan dijelaskan bagaimana cara memperoleh data dari internet dengan menggunakan teknik *web scraping*.

Sebelum proses *web scraping* dilakukan, beberapa *tools* yang harus dipersiapkan diantaranya sebagai berikut :

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

Pada penelitian ini penulis menggunakan sumber data dari salah satu situs pariwisata terbesar di dunia yaitu *TripAdvisor.com*. Dari situs tersebut akan diambil informasi yang dibutuhkan yaitu ulasan pengguna maskapai penerbangan Garuda Indonesia dalam bahasa inggris dengan menggunakan teknik *web scraping*. Data

yang akan di *scraping* tersebut memuat beberapa atribut berupa *id, quote, rating, date,* dan *review*.

Untuk melihat data ulasan maskapai penerbangan Garuda Indonesia pada halaman situs *TripAdvisor* dapat dilakukan dengan menuliskan *keyword* "Garuda Indonesia" pada kolom pencarian situs *TripAdvisor* menggunakan *web browser Google Chrome*, sehingga diperoleh halaman ulasan maskapai penerbangan Garuda Indonesia seperti terlihat pada gambar berikut :

Gambar 5.1 *Halaman review maskapai Garuda Indonesia pada situs tripadvisor*

Gambar 5.1 merupakan contoh tampilan halaman *TripAdvisor* untuk ulasan pengguna maskapai Garuda Indonesia dalam bahasa Inggris. Ulasan tersebut mengandung beberapa informasi berupa atribut diantaranya *quote, rating, date, review* dan *flights*. Untuk melakukan proses *scraping* dibutuhkan koneksi internet untuk menghubungkan *software R* dengan situs *TripAdvisor*. Adapun proses *scraping data* dilakukan dengan beberapa langkah sebagai berikut :

1. Tahap awal yang perlu dilakukan adalah menginstall *package* '`rvest`' pada *software R* yang berfungsi untuk membantu mengekstraksi data pada halaman *web* dengan cara menjalankan *script* `install.packages("rvest")`. Setelah *package* sukses terinstall pada *software R*, kemudian *package* tersebut dijalankan dengan perintah `library(rvest)`. Perintah "`library`" digunakan untuk mengaktifkan *package* yang ada pada *software R*.

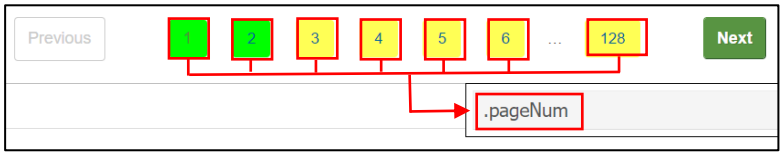
2. Melakukan proses *parsing* dokumen *html* atau pendefinisian *url* halaman *website* ulasan maskapai Garuda Indonesia pada situs *TripAdvisor*. Adapun *url* halaman ulasan maskapai Garuda Indonesia pada situs *TripAdvisor* adalah : https://www.tripadvisor.com/Airline_Review-d8729079-Reviews-Cheap-Flights-Garuda-Indonesia#REVIEWS.

   Untuk mendefinisikan *url* tersebut agar dapat dibaca pada *software R* maka dapat dilakukan dengan cara menjalankan *script* berikut :

```
url<-read_html("https://www.tripadvisor.com/
Airline_Review-d8729079-Reviews-Cheap-Flights-
Garuda-Indonesia#REVIEWS")
```

Gambar 5.2 *Script R untuk mendefinisikan URL website*

3. Situs *TripAdvisor* berisi banyak ulasan dari pengguna maskapai penerbangan udara Garuda Indonesia, namun pada setiap halaman situs hanya akan berisi 10 ulasan, sehingga data ulasan akan terbagi dalam beberapa halaman. Maka dari itu, untuk mendapatkan semua ulasan perlu dilakukan proses *looping* pada setiap halaman. Proses *looping* ini dimaksudkan agar program dapat berjalan lebih efektif, *scraping* data yang terdiri dari beberapa halaman tersebut dapat dijalankan secara otomatis tanpa harus melakukan *scraping* satu persatu dari halaman *website* tersebut. Namun, sebelum melakukan proses *looping*, terlebih dahulu perlu diketahui *nodes* atau *tag* yang menyatakan nomor halaman pada *website*. Untuk mengetahui letak *nodes* atau *tag* tersebut dapat digunakan *Add-extensions selector gadget* yang sebelumnya sudah terpasang pada *web browser google chrome*. Setelah *selector gadget* diaktifkan, untuk memperoleh kode dalam bahasa *CSS* maka dapat dilakukan dengan cara mengklik indeks halaman hingga diperoleh kode *CSS* yang menyatakan letak untuk nomor halaman. Dari hasil penyeleksian diperoleh kode CSS "`.pageNum`" yang menyatakan letak nomor halaman, seperti terlihat pada **Gambar 5.3** berikut :



Gambar 5.3 *Kode CSS letak nomor halaman*

4. Melakukan penghitungan atau pendefinisian jumlah halaman yang akan di *scraping* dan letak nomor halaman *website* kedalam 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. Melakukan pencarian indeks yang menyatakan halaman pada *url* situs *TripAdvisor*, untuk mengetahui indeks tersebut dapat dilihat dengan cara membandingkan *url* dari beberapa halaman yang memuat ulasan tentang maskapai Garuda Indonesia mulai dari halaman pertama dan seterusnya, seperti **Gambar 5.5** berikut :

```
URL Halaman Pertama
https://www.tripadvisor.com/Airline_Review-d8729079-Reviews-Cheap-Flights-Garuda-Indonesia#REVIEWS

URL Halaman Kedua
https://www.tripadvisor.com/Airline_Review-d8729079-Reviews-Cheap-Flights-or10-Garuda-Indonesia#REVIEWS

URL Halaman Ketiga
https://www.tripadvisor.com/Airline_Review-d8729079-Reviews-Cheap-Flights-or20-Garuda-Indonesia#REVIEWS
```

Gambar 5.5 *Mencari indeks nomor halaman pada URL Website*

Berdasarkan **Gambar 5.5** dapat diketahui bahwa pada setiap halamannya mempunyai *url* yang berbeda-beda dimana pada halaman pertama tidak terdapar indeks “-or-“ sedangkan pada halaman kedua dan ketiga berurutan terdapat indeks “-or10-” dan “-or20-”. Indeks tersebut menunjukkan bahwa untuk setiap penambahan satu nomor halaman, indeks angka akan bertambah sebanyak 10 dengan bubuhan “or” didepannya, dan setiap indeks halaman memiliki angka yang berada satu tingkat di bawah nomor halaman. Sehingga jika dikonfersikan kedalam Bahasa pemrograman *R* akan menjadi seperti **Gambar 5.6** berikut :

```
a<-0:(npages-1)
b<-10
res<-numeric(length=length(a))
for (i in seq_along(a)) {
res[i]<-a[i]*b
}
```

Gambar 5.6 *Script R untuk mendefinisikan indeks nomor halaman*

6. Membuat sebuah *template* dalam bentuk *data frame* atau tabel sebagai tempat untuk hasil data yang akan dilakukan *scraping* dengan nama "*tableout*" menggunakan perintah `tableout <- data.frame()`.
7. Melakukan proses *looping* dengan cara memberi inisial pada indeks angka halaman. Pada kasus ini indeks angka halaman dirubah menjadi "`,i,`" pada *url website* seperti **Gambar 5.7** berikut:

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

url <- paste ("https://www.tripadvisor.com/Airline_Review-
d8729079-Reviews-Cheap-Flights-or",i,"-Garuda-Indonesia#REVIEWS",sep="")
```

Gambar 5.7 *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 kembali digunakan *selector gadget*. Pada penelitian ini terdapat beberapa *tag* kelas *HTML* yang diperlukan, yaitu *tag HTML* untuk atribut *id*, *quote*, *rating*, *date*, *review* dan *flights* yang nanti akan digunakan untuk proses ekstrak informasi.

```
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()

flights <- reviews %>%
  html_node(".categoryLabel") %>%
  html_text()
```

Gambar 5.8 *Script R untuk mendefinisikan atribut yang akan di scraping dari halaman 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 kedalam 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, flights)
tableout <- rbind(tableout,temp.tableout)
}
```

