BAB 2
LANDASAN TEORI

2.1. Text Mining

Text mining adalah suatu proses yang bertujuan untuk menemukan informasi atau tren terbaru yang belum pernah terungkap dengan mengolah dan menganalisa data dalam jumlah besar [7]. Text mining dapat memberikan solusi untuk permasalahan seperti pengelompokan atau pengorganisasian, menganalisa *unstructured text* dalam jumlah besar, atau pemrosesan. Pada dasarnya text mining adalah bidang interdisiplin yang berhubungan dengan perolehan informasi (*information reveral*), data mining, pembelajaran mesin (*machine learning*), statistik, dan komputasi linguistik. Dalam menganalisa sebagian atau keseluruhan teks tidak terstruktur, text mining akan mengaitkan satu bagian teks dengan bagian yang lainnya menurut aturan-aturan tertentu. Text mining umumnya mencakup klasifikasi informasi atau teks, ekstraksi entitas atau konsep, pengembangan dan perumusan taksa umum, dan kelompok teks. Selain itu, text mining juga diartikan sebagai kegiatan mengekstraksi data dari data yang berupa teks atau dokumen, dengan tujuan menemukan kata-kata yang dapat mewakili apa yang ada dalam dokumen sehingga dapat dilakukan analisis hubungan dalam text mining [8].

2.2. Analisis Sentimen

Analisis sentimen atau disebut juga *opinion mining* adalah bidang studi yang menganalisis opini, sentimen, penilaian, sikap, dan emosi seseorang terhadap suatu entitas dan atribut yang dinyatakan dalam sebuah teks tertulis. Entitas dapat berupa produk, jasa, organisasi, individu, peristiwa, isu, atau topik [9]. Opini berbeda dengan informasi faktual yang bersifat objektif sedangkan opini dan sentimen bersifat subjektif. Pemeriksaan berbagai opini dari banyaknya pihak sangat diperlukan agar mendapatkan pandangan subjektif yang berasal lebih dari satu orang, sehingga diperlukan ringkasan untuk mewakili suatu opini. Analisis sentimen dibagi menjadi tiga tingkat, yaitu level dokumen, level kalimat,

dan level aspek dan entitas. Untuk mengklasifikasikan teks opini pada tingkat dokumen atau pada tingkat kalimat sebagai positif atau negatif tidak cukup untuk sebagian besar aplikasi, karena klasifikasi tersebut tidak mengidentifikasi sentimen, target opini, ataupun menetapkan sentimen pada sebuah target.

2.4.1. Analisis Sentimen Berbasis Aspek

Analisis sentimen berbasis aspek adalah proses mendapatkan informasi suatu sentimen dari sudut pandang tertentu mengenai aspek yang dibahas. Analisis sentimen berbasis aspek juga merupakan pengembangan dari analisis sentimen yang mengacu pada kalimat. Untuk mendapatkan aspek-aspek yang akan digunakan dalam analisis sentimen, dibutuhkan *aspect sentence labelling* agar kata yang memiliki kategori yang sama dapat ditentukan dan teridentifikasi aspek yang relevan dengan kata tersebut [10]. Terdapat dua tugas (*task*) yang paling banyak mendapat perhatian penelitian, yaitu *aspect extraction* dan *aspect sentiment classification*.

1. *Aspect Extraction*

   Tujuan tahap ini adalah untuk mengekstrak aspek dan entitas yang telah dievaluasi.

2. *Aspect Sentiment Classification*

   Tugas ini menentukan pendapat pada aspek yang berbeda termasuk ke dalam sentimen positif, negatif, atau netral.

Metode untuk melakukan analisis sentimen berbasis aspek terbagi menjadi dua, yaitu *Supervised Learning* dan *Lexicon-Based*. Terdapat dua pendekatan utama pada *Supervised Learning*. Pendekatan pertama adalah untuk menghasilkan satu set fitur yang bergantung pada identitas target atau aspek dalam kalimat. Pendekatan kedua adalah menentukan ruang lingkup penerapan setiap ekspresi sentimen untuk menentukan apakan mencakup entitas target atau aspek dalam kalimat.