Gambar 5.9 *Script R untuk menyusun data scraping kedalam 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, "D://Bismillah Skripsi/Data/Garuda.csv")
save.image()
```

Gambar 5.10 *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 *Contoh data hasil web scraping dari situs TripAdvisor*

no	*id*	*Quote*	*rating*	*date*	*reviewnospace*	*flights*
1	rn470805088	*What happen with Garuda*	4	3/28/2017	*Garuda is my number 1 choice to fly accross Indonesia, just in my laat two flight, the onboard service seemed to be shortened. There were only mineral water, tea, coffee, milk, and orange juice. It was more*	*Domestic*

					vary before.The food was soooo modest too, it was only rice with chicken.What happen with Garuda Indonesia?	
2	rn4707 31132	*Great airline*	4	3/27/ 2017	*Great airline, the service is exceptional by the crew.The seats could be a little more comfy for an overnight flight but it is only 5 and a half hours, great that blankets, pillows and water are waiting on your seat when you board.Good movie selection, but take you own headphones as the supplied ones are a little large...*	*Internati onal*
3	rn4704 34230	*Good service*	4	3/26/ 2017	*Good service, have space enough for leg..Perform good service, well maintainted, well steward/dess, clean and not too expensive.*	*Domestic*
4	rn4702 68810	*Best airlines I've flown ever...*	5	3/26/ 2017	*Don't know what else to say besides amazing flight, service attention to detail and more. And I only flew economy!! Prices are great too. I don't know how they do it. My girlfriend and I both said the best flight we've ever had!!*	*Internati onal*
5	rn4702 20569	*cheating flying time and a mess on terminal 3*	2	3/26/ 2017	*Its really the poor ground service and the wrong flying time.Jakarta Yogya is actually just a 35 min flight but Garuda makes it up to almost one hour, same all the other flights.Of course Garuda is always in time because the add 30% on the real flight time.*	*Domestic*

Setelah data diperoleh, proses selanjutnya adalah melakukan pelabelan data dan analisis klasifikasi dengan menggunakan metode NBC. Tabel 5.2 dan tabel 5.3 berikut merupakan penjelasan dari *script R* yang digunakan untuk melakukan pelabelan dan pengklasifikasian NBC.

Tabel 5.2 *Tahap-tahap pelabelan menggunakan software R*

Script R	Fungsi

`library(tm)`
`setwd("D://Bismillah Skripsi/Data/")`
`docs<-`
`  read.csv("Hasil_Cleaning.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("D://Bismillah Skripsi/SCRIPT/positive-words.txt",what="character",comment.char=";")`
`negatif <- scan("D://Bismillah Skripsi/SCRIPT/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(docs, kata.positif, kata.negatif, .progress='none')`
`{`
`require(plyr)`
`require(stringr)`
`scores = laply(docs, 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) - (1*sum(negatif.matches))`
`return(score)`
`}, kata.positif, kata.negatif, .progress=.progress )`
`scores.df = data.frame(score=scores, text=docs)`
`return(scores.df)`
`}`
	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 tabel

`hasil = score.sentiment(docs$text, kata.positif, kata.negatif)`
`View(hasil)`
 	
`#CONVERT SCORE TO SENTIMENT`
`hasil$klasifikasi<- ifelse(hasil$score<0, "Negatif","Positif")`
`hasil$klasifikasi`
`View(hasil)`
	6. Memanggil *function* hasil skoring yang telah dibuat
	7. Melakukan konversi nilai skor ke dalam kelas positif, negatif, dan netral

`#EXCHANGE ROW SEQUENCE`
`data <- hasil[c(3,1,2)]`
`View(data)`
`write.csv(data, file = "Label-english.csv")`
	8. Menyimpan file hasil pelabelan ke dalam format *csv.*

Tabel 5.3 *Tahap melakukan analisis NBC dengan software R*

Script R	Fungsi

`# Load Packages`
`library(tm)`
`library(e1071)`
`library(dplyr)`
`library(caret)`
`library(wordcloud)`
`library(RColorBrewer)`
`library(streamgraph)`
	1. Menjalankan *packages* yang sebelumnya telah terinstal pada program R

`# Import Data`
`df<- read.delim("clipboard", quote = "")`
`glimpse(df)`
	2. Mengimpor data kedalam program R dengan cara *copy-paste* dari *file excel*

`# Randomize dataset`
`set.seed(1)`
`df <- df[sample(nrow(df)), ]`
`glimpse(df)`
	3. Mengatur pegacakan pada *dataset*

`# preparing corpus`
`corpus <- Corpus(VectorSource(df$text))`
 	
`#data cleanup`
`corpus.clean <- corpus %>%`
`tm_map(content_transformer(tolower)) %>%`
`tm_map(removePunctuation) %>%`
`tm_map(removeWords, stopwords("english"))%>%`
`tm_map(stripWhitespace)`
 	
`# matrix representation`
`dtm <- DocumentTermMatrix(corpus.clean, control = list(weighting = weightTfIdf))`
	4. Proses pembentukan korpus
	5. Proses *cleaning* data
	6. Proses pembentukan *term* matriks

`# Split Data`
 	
`trainIndex <- createDataPartition(df$Polarity, p = 0.8, list = FALSE, times = 1)`
 	
`df.train <- df[trainIndex, ]`
`df.test <- df[-trainIndex, ]`
 	
`dtm.train <- dtm[trainIndex, ]`
`dtm.test <- dtm[-trainIndex, ]`
 	
`n_train<-length(df.train$Polarity)`
`n_test<-length(df.test$Polarity)`
 	
`corpus.clean.train <- corpus.clean[1:n_train]`
`corpus.clean.test <- corpus.clean[1:n_test]`
	7. Proses pembagian data *training* dan data *testing* serta membuat wadah untuk proses *training* dan *testing* data

`#feature selection`
 	
`term_control <- findFreqTerms(dtm.train, 1)`
	8. Membuat objek kelas *Document Term Matrix*

`dtm.train.nb <-`
`  DocumentTermMatrix(corpus.clean.train,`
`  control=list(dictionary = term_control))`
`dtm.test.nb <-`
`  DocumentTermMatrix(corpus.clean.test,`
`  control=list(dictionary = term_control))`
	9. Fungsi yang digunkan pada proses ini adalah `findFreqTerms` untuk mengidentifikasi kata-kata yang sering muncul lalu membatasi *Document Term Matrix* (DTM) hanya dengan kata-kata yang terdapat dalam kamus yang dibentuk dengan fungsi `findFreqTerms`

`# Function to convert the word frequencies to`
`  yes (presence) and no (absence) labels`
`convert_count <- function(x) {`
`  y <- ifelse(x > 0, 1,0)`
`  y <- factor(y, levels=c(0,1), labels=c("No",`
`  "Yes"))`
`  y`
`}`
	10. Melakukan pelabelan data dengan memperhatikan frekuensi kata

`# Apply the convert_count function to get final`
`  training and testing DTMs`
`trainNB <- apply(dtm.train.nb, 2,`
`  convert_count)`
`testNB <- apply(dtm.test.nb, 2, convert_count)`
`# Train the classifier`
`system.time( classifier <- naiveBayes(trainNB,`
`  df.train$Polarity, laplace = 1))`
 	
`# Use the NB classifier we built to make`
`  predictions on the test set.`
`system.time( pred <- predict(classifier,`
`  newdata=testNB) )`
 	
`# Create a truth table by tabulating the`
`  predicted class labels with the actual class`
`  labels`
`table("Predictions"= pred,  "Actual" =`
`  df.test$Polarity)`
	11. Melakukan *training* untuk mendapatkan model dengan algoritma NBC
	12. Menggunakan model data *training* untuk mengklasifikasikan data baru

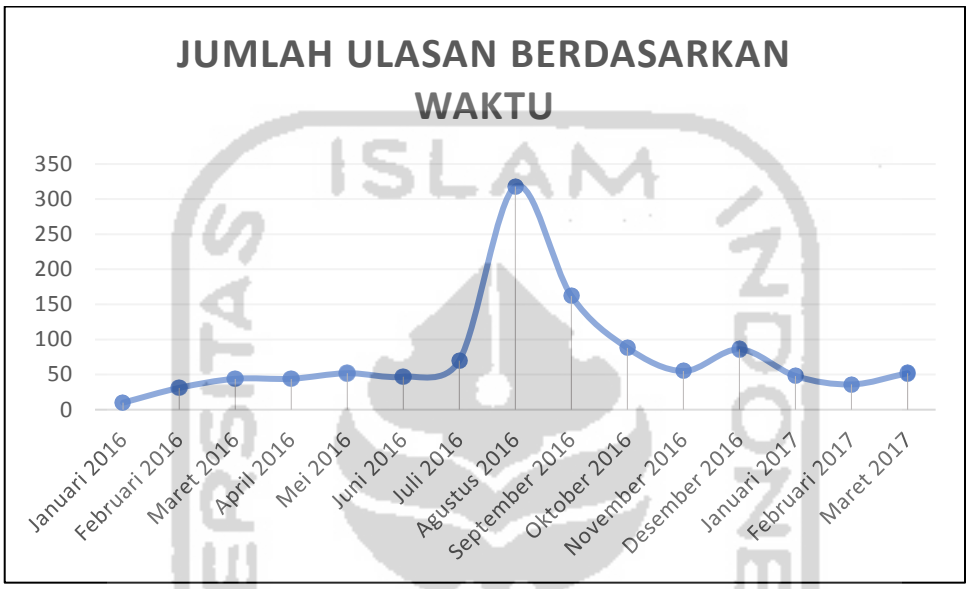
`# Prepare the confusion matrix`
`conf.mat <- confusionMatrix(pred,`
`  df.test$Polarity)`
`conf.mat`
 	
`conf.mat$byClass`
 	
`conf.mat$overall`
	13. Membuat tabel *confusion matrix*

`# Prediction Accuracy` `conf.mat$overall['Accuracy']`	14. Menghitung nilai akurasi hasil prediksi

5.2. Analisis Deskriptif

Analisis deskriptif dalam penelitian ini bertujuan untuk melihat gambaran secara umum informasi tentang maskapai Garuda Indonesia berdasarkan data

ulasan pengunjung dari situs *TripAdvisor* yang sebelumnya diperoleh dengan teknik *web scraping*. Dari data tersebut, secara umum dapat digambarkan beberapa aspek diantaranya jumlah ulasan yang masuk berdasarkan urutan waktu, jenis penerbangan dan *rating* maskapai Garuda Indonesia yang diberikan penumpang.

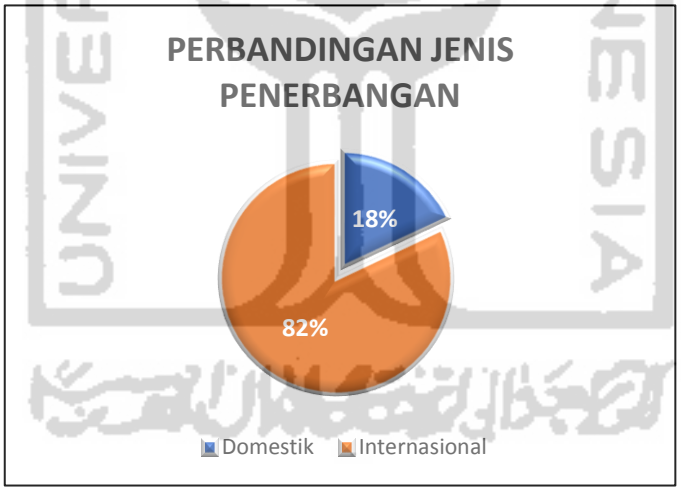


Gambar 5.11 *Grafik jumlah ulasan berbahasa inggris berdasarkan urutan waktu*

Gambar 5.11 menunjukkan grafik jumlah pengunjung situs *TripAdvisor* yang memberikan ulasan berbahasa inggris mengenai maskapai Garuda Indonesia terhitung sejak bulan Januari tahun 2016 hingga bulan Maret tahun 2017. Berdasarkan gambar tersebut dapat dilihat bahwa jumlah ulasan pada setiap bulannya cenderung mengalami fluktuasi, kenaikan jumlah ulasan meningkat secara signifikan pada bulan agustus tahun 2016 dengan jumlah ulasan yang masuk sebanyak 318 ulasan dan kembali menurun perlahan pada bulan-bulan selanjutnya. Kenaikan jumlah ulasan tersebut diduga karena bertepatan dengan jadwal pemberangkatan jamaah haji 2016. Garuda Indonesia telah memberangkatkan Calon Jemaah Haji Indonesia tahun 2016 (1437H) dari delapan embarkasi yaitu Banda Aceh, Medan, Padang, Jakarta, Solo, Balikpapan, Makassar dan Lombok, yang telah ditetapkan dan dalam dua tahap pelaksanaan, yaitu *phase* pertama (keberangkatan) mulai 9 Agustus 2016 serta *phase* kedua (pemulangan) mulai 17

September 2016. Pada musim Haji 2016 (1437 H) Garuda Indonesia telah menerbangkan 79.020 jamaah Indonesia yang tergabung dalam 205 kelompok terbang (kloter) dari delapan embarkasi (Garuda Indonesia, 2016).

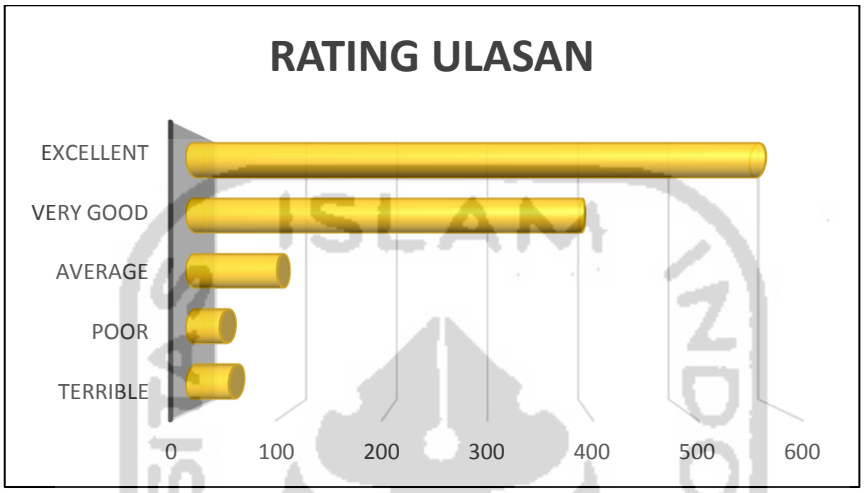
Seiring dengan meningkatnya penumpang atau penerbangan yang dilakukan maskapai Garuda Indonesia maka peluang meningkatnya jumlah ulasan mengenai maskapai Garuda Indonesia pada situs *TripAdvisor* juga akan semakin besar. Kemudian semakin banyaknya jumlah ulasan yang masuk pada situs *TripAdvisor*, maka akan semakin banyak pula informasi yang dapat diperoleh dari ulasan tersebut. Informasi berupa data ulasan penumpang tentu sangat bermanfaat bagi pihak maskapai Garuda Indonesia, karena dengan informasi tersebut pihak maskapai secara tidak langsung dapat mengetahui persepsi dan opini penumpang terhadap maskapai Garuda Indonesia, baik persepsi terhadap fasilitas, pelayanan, maupun kualitasnya sehingga dapat dijadikan sebagai kontrol dan bahan evaluasi ke arah yang lebih baik.



Gambar 5.12 *Grafik perbandingan jenis penerbangan*

Gambar 5.12 menunjukkan perbandingan jumlah jenis penerbangan yang dilakukan. Dari 1443 data ulasan, didapatkan bahwa pada ulasan berbahasa inggris tersebut didominasi oleh penumpang yang melakukan perjalanan internasional dimana 82% diantaranya adalah melakukan penerbangan internasional sedangkan sebanyak 18% melakukan penerbangan domestik (dalam negeri). Jumlah penerbangan internasional jauh lebih besar dari penerbangan domestik, hal ini

dikarenakan data yang digunakan adalah data ulasan berbahasa internasional (Bahasa Inggris). **Gambar 5.13** berikut merupakan *rating* yang menggambarkan penilaian diberikan penumpang terhadap maskapai Garuda Indonesia pada situs *TripAdvisor*:



Gambar 5.13 *Rating maskapai Garuda Indonesia berdasarkan ulasan penumpang pada situs TripAdvisor*

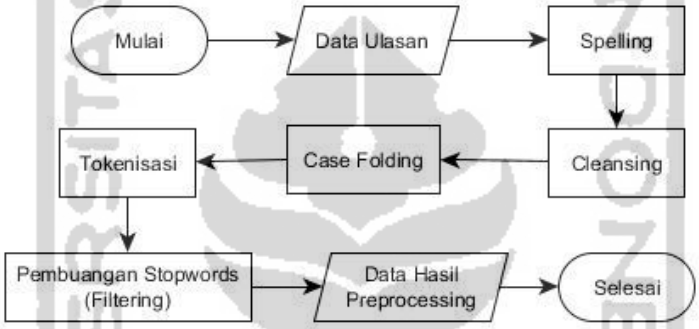
Rating pada situs *TripAdvisor* mempunyai skala 1-5 dengan kategori dari yang paling rendah ialah "*Terrible*" yang diberi skor "1", "*Poor*" dengan skor "2", "*Average*" dengan skor "3", "*Very Good*" dengan skor "4", dan "*Excelent*" dengan skor "5". Dari **Gambar 5.13** diatas dapat diketahui bahwa mayoritas penumpang maskapai Garuda Indonesia mempunyai penilain yang baik terhadap maskapai tersebut. Hal ini terbukti berdasarkan jumlah penilaian pengunjung dari 1143 ulasan, terdapat sebanyak 578 penumpang memberikan penilaian *Excelent,* 397 penumpang memberikan penilaian *Very Good,* dan 91 penumpang memberikan penilaian *Average,* sedangkan untuk penilaian buruk terhadap hanya berjumlah 34 pada kategori *Poor* dan 43 ulasan dengan kategori *Terrible.*

5.3. *Preprocessing* atau Prapemprosesan Data

Sebelum melakukan klasifikasi pada dokumen teks, perlu dilakukan *preprocessing*. Data ulasan yang diperoleh belum sepenuhnya siap digunakan untuk proses klasifikasi secara langsung karena data masih tidak terstruktur dengan baik dan terdapat banyak *noise*. Data masih memuat angka, tanda baca, *emoticon*, serta

kata-kata lain yang kurang bermakna untuk dijadikan fitur. Maka dari itu, perlu dilakukan *preprocessing* yang bertujuan untuk menyeragamkan bentuk kata, menghilangkan karakter-karakter selain huruf, dan mengurangi *volume* kosakata sehingga data akan lebih terstruktur.

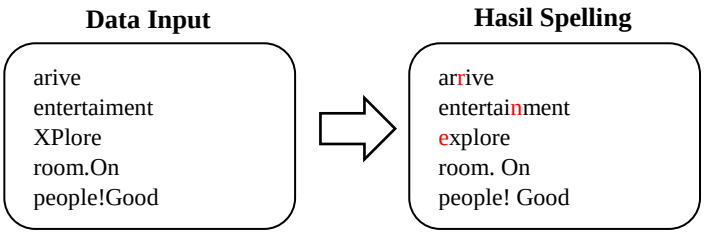
Dalam proses *preprocessing*, banyak tahap yang perlu dilalui diantaranya *spelling, cleansing, case folding, tokenizing* dan *filtering*. Pada tahap *spelling* akan dilakukan dengan bantuan *Microsoft Excel 2016* sedangkat pada tahap *cleansing, case folding, tokenizing* dan *filtering* akan dilakukan dengan bantuan *software R 3.3.2.* Alur dari *preprocessing* dapat dilihat pada gambar **Gambar 5.14** berikut :



Gambar 5.14 *Diagram alir preprocessing*

5.3.1. Spelling Normalization

Spelling adalah tahap awal yang perlu dilakukan untuk mendapatkan kualitas dokumen yang baik. Pada penelitian ini proses *spelling* dilakukan dengan bantuan *Microsoft Excel 2016*. *Spelling* merupakan fitur yang telah tersedia didalam *Microsoft Excel* yang sangat berguna dan *powerful* dalam menghasilkan dokumen yang baik. Fitur *spelling* ini dilengkapi dengan *dictionary* berbahasa inggris yang akan membantu dalam penyusunan dan kelengkapan huruf pembentuk kata / *frase* sehingga penggunaan kata dan tata bahasa kita menjadi teratur dan baik. Ini tentunya akan memberi efek kualitas dokumen kita, apakah layak diteruskan untuk dianalisis atau tidak. Contoh penerapan fitur *spelling* dapat dilihat pada **Gambar 5. 15** berikut :



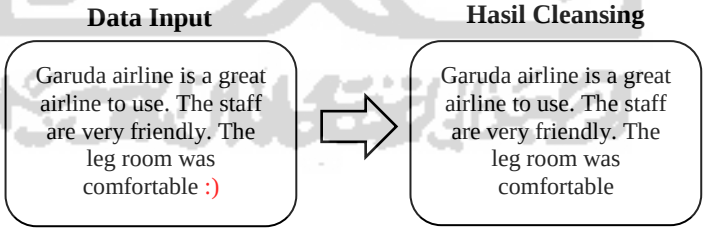
Gambar 5.15 *Proses Spelling*

5.3.2. Cleansing

Proses *cleansing* dilakukan untuk membersihkan dokumen dari kata-kata yang tidak berarti untuk mengurangi *noise* sehingga proses klasifikasi lebih efektif. Adapun katakata yang dihilangkan antara lain (Putranti, 2013) :

a. HTML karakter (<, >, dll)
b. Kata kunci pencarian (*blackberry*, *iphone*, dll)
c. Ikon emosi (“ :) ”, “ :-) ”, “ : D”,“ : (“, dan “ :- (“)
d. *Hashtag* (#)
e. *Username* (@username)
f. Alamat situs (url) (http://situs.com)
g. Alamat email (nama@situs.com)

Contoh penggunaan *case folding* dapat dilihat contoh pada **Gambar 5.16** berikut :

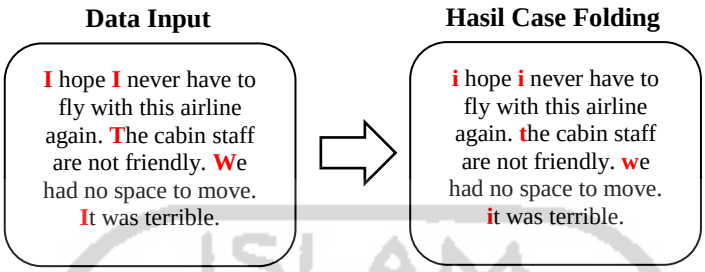


Gambar 5.16 *Proses cleansing*

5.3.3. Case Folding

Case folding adalah proses penyeragaman bentuk huruf dimana dalam proses ini hanya menerima huruf latin antara “a” sampai “z”. Karakter lain selain huruf dianggap sebagai delimiter sehingga karakter tersebut akan dihapus dari dokumen. Kemudian penyeragaman dilakukan dengan mengubah isi dokumen menjadi huruf kecil secara keseluruhan (dari “a” sampai dengan “z”). Hal ini

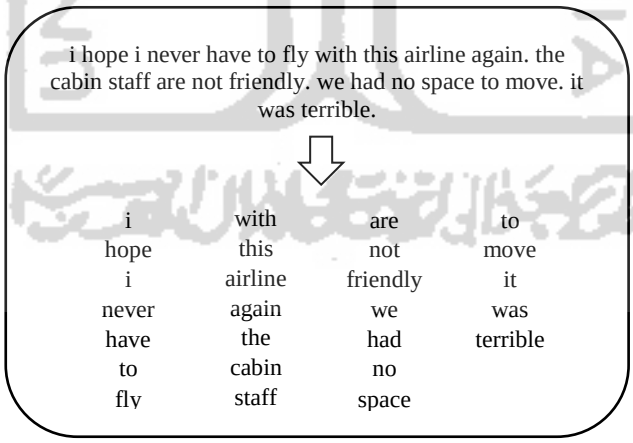
bertujuan agar kata yang ditulis dengan huruf awal *capital* dan huruf *non capital* tidak terdeteksi memiliki arti yang berbeda. Contoh hasil proses *case folding* dapat dilihat pada gambar **Gambar 5.17** berikut :



Gambar 5.17 *Proses case folding*

5.3.4. Tokenizing

Tokenizing atau tokenisasi adalah proses memisahkan kata per kata pada sebuah dokumen menjadi kata – kata yang saling independen. *Tokenizing* dilakukan untuk mendapatkan token atau potongan kata yang akan menjadi entitas yang memiliki nilai dalam penyusunan matriks dokumen pada proses selanjutnya. Tokenisasi dapat memudahkan proses perhitungan keberadaan kata tersebut dalam dokumen ataupun untuk menghitung frekuensi kemunculan kata tersebut dalam *corpus*. Contoh proses *tokenizing* ditunjukkan pada **Gambar 5.18** berikut :



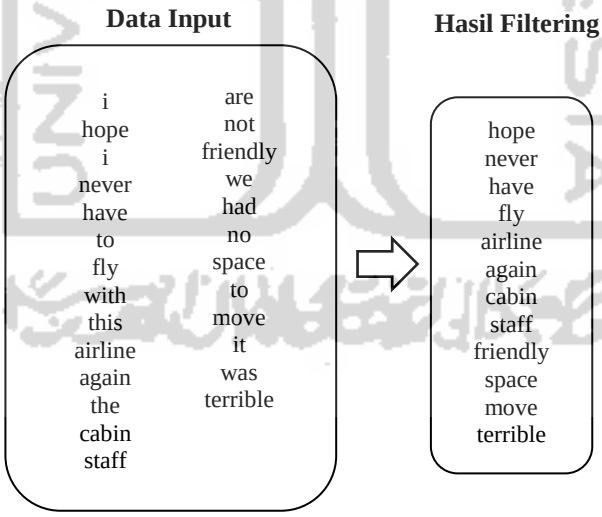
Gambar 5.18 *Proses tokenizing*

Setelah tokeninasi selesai, dokumen akan dilanjutkan ke tahap *stemming* yaitu proses merubah kata yang telah ditokenisasi menjadi kata dasarnya. Namun tahap *stemming* tidak sering digunakan karena mengakibatkan kerancuan dan menjadi tidak spesifik dalam merepresentasikan arti yang sebenarnya dari kata hasil

stemming. Oleh sebab itu, pada penelitian ini juga tidak akan menggunakan tahap *stemming*.

5.3.5. Filtering

Tahap penyaringan atau *filtering* merupakan tahap dilakukannya pemilihan kata pada dokumen atau pengurangan dimensi kata di dalam *corpus* yang disebut *stopwords*. Pada penelitian ini, *Stopwords* yang digunakan adalah *stopwords* yang telah disusun oleh (Bouge, Kevin., 2011) yang berisi sebanyak 571 kata. *Stopwords* merupakan tahap untuk menghilangkan kata-kata yang tidak berpengaruh / tidak informatif namun seringkali muncul dalam dokumen. Kata-kata tersebut seperti kata penghubung, kata ganti orang, kata seruan dan kata lainnya yang tidak begitu memiliki arti dalam penentuan kelas topik suatu dokumen. Kata-kata *stopwords* akan didata didalam *stoplist*. Setiap bahasa mempunyai *stoplist* masing-masing. Contoh isi *stoplist* dalam bahasa inggris adalah "i", "you", "and", "to", "it", "the", "was" dan lain-lain. Contoh proses *filtering* dapat dilihat pada **Gambar 5.19** berikut:



Gambar 5.19 *Proses filtering*

5.4. Pelabelan Kelas Sentimen

Setelah melewati proses *preprocessing*, selanjutnya akan dilakukan pelabelan kelas sentimen. Tahap ini juga merupakan salah satu proses untuk mendapatkan representasi *corpus* yang diharapkan. Pendekatan representasi *corpus*

yang sering digunakan adalah model *bag-of-words*. Model *bag-of-words* akan merepresentasikan *corpus* menjadi kata perkata lalu menjumlahkan kata yang sama dalam *corpus* tersebut. Dalam *bag-of-words* representasi dari setiap kata diwakili oleh variabel terpisah yang memiliki besaran numerik. Cara menghitung besaran numerik yaitu dengan pembobotan. Pembobotan yang digunakan adalah pembobotan otomatis berbasis kamus *(lexicon based)*. Dalam penelitian ini, kamus *lexicon* yang digunakan untuk pembobotan data adalah kamus yang disusun oleh (Hu and Liu, 2004) yang berisi 6800 kata. Pembobotan kata dilakukan dengan menghitung frekuensi kemunculan kata pada sebuah dokumen teks. Semakin sering sebuah kata muncul pada sebuah dokumen teks, maka bobot kata tersebut semakin besar dan kata tersebut dianggap sebagai kata yang sangat merepresentasikan dokumen teks tersebut (Yates dan Neto, 1999 dikutip dalam Basnur, 2009).

Pada umumnya, proses pelabelan dibagi kedalam tiga kelas sentimen, yaitu sentimen positif, negatif dan netral dengan cara skoring. Penilaian dokumen masuk ke dalam suatu kelas segmentasi positif atau negatif ditentukan dengan memanfaatkan kumpulan kata berbahasa Inggris yang terdiri dari *positive words* yaitu kumpulan kata-kata positif dan *negative words* yaitu kumpulan kata-kata negatif. Berdasarkan kumpulan kata berbahasa inggris tersebut kemudian akan dilakukan pelabelan otomatis oleh *program R* dengan cara menghitung skor jumlah kata positif dikurangi skor jumlah kata negatif dalam setiap kalimat ulasan. 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. Adapun hasil pelabelan kelas sentimen diperoleh perbandingan jumlah data seperti berikut :

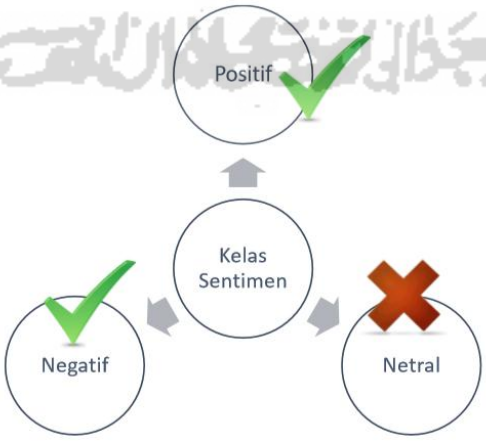
Tabel 5.4 *Perbandingan jumlah data pada kelas sentimen*

Sentimen Sementara	Jumlah Ulasan
Positif	911
Negatif	140
Netral	92

Klasifikasi data pada penelitian ini dibagi menjadi sentimen positif, negatif dan netral, namun data yang akan digunakan hanyalah data dengan sentimen positif dan negatif (Gambar 3.20). Hal ini dilakukan menimbang bahwa kelas sentimen netral dianggap kurang memberikan manfaat bagi pihak maskapai. Suatu ulasan diklasifikasikan sebagai sentimen positif bila mengandung pernyataan positif seperti pujian, ungkapan terima kasih, atau testimoni positif tentang maskapai penerbangan Garuda Indonesia. Suatu ulasan diklasifikasikan sebagai sentimen negatif bila mengandung pernyataan-pernyataan negatif seperti ketidakpuasan, penghinaan, laporan kegagalan layanan, dan sebagainya. Terakhir, sentimen netral adalah klasifikasi untuk ulasan yang diperoleh karena beberapa kemungkinan, diantaranya yaitu :

1. Dalam satu kalimat ulasan tidak terdapat kata sentimen yang teridentifikasi oleh kamus *lexicon*, baik sebagai kata positif maupun kata negatif.
2. Jumlah bobot antara kata positif dan kata negatif seimbang, sehingga jika dikalkulasikan akan menghasilkan skor yang bernilai 0.
3. Ulasan yang tidak mengandung sentimen, contohnya iklan, pertanyaan tanpa sentimen, dan sebagainya.

Sehingga dalam kasus ini, akan dilakukan reduksi kelas, yaitu dengan cara mengkategorikan kelas sentimen netral kedalam sentimen positif atau negatif yang dilakukan dengan cara manual.



Gambar 5.20 *Pembagian Kelas Sentimen*

Jika dalam kelas sentimen netral tidak teridentifikasi kata sentimen baik positif maupun negatif, maka akan diklasifikasikan ke dalam kelas positif. Sedangkan jika sentimen netral diidentifikasi memiliki bobot sentimen positif dan sentimen negatif yang seimbang, maka akan diklasifikasikan ke dalam kelas negatif. Hal ini dilakukan dengan pertimbangan bahwa informasi negatif dapat diekstraksi dengan lebih mudah, informasi negatif dapat juga diterjemahkan sebagai keluhan / ketidakpuasan pengguna, sehingga dengan mengetahui keluhan/ketidakpuasan tersebut diharapkan pihak Garuda Indonesia dapat melakukan evaluasi perbaikan ke arah yang lebih baik. Contoh hasil pelabelan data ulasan dapat dilihat pada **Tabel 5.5** berikut :

Tabel 5.5 *Hasil pelabelan menggunakan kamus lexicon dan proses manual*

Kelas Sentimen	Skor	Ulasan
Positif	11	*amazed boarded singapore aircraft brand works nice entertainment system spacious good legrrom good food good service good passenger announcement system totally recommend fly back booked garuda check fast efficient*
Negatif	-4	*ruin holidays flights delayed managed lose luggage arrival office people claims passengers provide incomplete number track luggage answer phone email provided lost luggage lied delivering luggage day hotel stay staff office compensation provide called insurance company day info trace luggage fly back ow ridiculously bad staff mismanage information helping*

5.4.1. Simulasi Perhitungan Skor Sentimen

Berdasarkan teks ulasan "*traveling garuda indonesia times* ***honestly*** ***hate*** *garuda indonesia* ***terrible*** *service past time admit developing service quality better* ***smiles*** *crews sincere* ***greet***", terdapat 3 kata poasitif dan 2 kata negatif yang terdeteksi pada kamus *lexicon*, yakni *"honestly", "smiles",* dan *"greet"* sebagai kata positif, kemudian *"hate"* dan *"terrible"* sebagai kata postif. Adapun rumus yang digunakan dalam proses perhitungan skor sentimen adalah sebagai berikut:

Skor = (Jumlah kata positif) – (Jumlah kata negatif) **(5.1)**

Tabel 5.6 *Simulasi perhitungan skor sentimen*

Teks ulasan	Kata positif	Kata negatif
traveling garuda indonesia times **honestly** **hate** garuda indonesia **terrible** service past time admit developing service quality better **smiles** crews sincere **greet**	**honestly** **smiles** **greet**	**hate** **terrible**
Jumlah	**3**	**2**

Sehingga dengan demikian diperoleh perhitungan sebagai berikut :

Skor = (Jumlah kata positif) – (Jumlah kata negatif)

Skor = 3−2

Skor = 1

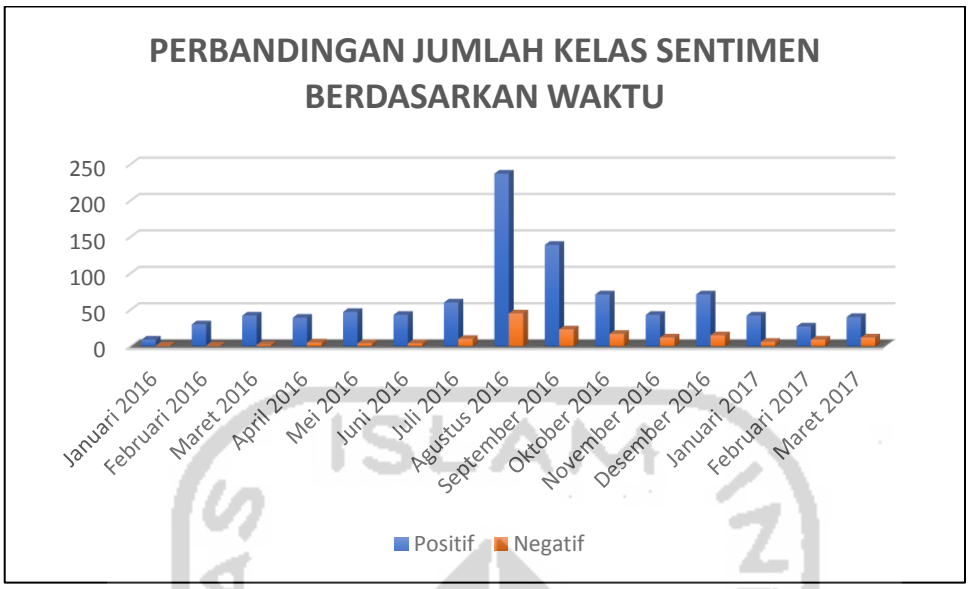
Skor akhir yang diperoleh dari simulasi perhitungan bernilai > 0, sehingga hasil klasifikasi ulasan adalah positif.

Hasil pelabelan kelas sentimen pada dua kelas sentimen positif dan sentimen negatif dapat dilihat pada tabel berikut :

Tabel 5.7 *Jumlah ulasan pada kelas sentimen*

Kelas Sentimen	Jumlah Ulasan
Positif	976
Negatif	167
Total	**1143**

Berdasarkan **Tabel 5.7,** diperoleh hasil pelabelan kelas sentimen dengan jumlah ulasan positif memiliki frekuensi yang lebih tinggi dibandingkan jumlah ulasan negatif. Dari total 1143 ulasan, jumlah ulasan positif adalah sebanyak 976 ulasan, dan ulasan negatif adalah sebanyak 167 ulasan. Berikut adalah grafik perbandingan jumlah kelas sentimen berdasarkan waktu :



Gambar 5.21 *Grafik perbandingan jumlah kelas sentimen berdasarkan urutan waktu*

Berdasarkan **Gambar 5.21** diatas dapat diketahui bahwa pada setiap bulannya sentimen positif selalu lebih banyak daripada sentimen negatif. Banyaknya sentimen positif menunjukkan bahwa penumpang maskapai Garuda Indonesia memiliki persepsi yang baik terhadap maskapai asal Indonesia ini. 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 memiliki perbandingan 80% : 20%. Suthaharan, Shan (2015) menyatakan bahwa

meskipun penelitian ekstensif belum dilakukan dalam pemilihan rasio yang optimal antara kumpulan data ini, ada beberapa praktik umum dalam memilih ukuran kumpulan data ini. Berdasarkan *Paretto Principle*, Rasio yang umum digunakan adalah 80:20 untuk *data sets training* dan *testing*. Perbandingan jumlah data latih dan data uji dapat dilihat pada **Tabel 5.8** berikut :

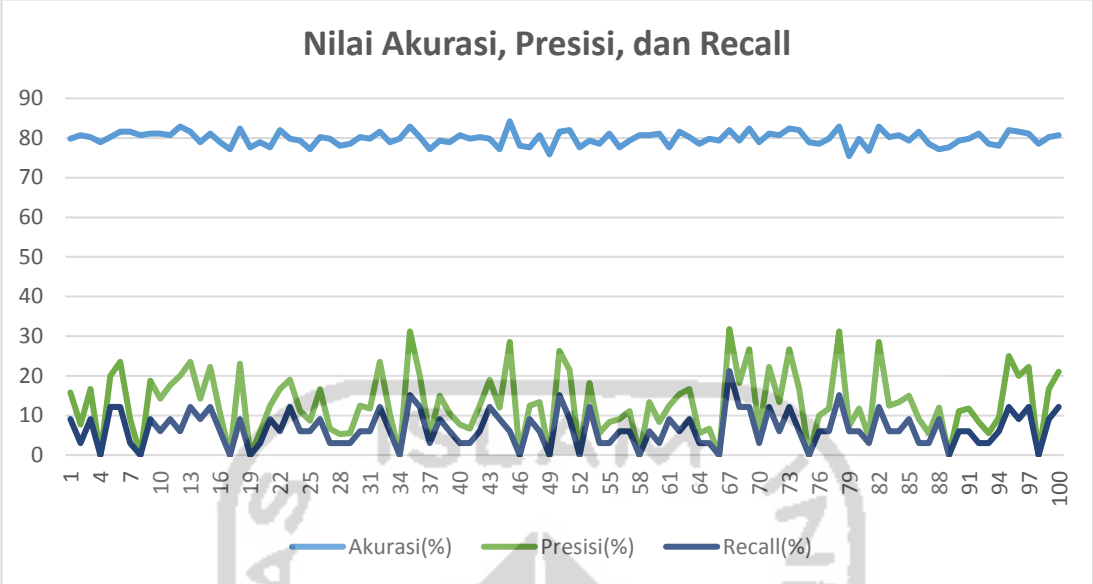
Tabel 5.8 *Perbandingan data latih dan data uji*

Klasifikasi	Jumlah	Data Latih (80%)	Data uji (20%)
Positif	976	780.80 ≈ 781	195.20 ≈ 195
Negatif	167	133.60 ≈ 134	33.40 ≈ 33
Total	**1143**	**915**	**228**

Berdasarkan **Tabel 5.8**, dengan perbandingan data latih dan data uji sebesar 80% : 20%, dari total 1143 data ulasan berbahasa Inggris, digunakan sebanyak 915 data sebagai data latih dan 228 data sebagai data uji.

5.6. Klasifikasi dengan Metode *Naïve Bayes Classifier*

Proses klasifikasi dilakukan dengan cara membuat *machine learning* menggunakan data latih dan data uji secara acak. Penelitian ini menggunakan metode *confusion matrix* dalam proses evaluasi. *Confusion matrix* merupakan salah satu *tools* penting dalam metode evaluasi yang digunakan pada *machine learning* yang biasanya memuat dua kategori atau lebih (Manning, dkk, 2009). Setiap unsur matiks menunjukkan jumlah contoh data uji untuk kelas sebenarnya yang digambarkan dalam bentuk baris sedangkan kolom menggambarkan kelas yang diprediksi. Untuk melakukan evaluasi model, pada percobaan ini dilakukan dengan iterasi pada *dataset* sebagai *cross validation* untuk menemukan nilai akurasi prediksi yang terbaik. Berikut grafik nilai akurasi dari 100 kali iterasi atau percobaan :



Gambar 5.22 *Grafik nilai akurasi, presisi, dan Recall dengan metode NBC*

Gambar 5.22 di atas memperlihatkan hasil akurasi dengan metode *Naïve Bayes Classifier* yang memberikan kisaran atau rata-rata hasil akurasi sebesar 80%. Variasi kenaikan akurasi menunjukkan selisih yang kecil, sehingga dapat dikatakan metode *Naïve Bayes Classifier* merupakan metode yang baik dalam pengklasifikasian teks berbahasa inggris dengan topik multikelas. Model klasifikasi dikatakan baik jika ketiga parameter akurasi, *recall*, dan presisi memiliki nilai yang seimbang. Model yang akurat belum tentu baik jika nilai *recall* dan presisinya rendah. Selain akurasi, *recall*, dan presisi, *confusion matrix* juga bisa digunakan untuk melakukan evaluasi kinerja dari *classifier*.

Berdasarkan **Gambar 5.22** diatas, dari 100 kali iterasi yang dilakukan dilakukan menggunakan metode *Naïve Bayes Classifier* didapatkan tingkat akurasi tertinggi yaitu pada iterasi ke-45 yakni sebesar 84% namun masih memiliki nilai *recall* dan presisi yang rendah yaitu sebesar 6% dan 29% serta memiliki *confusion matrix* yang kurang baik. Hasil perhitungan tingkat akurasi tersebut diperoleh dari jumlah data uji yang terklasifikasi dengan benar dibandingkan dengan total semua data yang di uji. Setelah dilakukan pengamatan, didapatkan model klasifikasi yang paling baik adalah pada iterasi ke-67 yaitu dengan nilai akurasi sebesar 82%, nilai *recall* sebesar 21%, dan nilai presisi sebesar 32%.

Pada penelitian ini juga menggunakan metode *confusion matrix* dalam proses evaluasi. *Confusion matrix* merupakan salah satu *tools* penting dalam metode visualisasi yang digunakan pada mesin pembelajaran yang biasanya memuat dua kategori atau lebih (Manning, dkk, 2009). Setiap unsur matiks menunjukkan jumlah contoh data uji untuk kelas sebenarnya yang digambarkan dalam bentuk baris sedangkan kolom menggambarkan kelas yang diprediksi. **Tabel 5.9** berikut menggambarkan hasil *confusion matrix* prediksi dua kelas sentimen dengan nilai akurasi, *recall*, dan presisi terbaik.

Tabel 5.9 *Hasil confusion matrix*

Prediksi	**Aktual**		**Class Precision**
	Positif	**Negatif**	
Positif	180	26	87,38%
Negatif	15	7	31,82%
Class Recall	92,31%	21,21%	
Akurasi 82,02%			

Berdasarkan **Tabel 5.9**, dengan menggunakan metode *Naïve Bayes Classifier* diperoleh hasil prediksi bahwa pada kelas positif, dari 195 ulasan positif, terdapat 180 ulasan yang sudah terklasifikasi dengan benar dan terdapat kesalahan prediksi sebesar 15 ulasan yang masuk kedalam ulasan negatif sehingga diperoleh nilai presisi untuk kelas positif sebesar 87,38%. Sedangkan pada ulasan negatif, dari total 33 ulasan terdapat 7 ulasan yang sudah terklasifikasi dengan benar sebagai ulasan negatif dan terdapat kesalahan prediksi sebanyak 26 ulasan yang masuk ke dalam ulasan positif, sehingga diperoleh nilai presisi kelas negatif sebesar 31,82%. Kemudian dari nilai *confusion matrix* tersebut diperoleh tingkat akurasi sebesar 82,02%, artinya dari 228 data ulasan yang diujikan, terdapat 187 ulasan yang benar pengklasifikasiannya oleh model *Naïve Bayes Classifier* (NBC).

Kecilnya nilai akurasi pada kelas negatif disebabkan karena ketidakseimbangan (*imbalanced*) antara jumlah data positif dan data negatif, jumlah data latih negatif yang relatif sedikit akan membuat *machine* lebih sulit dalam memahami pola data yang bervariasi, sehingga berpengaruh terhadap

ketepatan hasil prediksi. Untuk memperoleh ketepatan prediksi dalam klasifikasi, diperlukan jumlah data yang relatif seimbang pada masing-masing kelas, artinya antara kelas positif dan kelas negatif jumlah data yang digunakan memiliki perbandingan yang tidak terlampau jauh.

Adapun proses perhitungan nilai akurasi dilakukan dengan menggunakan rumus berikut :

$$\text{Akurasi} = \frac{\text{Jumlah data yang terprediksi dengan benar}}{\text{Jumlah semua data yang di uji}} \; x \; 100\% \qquad \textbf{(5.2)}$$

Sehingga, untuk ulasan berbahasa Inggris, nilai akurasi diperoleh dari perhitungan berikut:

$$\text{Akurasi} = \frac{180 + 7}{180 + 15 + 26 + 7} \; x \; 100\%$$

$$\text{Akurasi} = \frac{187}{228} \; x \; 100\%$$

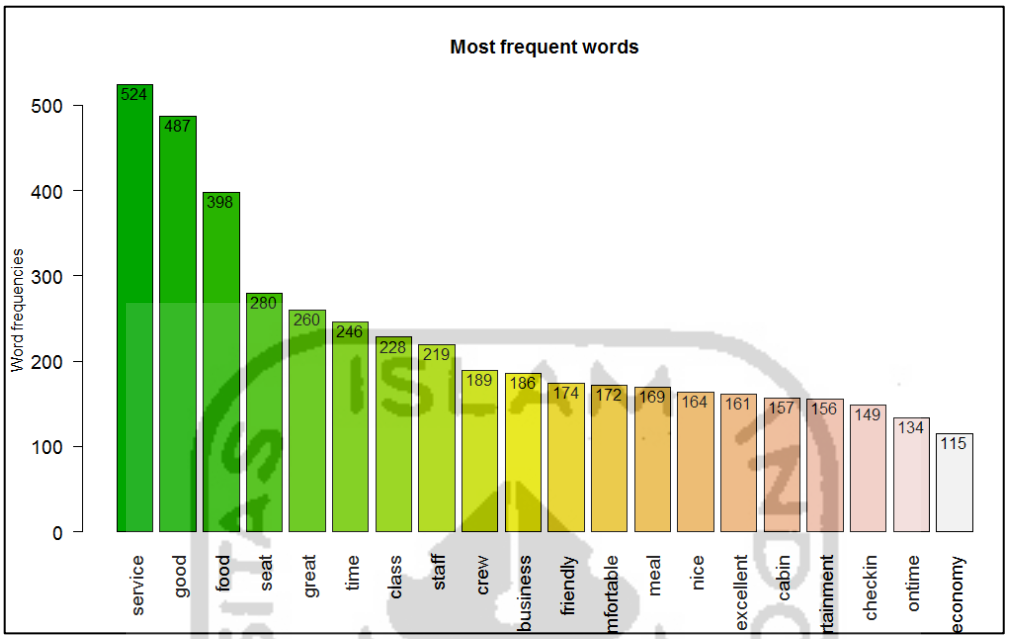
$$\text{Akurasi} = 82{,}02\%$$

5.7. Visualisasi dan Asosiasi

Visualisasi dilakukan terhadap masing-masing klasifikasi kelas sentimen. Adapun tujuan visualisasi adalah untuk mengekstraksi informasi berupa topik yang paling sering di bicarakan / diulas oleh penumpang maskapai Garuda Indonesia, 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 penjelasan hasil visualisasi dan asosiasi kata dari setiap klasifikasi kelas sentimen.

5.7.1. Ulasan Positif

Data ulasan positif yang digunakan adalah data hasil pelabelan yang dilakukan baik menggunakan kamus *lexicon* maupun secara manual. Ekstraksi informasi pada ulasan positif dilakukan secara berulang-ulang hingga mendapatkan informasi tentang ulasan positif penumpang maskapai Garuda Indonesia yang paling sering diulas / 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 5.23 *Kata yang paling banyak muncul pada kelas positif*

Berdasarkan hasil klasifikasi ulasan positif, dari jumlah ulasan positif sebanyak 976 ulasan, diperoleh beberapa kata yang paling banyak muncul diantaranya adalah kata "*service*" dengan frekuensi sebanyak 524 kali, "*good*" sebanyak 487 kali, "*food*" 398 kali, dan seterusnya. Kata-kata yang muncul seperti pada **Gambar 5.23** merupakan kata yang memiliki sentimen positif dan merupakan topik pembicaraan yang paling banyak diulas oleh pengunjung. 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 juga dapat ditampilkan dalam bentuk *wordcloud* seperti terlihat pada **Gambar 5.24.**

Gambar 5.24 *Wordcloud ulasan positif*

Berdasarkan 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. Selanjutnya, dilakukan pencarian asosiasi antar kata yang sering muncul secara bersamaan dan diperoleh hasil sebagai berikut :

Tabel 5.10 *Asosiasi kata pada kelas sentimen positif*

service		food		seat		time	
good	0,19	drinks	0,21	comfortable	0,26	excited	0,13
excellent	0,18	good	0,20	exclusively	0,25	luckily	0,11
exceptional	0,14	impressive	0,15	length	0,25	satisfy	0,11
executive	0,13	escorted	0,14	nature	0,25	sooner	0,11
faultless	0,13	freshly	0,14	scenery	0,25	cheerful	0,11
quality	0,12	stunning	0,14	large	0,20	passionate	0,11
exclusively	0,12	heat	0,14	simple	0,15	loyalty	0,11
		delicious	0,13	elite	0,15	flawless	0,11
		classy	0,12	happy	0,13	joyful	0,11
		flavors	0,12	entertained	0,13	beautiful	0,10
		variety	0,11	comfy	0,12		

		local	0,10	easy	0,12		
		great	0,10	roomy	0,12		
staff		**entertainment**		**checkin**		**cabin**	
friendly	0,19	movie	0,17	easy	0,30	temperature	0,22
great	0,17	fulfill	0,16	rushing	0,29	consistently	0,19
helpful	0,14	nicest	0,16	priority	0,22	elegant	0,19
assistance	0,14	religious	0,16	lounge	0,20	lighting	0,19
professional	0,13	privacy	0,14	quick	0,20	clean	0,16
rushing	0,12	music	0,14	efficiently	0,20	satisfying	0,16
wonderful	0,12	variety	0,13	online	0,20	wonderfully	0,14
dearer	0,12	recommend	0,11	dedicated	0,15	properly	0,11
		complimentary	0,11	smoothly	0,13	outstanding	0,11
		games	0,10	seamless	0,11	spacious	0,10
		happily	0,10				

Berdasarkan **Tabel 5.10**, 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 diulas. Dari **Tabel 5.10** diatas, jika dilihat asosiasi kata yang berkaitan dengan kata "*service*", dapat diperoleh informasi tentang *service* atau pelayanan yang sangat bagus, luar biasa, eksklusif dan sempurna.

Kata-kata yang berasosiasi dengan kata "*food*" juga memberikan informasi makanan yang beranekaragam, mulai dari makanan yang baik, segar, lezat, bervariasi, berkelas dan mengesankan.

Kata-kata yang berasosiasi dengan kata "*seat*" memberikan informasi tentang tempat duduk yang nyaman, eksklusif, panjang, luas, dan dapat melihat pemandangan alam sehingga penumpang merasakan senang dan terhibur.

Kata-kata yang berasosiasi dengan kata "*time*" yang berarti waktu, baik waktu pemberangkatan atau waktu sampai. Dari **Tabel. 5.10** dapat dilihat bahwa kata *time* atau waktu memberikan informasi bahwa lebih cepat, memuaskan, menyenangkan, beruntung, dan sempurna.

Kata-kata yang berasosiasi dengan kata "*staff*" memberikan informasi tentang kinerja *staff* maskapai Garuda Indonesia dinilai ramah, professional, hebat, perhatian, peduli dan bergegas membantu penumpang.

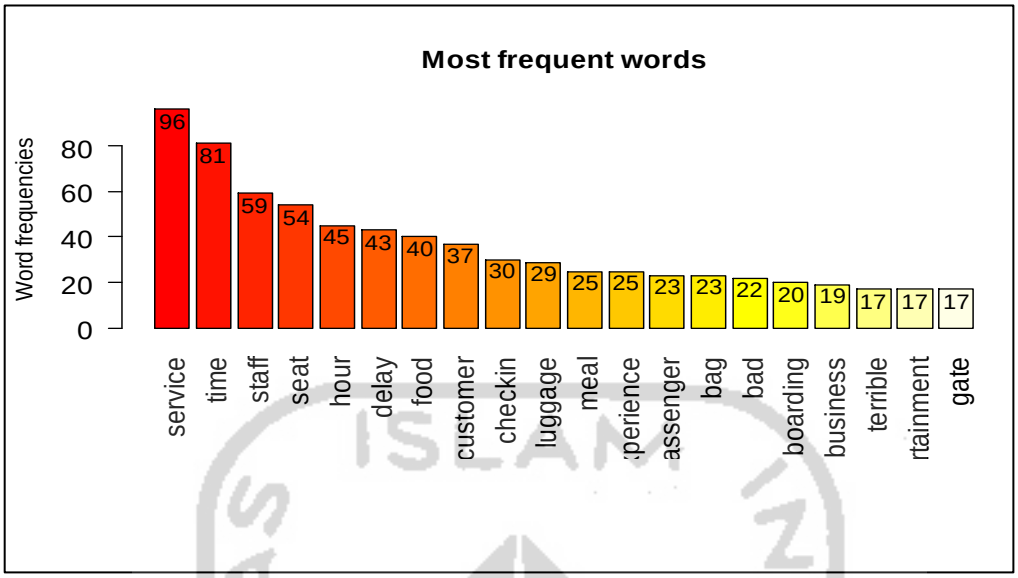
Kata-kata yang berasosiasi dengan kata "*entertainment*" memberikan informasi tentang hiburan yang diberikan maskapai Garuda Indonesia dinilai sangat memenuhi atau memuaskan dan juga paling bagus, banyak variasi mulai dari film, musik dan *games*. Hiburan yang diberikan juga bersifat pribadi untuk masing-masing penumpang serta bersifat gratis sehingga penumpang merasa gembira dan sangat menyarankan untuk menggunakan maskapai ini.

Kata-kata yang berasosiasi dengan kata "*checkin*" memberikan informasi tentang proses *check-in* berlangsung dengan mudah, cepat, lancar, mulus, efisien, dan juga diprioritaskan.

Kata-kata yang berasosiasi dengan kata "*cabin*" memberikan informasi tentang kabin pesawat yang luas, bersih, sangat memuaskan, luar biasa, dan memiliki suhu dan penerangan yang tepat.

5.7.2. Ulasan Negatif

Ekstraksi informasi pada ulasan negatif dilakukan secara berulang-ulang hingga mendapatkan informasi tentang ulasan negatif maskapai Garuda Indonesia yang paling sering diulas / dibicirakan. Berdasarkan hasil pelabelan, ulasan negatif pengunjung terhadap maskapai cukup sedikit bila dibandingkan dengan jumlah ulasan positif. Dari total ulasan sebanyak 1143 ulasan, hanya teridentifikasi sebanyak 167 ulasan negatif. Hal tersebut menunjukkan bahwa mayoritas penumpang maskapai Garuda Indonesia mempunyai persepsi yang baik terhadap masakapai tersebut. 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 penumpang dengan klasifikasi ulasan negatif.



Gambar 5.25 *Kata yang paling banyak muncul pada kelas negatif*

Berdasarkan hasil klasifikasi ulasan negatif, diperoleh beberapa kata yang paling banyak muncul dengan topik yang dianggap relevan sebagai sentimen negatif diantaranya adalah kata "*service*" dengan frekuensi sebanyak 96 kali, "*time*" sebanyak 81 kali, "*staff*" sebanyak 59 kali, "*seat*" sebanyak 54 kali, "*hour*" sebanyak 45 kali, dan seterusnya. Kata-kata yang muncul seperti pada **Gambar 5.25** merupakan kata yang memiliki sentimen negatif berbahasa Inggris dan merupakan topik pembicaraan yang paling banyak diulas oleh penumpang. 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 5.26.**

Gambar 5.26 *Wordcloud ulasan negatif*

Visualisasi *wordcloud* pada **Gambar 5.26** 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 *service*, *time*, *staff*, *seat*, *hour* dan sebagainya. Selanjutnya, dilakukan pencarian asosiasi antar kata yang sering muncul secara bersamaan dan diperoleh hasil sebagai berikut :

Tabel 5.11 *Asosiasi kata pada kelas sentimen negatif*

service		**staff**		**seat**		**food**	
distance	0,46	catch	0,43	begging	0,38	inedible	0,52
catering	0,46	missed	0,41	distance	0,38	pasta	0,51
postponed	0,46	apologies	0,41	separate	0,38	dried	0,44
price	0,36	confused	0,40	full	0,34	hungry	0,44
full	0,35	misunderstanding	0,40	cramped	0,28	local	0,44
ridiculous	0,31	unfriendly	0,33	worse	0,27	serving	0,44
worse	0,29	arguing	0,29	smaller	0,27	western	0,29
anger	0,21	uninformed	0,29	troubled	0,27	hideous	0,28
convoluted	0,21	disinterested	0,29	complain	0,24	lousy	0,28
fiasco	0,21	undelivered	0,29	flat	0,19	wine	0,26
pathetic	0,21	lose	0,26	mistake	0,18	disappointing	0,23
confusing	0,17	mess	0,26	problems	0,18	vegetarian	0,20
compensation	0,16	communication	0,19	ankle	0,17	tasted	0,17

hour		checkin		luggage		boarding	
delay	0,49	crashes	0,52	compensation	0,79	queue	0,42
communicated	0,36	resigned	0,52	mistakenly	0,72	crazy	0,39
waiting	0,23	tools	0,52	unfortunate	0,72	insulting	0,37
distance	0,23	response	0,45	careless	0,50	loud	0,37
postponed	0,23	irregularity	0,38	fuss	0,50	shocking	0,32
reserve	0,23	information	0,36	missing	0,50	rush	0,25
appalling	0,23	feature	0,35	claims	0,45	apologies	0,25
prioritized	0,23	worse	0,26	wait	0,44	priority	0,19
disinterested	0,23	uninformed	0,25	hassle	0,34	security	0,19
urgent	0,23	expense	0,25	lose	0,30	employee	0,17
late	0,21	impolite	0,25	deteriorated	0,26	anxiety	0,17
		joke	0,25	unacceptable	0,26	disrupted	0,17
		notorious	0,25	flip	0,26	sloppy	0,17
		difficult	0,19	errors	0,17	anger	0,17
		trouble	0,19	expensive	0,17	convoluted	0,17
		poor	0,16	terrible	0,16	fiasco	0,17
		uncomfortable	0,16			pathetic	0,17
		stuck	0,16			sadly	0,17

Tabel 5.11 menunjukkan asosiasi antar kata pada ulasan negatif, 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 "*service*" pada ulasan negatif memberikan informasi tentang keluhan pelanggan atau penumpang terhadap buruknya pelayanan, konyol, membelit, menyedihkan, membingungkan, kegagalan, penundaan jadwal, masalah harga serta kompensasi yang diberikan.

Kata-kata yang berasosiasi dengan kata "*staff*" pada ulasan negatif memberikan informasi tentang keluhan pelanggan atau penumpang terhadap kinerja *staff* maskapai Garuda Indonesia yang dinilai bingung, tidak ramah, kacau, tidak menjawab serta kurangnya komunikasi dan penyampaian informasi sehingga banyak terjadi kesalahpahaman.

Kata-kata yang berasosiasi dengan kata "*seat*" pada ulasan negatif memberikan informasi tentang tempat duduk yang penuh, sempit, lebih kecil, lebih buruk, tempat buat pergelangan kaki serta banyak yang mengeluh jarak yang terpisahkan antara rekan seperjalanannya.

Kata-kata yang berasosiasi dengan kata "*food*" pada ulasan negatif memberikan informasi tentang ketidakpuasan beberapa penumpang terhadap makanan yang disajikan maskpai Garuda Indonesia yang dinilai mengecewakan, mengerikan, masalah rasa, porsi, sehingga kebanyakan tidak termakan. Selain itu penumpang mengeluhkan karena makanan yang disediakan kebanyakan menu lokal dan tidak terdapat *wine* atau anggur.

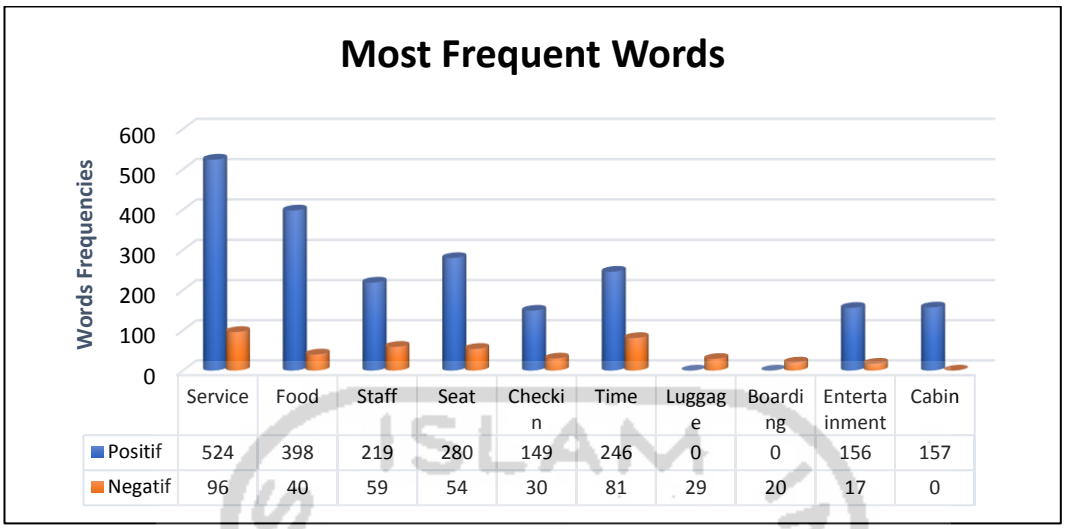
Kata-kata yang berasosiasi dengan kata "*hour*" tentunya akan memberikan informasi tentang keluhan penumpang terhadap masalah waktu pada penerbangan dengan maskapai Garuda Indonesia. Kata-kata yang berasosiasi dengan kata "*hour*" diantaranya adalah kata *delay, waiting, late*, dan sebagainya. Kata-kata tersebut mengandung arti bahwa pada penerbangan Garuda Indonesia masih memiliki masalah pada waktu seperti penundaan penerbangan sehingga penumpang harus menunggu dengan jarak yang cukup lama dan terlambat sampai ketempat tujuannya dan melewati kegiatan penting yang sudah dijadwalkan sebelumnya.

Kata-kata yang berasosiasi dengan kata "*checkin*" memberikan informasi tentang keluhan penumpang terhadap proses *check-in* diantaranya ketidakteraturan, kurangnya informasi, kesulitan, dan terjebak. Selain itu masalah biaya dan keramahan petugas *check-in* pun menjadi keluhan penumpang.

Kata-kata yang berasosiasi dengan kata "*luggage*" memberikan informasi tentang keluhan penumpang terhadap pelayanan bagasi yang buruk, rumit, haus menunggu lama, terjadi kesalahan dan hilang. Selain itu masalah harga dan kecerobohan serta susahnya mendapatkan kompensasi menjadi keluhan dan sangat disayangkan oleh penumpang.

Kata-kata yang berasosiasi dengan kata "*boarding*" memberikan informasi tentang keluhan penumpang terhadap proses *boarding* yang penuh antrian, gila, membelit, menyedihkan, mengejutkan, gelisah, serta masalah keamanan juga ikut menjadi keluhan pada proses ini.

Selanjutnya jika kedua *barplot* pada kedua kelas sentimen disandingkan, maka akan terlihat kata-kata yang sering muncul pada kedua kelas sentimen seperti yang terlihat pada **Gambar 2.27** berikut :



Gambar 5.27 *Kata benda yang paling banyak muncul pada kedua kelas sentimen*

Sepuluh kata benda yang paling sering muncul pada hasil klasifikasi sentimen pada kelas positif dan negatif yaitu kata *service*, *food*, *staff*, *seat*, dan seterusnya seperti pada **Gambar 5.27**. Kata yang paling banyak muncul pada klasifikasi tersebut adalah *service*, kata *service* selalu muncul dengan frekuensi tertinggi pada setiap kelas sentiment baik kelas positif maupun negatif. Kemudian dari sepuluh kata tersebut terdapat beberapa kata yang muncul hanya pada satu kelas saja, yaitu kata *luggage* yang hanya muncul pada kata kelas negatif, sedangkan kata *cabin* yang hanya muncul pada kata kelas positif.

BAB VI
PENUTUP

6.1. Kesimpulan

Berdasarkan hasil analisis yang telah dilakukan, diperoleh beberapa kesimpulan sebagai berikut :

1. Untuk melakukan proses pengambilan data ulasan dengan menggunakan teknik *web scraping* dibutuhkan bantuan beberapa *tools* seperti *software R* dan *browser google chrome* yang sudah harus terinstal pada komputer. Salain itu *packages* `rvest` juga harus terinstal pada *software R* dan *selector gadget* telah terpasang pada *google chrome*. Kemudian proses *scraping* dilakukan pada *software R* dan simpan dalam format *Comma Separated Value (CSV)*. Dengan menggunakan teknik *web scraping* didapatkan data yang berupa ulasan penumpang tentang maskapai Garuda Indonesia sejak bulan Januari 2016 sampai dengan bulan Maret 2017 sebanyak 1143 ulasan yang terdiri dari variabel *ID, Quote, Rating, Date, Review,* dan *Flights*.
2. Berdasarkan analisis yang telah dilakukan maka dapat diketahui bahwa jumlah ulasan pada setiap bulannya sejak Januari 2016 hingga Maret 2017 cenderung mengalami fluktuasi, kenaikan jumlah ulasan meningkat secara signifikan pada bulan agustus tahun 2016. Kenaikan jumlah ulasan tersebut diduga karena bertepatan dengan jadwal pemberangkatan jama'ah haji 2016. Pada ulasan tersebut didominasi oleh ulasan dari penumpang yang melakukan perjalanan internasional yaitu sebanyak 82%. Hal ini dikarenakan data yang digunakan adalah data ulasan berbahasa internasional (Bahasa Inggris). Kemudian berdasarkan rating dan juga hasil pelabelan kelas sentimen dapat diketahui bahwa mayoritas penumpang maskapai Garuda Indonesia mempunyai penilaian ataupun presepsi yang baik terhadap maskapai tersebut yang dibuktikan dengan jumlah ulasan positif jauh lebih banyak dari pada ulasan negatif yaitu sebanyak 976 ulasan dari total 1143 ulasan.
3. Dengan menggunakan perbandingan data latih dan data uji sebesar 80% : 20% diperoleh hasil klasifikasi sentimen menggunakan model *Naïve Bayes*

Classifier diperoleh tingkat akurasi sebesar 82,02%, artinya dari 228 data ulasan yang diujikan, terdapat 187 ulasan yang benar pengklasifikasiannya.

4. Berdasarkan hasil klasifikasi dan asosiasi teks yang dilakukan, secara umum dapat diketahui bahwa penumpang maskapai Garuda Indonesia mayoritas membicarakan mengenai *service*, *staff, food,* dan *check-in* karena selalu muncul baik pada kelas sentimen positif maupun negatif. Secara umum metode asosiasi teks yang digunakan menunjukkan hasil ekstraksi informasi pada kelas positif diantaranya terkait *service, food, seat, time, staff, entertainment, check-in,* dan *cabin*. Sedangkan pada kelas negatif yang sering dikeluhkan diantaranya *service, staff, seat, food, hour, check-in, luggage,* dan *boarding*.

6.2. Saran

Berdasarkan kesimpulan di atas, dapat diberikan beberapa saran sebagai berikut :

1. Bagi pihak manajemen maskapai Garuda Indonesia, hasil ekstraksi informasi dari ulasan-ulasan yang telah diberikan oleh pelanggan khususnya ulasan yang berbentuk negatif dapat dijadikan bahan evaluasi dalam peningkatan kepuasan pelanggan dan memberikan pelayanan semaksimal mungkin, serta untuk pengembangan bisnis selanjutnya.
2. Penanganan negasi belum menjadi fokus utama dalam analisis sentimen pada penelitian ini, sehingga kalimat yang memiliki kata negasi belum dapat ditentukan polaritasnya secara optimal. Untuk penelitian selanjutnya diharapkan dapat melakukan penanganan khusus terhadap kata negasi agar hasil yang didapatkan lebih akurat.
3. Dalam penelitian ini data ulasan yang digunakan masih dibatasi untuk ulasan yang berbahasa Inggris saja, sehingga perlu dikembangkan pada penelitian selanjutnya dengan menggunakan ulasan berbagai bahasa.
4. Bagi peneliti selanjutnya, dapat menggunakan pendekatan *machine learning* lain sebagai pembanding performa algoritma *Naïve Bayes Classifier* untuk mengklasifikasikan ulasan penumpang maskapai garuda Indonesia pada situs *TripAdvisor*.

DAFTAR PUSTAKA

Arifin, Oki. 2016. *Penentuan Prioritas Pemasangan Internet Untuk Pelanggan Baru Perusahaan Menggunakan Naive Bayes (Studi Kasus : Pt. Time Excelindo).* Tesis. Program Studi Magister Ilmu Komputer FMIPA UGM Yogyakarta.

Basnur, P.W., 2009. *Pengklasifikasian Artikel Berita Berbahasa Indonesia Secara Otomatis Menggunakan Ontologi*. Tugas Akhir. Program Ilmu Komputer Fakultas Ilmu Komputer, Universitas Indonesia, Depok.

Bishop. C. M., 2005. *Neural Network for Pattern Recognition*. United States: Oxford University Press.

Boiy, E. et.al. 2007. *Automatic Sentiment Analysis in Online Text*. Proceedings of the Conference on Electronic Publishing(ELPUB-2007), pp. 349-360.

Bouge, Kevin. 2011. *Download stop words*. https://sites.google.com/site/kevinbouge/stopwords-lists. Diakses pada tanggal 5 Maret 2017 pukul 15:30 WIB.

Cunningham P, Smyth B, Wu G, et al. 2010. *Does TripAdvisor make hotels better?*. Technical Report UCD-CSI-2010-06.

Dave, K., Lawrence, S., dan Pennock, D.M. 2003. *Mining the peanut gallery: opinion extraction and semantic classification of product reviews*. Proceeding of WWW '03 Proceedings of the 12th international conference on World Wide Web (pp. 519-528). Hungary: ACM.

Dewantoro, P.R. 2016. *Implementasi Teknik Web Scraping Pada Proses Topic Modelling Portal Berita.* Skripsi. Program Studi Ilmu Komputer FMIPA UGM Yogyakarta.

Even, Y., dan Zohar. 2002. *Introduction to Text Mining,Automated Learning Group National Center for Supercomputing Applications*. University of Illionis.

Fadlisyah, B. D. A. 2014. *Statistika : Terapannya di Informatika, 1st ed*. Yogyakarta: Graha Ilmu.

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

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

Garuda Indonesia. 2016. *Garuda Indonesia Siap Berangkatkan 79.020 Calon Jemaah Haji Indonesia Mulai 9 Agustus 2016*. https://www.garuda-indonesia.com/id/id/corporate-partners/company-profile/about/index.page?. Diakses pada tanggal 12 Maret 2017 pukul 09:30 WIB.

Garuda Indonesia. 2016. *Tentang Garuda Indonesia*. https://www.garuda-indonesia.com/id/id/news-and-events/garuda-siap-berangkatkan-calon-jemaah-haji.page. Diakses pada tanggal 20 April 2017 pukul 19:45 WIB.

Ginanjar, Gilang. 2009. *Analisis Kualitas Pelayanan Maskapai Garuda Indonesia Terhadap Kepuasan Penumpang*. Skripsi. Jurusan Teknik Penerbangan Sekolah Tinggi Teknologi Adisutjipto Yogyakarta.

Go, A., Bhayani, R. dan Huang, L. 2009. *Twitter Sentiment Classification using Distant Supervision*. Stanford : CS224N Project Report.

Gretzel, U. et.al. 2007. *Online Travel Review Study: Role and Impact of Online Travel Reviews*. Texas : Laboratory for Intelligent Systems in Tourism, Texas A & M University.

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

Hardjana, Agus M. 2003. *Komunikasi intrapersonal & Komunikasi Interpersonal*. Yogyakarta: Penerbit Kanisius.

Hamzah, Amir. 2012. *Klasifikasi Teks Dengan Naïve Bayes Classifier (Nbc) Untuk Pengelompokan Teks Berita Dan Abstract Akademis.* Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III. ISSN:1979-911X.

Harrison-Walker, L. J., 2001. *The measurement of word-of-mouth communication and investigation of service quality and customer commitment as potential antecedents*. Journal of Service Research 4 (1), 60-75.

Hennig-Thurau, T., et.al. 2014. *Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet?*. Journal of Interactive Marketing, 2004, 18(1): 38-52.

Hu and Liu. 2004. *Opinion Lexicon: A list of English positive and negative opinion words or sentiment words*. http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar. Diakses pada tanggal 5 Maret 2017 pukul 14:00 WIB.

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, hlm. 159-164.

Juliasari, N., & Sitompul, J. C. (2012). *Aplikasi Search Engine dengan Metode Depth First Search (DFS)*. BIT Numerical Mathematics, 9.

Kaji, N., and Kitsuregawa, M. 2007. *Building Lexicon For Sentiment Analysis From Massive Collection Of Html Documents*. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1075–1083.

Kurniawan, Bambang, dkk. 2012. *Klasifikasi Konten Berita Dengan Metode Text Mining*. Jurnal dunia teknologi informasi vol. 1, no.1, (2012) 14 – 19.

Kohavi, R. 1995. *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. International Joint Conference on Artificial Intelligence 14.12, pp. 1137–1143.

Kotler, P. dan Amstrong, G. 2012 . *Principles Of Marketing*. Global Edition,. 14 Edition, Pearson Education.

Larose, D.T. 2006. *Naïve Bayes Estimation and Bayesian Networks, in Data Mining Methods and Models*. USA: John Wiley & Sons Inc.

Lim, S. Y, Song, M.H., & Lee, S.J. 2006. *Ontology-based automatic classification of web documents*. SPringer-Verlag, 690-700.

Liu, B., 2012. *Sentiment Analysis and Opinion Mining*. San Rafael: Morgan & Claypool Publishers.

_____. 2010. *Sentiment Analysis and Subjectivity*. Chicago: University of Illinois.

Ma'arif, M.R. 2016. *Integrasi Laman Website tentang Pariwisata Daerah Istimewa Yogyakarta Memanfaatkan Teknologi Web Scraping dan text Mining.* Jurnal Teknomatika, 9(1): 71-80.

Manning, C. D., et.al. 2009. *An Introduction to Information Retrieval – Online Edition*. Cambridge: Cambridge University Press.

Masithoh, Nurul. 2016. *Analisis Klasifikasi Topik Menggunakan Metode Naïve Bayes Classifier, Naïve Bayes Multinomial Classifier, Dan Maximum Entropy Pada Artikel Berita.* Tugas Akhir. Program Studi Statistika FMIPA UGM Yogyakarta.

Miguéns, J., et.al. 2008. *Social media and tourism destinations: TripAdvisor case study*. Advances in Tourism Research (Aveiro).

Mohri, et al. 2012. *Foundations of machine learning*. Cambridge: MIT Press.

Nasukawa, T. & Yi, J., 2003. *Sentiment Analysis: Capturing Favorability Using Natural Language Processing*. In Proceedings of the 2nd International Conference on Knowledge Capture. pp. 70–77.

O'keefe, T. Koprinska, I. 2009. *Feature selection and weighting methods in sentiment analysis. In: Proceedings of the 14th Australasian document computing symposium*. Sydney, pp 67–74.

Pang, Bo., et. al. 2002 . *Thumbs up? Sentiment Classification using Machine Learning Techniques*. Proceedings of EMNLP 2002, pp. 79–86.

_____. 2008. *Opinion Mining and Sentiment Analysis*. Foundation and Trends in Information Retrieval. Vol. 2, No 1-2 (2008) 1–135.

Prasetyo, E., 2012. *Data Mining Konsep dan Aplikasi menggunakan Matlab*. Yogyakakara: Andi.

Putranti, N.D. 2013. *Analisis Sentimen Twitter Untuk Teks Berbahasa Indonesia Dengan Maximum Entropy Dan Support Vector Machine*. Tesis. Magister Ilmu Komputer UGM Yogyakarta

Rafaeilzadeh, dkk., 2008. *Cross-Validation*. Encyclopedia of Database Systems, 532-538.

Ramadhani, Tiara G. 2015. *Analisis Sentimen Menggunakan Metode Naïve Bayes Classifier Dengan Model Dokumen Bernoulli Dan Support Vector Machine*. Program Studi Statistika FMIPA UGM Yogyakarta.

Rianto, Bagus.2016. *Implementasi dan Perbandingan Metode Prapemprosesan Pada Analisis Sentimen Gubernur DKI Jakarta Menggunakan Metode Support Vector Machine dan Naïve Bayes.* Tugas Akhir. Program Studi Ilmu Komputer FMIPA UGM Yogyakarta.

Saks, Greg. 2006. *Travel: The emergence of Meta Search*. Compete.

Saraswati, N.W.S. 2011. *Text Mining dengan Metode Naïve Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis*. Thesis. Program Pascasarjana Universitas Udayana.

Schiffman, L.G., dan Kanuk, L. L. 2010. *Consumer Behavior*. 10th edition. New Jersey: Pearson Prentice Hall.

Sen, S. dan Lerman, D. 2007. *Why Are You Telling Me This? An Examination into Negative Consumer Reviews on the Web*. Journal of Interactive Marketing (21:4), pp. 76-94.

Shah C. and A. G. Jivani. 2013. *Comparison of data mining classification algorithms for breast cancer prediction*. Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on, 2013, pp. 1–4.

Statsoft. 2015. *Naive Bayes Classifier Introductory Overview*. http://www.statsoft.com/textbook/naivebayes-classifier. Diakses pada tanggal 15 April 2017 pukul 10:30 WIB.

Sugiama, A. Gima. 2011. *Eco Tourism*. Bandung : Guardaya Intimarta.

Suryana, T. dan Koesheryatin. 2014. *Aplikasi internet menggunakan HTML, CSS & JavaScript*. Jakarta: Elex Media Komputindo

Suthaharan, Shan. 2015. *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*. Springer. p. 10. ISBN 9781489976413.

Suwantoro, Gamal. 2004. *Dasar – Dasar Pariwisata*. Yogyakarta: Andi Publisher.

Tan, et al. 2006. *Introduction To Data Mining*. USA: Addison-Wesley.

Turban, E., et.al. 2005. *Decision Support System And Intelligent System*. Upper Saddle River, New Jersey USA: Prentice Hall

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

Ulwan, M. N. 2016. *Pattern Recognition Pada Unstructured Data Teks Menggunakan Support Vector Machine Dan Association.* Tugas Akhir. Program Studi Statistika FMIPA UII Yogyakarta.

Wickham, H., dan RStudio. 2016. *Easily Harvest (Scrape) Web Pages*. https://cran.r-project.org/web/packages/rvest/index.html. Diakses pada tanggal 15 April 2017 pukul 18:25 WIB.

Yoeti, O. A. 2001. *Pemasaran Pariwisata*. Bandung : Angkasa.

LAMPIRAN

Lampiran 1 *Script R Web Scraping Situs TripAdvisor*

```
library(rvest)

url<-read_html("https://www.tripadvisor.com/Airline_Review-d8729079-
Reviews-Cheap-Flights-Garuda-Indonesia#REVIEWS")

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

a<-0:(npages-1)
b<-10
res<-numeric(length=length(a))
for (i in seq_along(a)) {
res[i]<-a[i]*b
}

tableout <- data.frame()

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

url <- paste ("https://www.tripadvisor.com/Airline_Review-d8729079-Reviews-
Cheap-Flights-or",i,"-Garuda-Indonesia#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()

flights <- reviews %>%
  html_node(".categoryLabel") %>%
  html_text()

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

write.csv(tableout, "D://Bismillah Skripsi/Data/Garuda.csv")
save.image()
```

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("D://Bismillah Skripsi/Data")
docs<-readLines("Data Garuda.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_english.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("i","you","we","also","airline","Garuda","Indonesia"))
```

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

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

#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, 10)

dataframe<-data.frame(text=unlist(sapply(docs, `[`)), stringsAsFactors=F)

write.csv(dataframe, "D://Bismillah Skripsi/Data/Hasil_Cleaning.csv")
save.image()
```

Lampiran 3 *Script R Pelabelan dan Pembobotan*

```
library(tm)
setwd("D://Bismillah Skripsi/Data/")
docs<-read.csv("Hasil_Cleaning.csv",header=TRUE)

#skoring
positif <- scan("D://Bismillah Skripsi/SCRIPT/positive-
words.txt",what="character",comment.char=";")
negatif <- scan("D://Bismillah Skripsi/SCRIPT/negative-
words.txt",what="character",comment.char=";")
kata.positif = c(positif, "is near to")
kata.negatif = c(negatif, "cant")
score.sentiment = function(docs, kata.positif, kata.negatif,
.progress='none')
{
 require(plyr)
 require(stringr)
 scores = laply(docs, 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) - (1*sum(negatif.matches))
 return(score)
 }, kata.positif, kata.negatif, .progress=.progress )
 scores.df = data.frame(score=scores, text=docs)
 return(scores.df)
}

hasil = score.sentiment(docs$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 = "Label-english.csv")
```

Lampiran 4 *Script R Klasifikasi dengan menggunakan metode Naïve Bayes Classifier (NBC)*

```
# Load Packages
library(tm)
library(e1071)
library(dplyr)
library(caret)
library(wordcloud)
library(RColorBrewer)
library(streamgraph)


# Import Data
df<- read.delim("clipboard", quote = "")
glimpse(df)

# Randomize dataset
set.seed(1)
df <- df[sample(nrow(df)), ]
glimpse(df)

# preparing corpus
corpus <- Corpus(VectorSource(df$text))

#data cleanup
corpus.clean <- corpus %>%
  tm_map(content_transformer(tolower)) %>%
  tm_map(removePunctuation) %>%
  tm_map(removeWords, stopwords("english"))%>%
  tm_map(stripWhitespace)

# matrix representation
dtm <- DocumentTermMatrix(corpus.clean, control = list(weighting =
weightTfIdf))


# Split Data

trainIndex <- createDataPartition(df$Polarity, p = 0.8,
                                  list = FALSE,
                                  times = 1)

df.train <- df[trainIndex, ]
df.test <- df[-trainIndex, ]

dtm.train <- dtm[trainIndex, ]
dtm.test <- dtm[-trainIndex, ]

n_train<-length(df.train$Polarity)
n_test<-length(df.test$Polarity)

corpus.clean.train <- corpus.clean[1:n_train]
corpus.clean.test <- corpus.clean[1:n_test]

#feature selection

term_control <- findFreqTerms(dtm.train, 1)

dtm.train.nb <- DocumentTermMatrix(corpus.clean.train,
control=list(dictionary = term_control))
```

```
dtm.test.nb <- DocumentTermMatrix(corpus.clean.test,
control=list(dictionary = term_control))

# Function to convert the word frequencies to yes (presence) and no
(absence) labels
convert_count <- function(x) {
  y <- ifelse(x > 0, 1,0)
  y <- factor(y, levels=c(0,1), labels=c("No", "Yes"))
  y
}

# Apply the convert_count function to get final training and testing DTMs
trainNB <- apply(dtm.train.nb, 2, convert_count)
testNB <- apply(dtm.test.nb, 2, convert_count)

# Train the classifier
system.time( classifier <- naiveBayes(trainNB, df.train$Polarity, laplace =
1))

# Use the NB classifier we built to make predictions on the test set.
system.time( pred <- predict(classifier, newdata=testNB) )

# Create a truth table by tabulating the predicted class labels with the
actual class labels
table("Predictions"= pred,  "Actual" = df.test$Polarity)

# Prepare the confusion matrix
conf.mat <- confusionMatrix(pred, df.test$Polarity)
conf.mat

conf.mat$byClass

conf.mat$overall

# Prediction Accuracy
conf.mat$overall['Accuracy']
```

Lampiran 5 *Script R Visualisasi dan Asosiasi Teks (Untuk Ulasan Positif)*

```
# 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("D://Bismillah Skripsi/Data")
docs<-readLines("Label Positif.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("garuda","indonesia","yogyakarta","jogja","jakarta","cgk","bali",
"dont","evenings","nights","morning","flew","back","ill","goid","fly","flig
ht","flying","youre",
"we","pass","passed","left","left","denpasar","not","does","wow","add","one
","im","kid","id","have","par",
"flights","return","kg","adds","st","part","lombok","night","atr","plane","
added","london","mins",
"singapore","it","tokyo","hnd","gut","ga","crj","avod","year","aircraft","c
itilink","dps","the","doubt",
"airlines","if","from","vey","guy","coz","ve","did","due","years","dept","k
ul","because"))

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

#Replace words
docs <- tm_map(docs, gsub, pattern="helpfull",replacement="helpful")
docs <- tm_map(docs, gsub, pattern="especialy",replacement="especially")
docs <- tm_map(docs, gsub,
pattern="profesionel",replacement="professional")
docs <- tm_map(docs, gsub, pattern="hommy",replacement="homy")
docs <- tm_map(docs, gsub, pattern="foods",replacement="food")
docs <- tm_map(docs, gsub, pattern="speciall",replacement="special")
docs <- tm_map(docs, gsub, pattern="movies",replacement="movie")
docs <- tm_map(docs, gsub, pattern="prompty",replacement="prompt")
docs <- tm_map(docs, gsub, pattern="nicely",replacement="nice")
docs <- tm_map(docs, gsub, pattern="meals",replacement="meal")
docs <- tm_map(docs, gsub, pattern="seats",replacement="seat")
```

```
docs <- tm_map(docs, gsub, pattern="hours",replacement="hour")
docs <- tm_map(docs, gsub, pattern="times",replacement="time")
docs <- tm_map(docs, gsub, pattern="largest",replacement="large")
docs <- tm_map(docs, gsub, pattern="fariasi",replacement="variety")

#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)

#Find related words
v<-as.list(findAssocs(dtm, terms =c("service","food","seat","time","staff",
"entertainment","checkin","experience"),
corlimit = c(0.15,0.15,0.15,0.15,0.15,0.15,0.15,0.15)))
v

#Find related words (one by one)
v<-as.list(findAssocs(dtm, terms =c("service"),
corlimit = c(0.15)))
View(v$service)

#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 = terrain.colors(20))

termFrequency <- rowSums(as.matrix(dtm))
termFrequency <- subset(termFrequency, termFrequency>=115)

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

Lampiran 6 *Script R Visualisasi dan Asosiasi Teks (Untuk Ulasan Negatif)*

```
# 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("D://Bismillah Skripsi/Data")
docs<-readLines("Label Negatif.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("garuda","indonesia","yogyakarta","jogja","yogya","jakarta","cgk","bali",
"also","dont","evenings","nights","morning","place","stay","flew","back","m
ake","made","class",
"ill","goid","fly","flight","flying","youre","we","pass","passed","left","l
eft","flown","ontime",
"denpasar","not","does","trip","add","ons","im","kid","id","have","pee","fl
ights","told","great",
"kg","adds","kind","st","part","lombok","night","atr","plane","er","indones
ian","didn","day","days",
"added","london","mins","singapore","it","don","crew","gut","pay","ga","crj
","avod","singapore",
"year","airport","citilink","good","location","dps","the","doubt","aircraft
","found","thing",
"airlines","if","short","dps","leg","pax","long","coz","ve","did","due","ye
ars"))

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

#Replace words
docs <- tm_map(docs, gsub, pattern="hours",replacement="hour")
docs <- tm_map(docs, gsub, pattern="delayed",replacement="delay")
docs <- tm_map(docs, gsub, pattern="delays",replacement="delay")
docs <- tm_map(docs, gsub, pattern="meals",replacement="meal")
docs <- tm_map(docs, gsub, pattern="foods",replacement="food")
docs <- tm_map(docs, gsub, pattern="times",replacement="time")
docs <- tm_map(docs, gsub, pattern="passengers",replacement="passenger")
docs <- tm_map(docs, gsub, pattern="customers",replacement="customer")
```

```
docs <- tm_map(docs, gsub, pattern="prices",replacement="price")
docs <- tm_map(docs, gsub, pattern="services",replacement="service")
docs <- tm_map(docs, gsub, pattern="seats",replacement="seat")
docs <- tm_map(docs, gsub, pattern="bags",replacement="bag")

#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)

#Find related words
v<-as.list(findAssocs(dtm, terms
=c("service","staff","seat","food","costumer","checkin",
"luggage","meal","experience","passenger","gate","bag","boarding"),
corlimit =
c(0.15,0.15,0.15,0.15,0.15,0.15,0.15,0.15,0.15,0.15,0.15,0.15,0.15)))

#Find related words (one by one)
v<-as.list(findAssocs(dtm, terms =c("service"),
corlimit = c(0.15)))
v
View(v$service)

#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 = heat.colors(20))

termFrequency <- rowSums(as.matrix(dtm))
termFrequency <- subset(termFrequency, termFrequency>=17)

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

Lampiran 7 *Stopwords Berbahasa Inggris (Bouge, Kevin. 2011)*

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

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

Lampiran 8 *Hasil iterasi pada proses klasifikasi dengan NBC*

Iterasi ke-	Akurasi	Presisi	Recall	Iterasi ke-	Akurasi	Presisi	Recall
1	0.79825	0.15789	0.09091	44	0.77193	0.12000	0.09091
2	0.80702	0.07692	0.03030	45	0.84211	0.28571	0.06061
3	0.80263	0.16667	0.09091	46	0.78070	0.00000	0.00000
4	0.78947	0.00000	0.00000	47	0.77632	0.12500	0.09091
5	0.80263	0.20000	0.12121	48	0.80702	0.13333	0.06061
6	0.81579	0.23529	0.12121	49	0.75877	0.00000	0.00000
7	0.81579	0.09091	0.03030	50	0.81579	0.26316	0.15152
8	0.80702	0.00000	0.00000	51	0.82018	0.21429	0.09091
9	0.81140	0.18750	0.09091	52	0.77632	0.00000	0.00000
10	0.81140	0.14286	0.06061	53	0.79386	0.18182	0.12121
11	0.80702	0.17647	0.09091	54	0.78509	0.05556	0.03030
12	0.82895	0.20000	0.06061	55	0.81140	0.08333	0.03030
13	0.81579	0.23529	0.12121	56	0.77632	0.09091	0.06061
14	0.78947	0.14286	0.09091	57	0.79386	0.11111	0.06061
15	0.81140	0.22222	0.12121	58	0.80702	0.00000	0.00000
16	0.78947	0.10526	0.06061	59	0.80702	0.13333	0.06061
17	0.77193	0.00000	0.00000	60	0.81140	0.08333	0.03030
18	0.82456	0.23077	0.09091	61	0.77632	0.12500	0.09091
19	0.77632	0.00000	0.00000	62	0.81579	0.15385	0.06061
20	0.78947	0.05882	0.03030	63	0.80263	0.16667	0.09091
21	0.77632	0.12500	0.09091	64	0.78509	0.05556	0.03030
22	0.82018	0.16667	0.06061	65	0.79825	0.06667	0.03030
23	0.79825	0.19048	0.12121	66	0.79386	0.00000	0.00000
24	0.79386	0.11111	0.06061	67	0.82018	0.31818	0.21212
25	0.77193	0.08696	0.06061	68	0.79386	0.18182	0.12121
26	0.80263	0.16667	0.09091	69	0.82456	0.26667	0.12121
27	0.79825	0.06667	0.03030	70	0.78947	0.05882	0.03030
28	0.78070	0.05263	0.03030	71	0.81140	0.22222	0.12121
29	0.78509	0.05556	0.03030	72	0.80702	0.13333	0.06061
30	0.80263	0.12500	0.06061	73	0.82456	0.26667	0.12121
31	0.79825	0.11765	0.06061	74	0.82018	0.16667	0.06061
32	0.81579	0.23529	0.12121	75	0.78947	0.00000	0.00000
33	0.78947	0.10526	0.06061	76	0.78509	0.10000	0.06061
34	0.79825	0.00000	0.00000	77	0.79825	0.11765	0.06061
35	0.82895	0.31250	0.15152	78	0.82895	0.31250	0.15152
36	0.80263	0.20000	0.12121	79	0.75439	0.07407	0.06061
37	0.77193	0.04762	0.03030	80	0.79825	0.11765	0.06061
38	0.79386	0.15000	0.09091	81	0.76754	0.04545	0.03030
39	0.78947	0.10526	0.06061	82	0.82895	0.28571	0.12121
40	0.80702	0.07692	0.03030	83	0.80263	0.12500	0.06061
41	0.79825	0.06667	0.03030	84	0.80702	0.13333	0.06061
42	0.80263	0.12500	0.06061	85	0.79386	0.15000	0.09091
43	0.79825	0.19048	0.12121	86	0.81579	0.09091	0.03030

Iterasi ke-	Akurasi	Presisi	Recall
87	0.78509	0.05556	0.03030
88	0.77193	0.12000	0.09091
89	0.77632	0.00000	0.00000
90	0.79386	0.11111	0.06061
91	0.79825	0.11765	0.06061
92	0.81140	0.08333	0.03030
93	0.78509	0.05556	0.03030
94	0.78070	0.09524	0.06061
95	0.82018	0.25000	0.12121
96	0.81579	0.20000	0.09091
97	0.81140	0.22222	0.12121
98	0.78509	0.00000	0.00000
99	0.80263	0.16667	0.09091
100	0.80702	0.21053	0.12121