2.3. Preprocessing

Data yang telah didapatkan dalam tahap pengumpulan data, kemudian dilanjutkan dengan tahapan preprocessing yaitu melakukan pembersihan data dengan beberapa tahap agar menjadi data yang siap diolah oleh *machine learning* [4]. Data yang diolah umumnya memiliki beberapa karakteristik, berdimensi tinggi, terdapat *noise,* dan berstruktur tidak baik. Tahapan dalam preprocessing adalah sebagai berikut:

2.4.1. Case Folding

Case folding yaitu proses menyeragamkan bentuk huruf menjadi *lowercase* atau *uppercase* [11].

2.4.2. Cleaning

Proses cleaning bertujuan untuk membersihkan data dari hal yang tidak diperlukan seperti simbol, *emoticon*, angka, dan tanda baca [12].

2.4.3. Tokenization

Tokenization adalah tahapan untuk memisahkan kalimat menjadi bagian-bagian kata yang disebut token [2]. Karakter-karakter yang menjadi pemisah kata seperti *whitespace* akan dihilangkan karena tidak memiliki pengaruh terhadap pemrosesan teks.

2.4.4. Spelling Normalization

Spelling Normalization merupakan proses perbaikan kata-kata yang salah eja atau disingkat dalam bentuk tertentu. Tahap ini bertujuan untuk memperkecil dimensi kata yang memiliki arti yang sama tetapi memiliki ejaan yang salah atau disingkat dalam bentuk tertentu [13].

2.4.5. Filtering

Filtering merupakan sebuah proses membuang kata-kata yang tidak berguna dalam proses klasifikasi [5]. Proses Filtering dapat dilakukan dengan dua cara yaitu filtering berdasarkan *wordlist* atau *stoplist*. Tahapan ini menggunakan algoritma *stoplist*. *Stoplist* berisi sekumpulan kata yang tidak memiliki makna atau biasa disebut dengan *stopword*. Dalam

penerapannya, penggunaan *stoplist* lebih efektif daripada *wordlist*. Karena banyaknya kata yang tidak penting berjumlah lebih sedikit dari kata yang penting.

2.4.6. Stemming

Tujuan dari proses stemming adalah mengganti kata yang masih berimbuhan menjadi kata dasar dengan menghilangkan semua imbuhan kata (affixes) meliputi awalan kata (prefixes), akhiran kata (suffixes), atau menghilangkan awalan kata dan akhiran kata (confixes) pada kata turunan [2]. Library yang akan digunakan pada proses stemming adalah Sastrawi.

2.4. N-Gram

N-Gram adalah sekumpulan n-kata yang muncul dalam urutan pada kalimat atau kumpulan teks [14]. Pemodelan N-Gram adalah pendekatan identifikasi dan analisis fitur yang popular digunakan dalam pemodelan bahasa dan bidang pemrosesan bahasa alami. *Characters* dan *Word* merupakan model N-Gram yang paling banyak digunakan dalam ketegori teks. N-Gram dalam analisis sentimen membantu menganalisis sentimen teks atau dokumen [6]. Karakteristik pada N-Gram adalah sebagai berikut:

1. Walaupun terdapat kesalahan tekstual, N-Gram masih dapat berfungsi dengan baik.
2. Membutuhkan penyimpanan yang sederhana dan berjalan dengan efisien.
3. Waktu proses yang dibutuhkan relatif cepat.

Dalam bahasa Indonesia terdapat banyak frase yang tidak hanya terdiri dari satu kata oleh karena itu N-Gram sangat dibutuhkan [15]. N-Gram yang umum digunakan adalah Unigram, Bigram, dan Trigram. Unigram adalah token yang terdiri dari satu kata, Bigram merupakan token yang terdiri dari dua kata, dan Trigram adalah token yang terdiri dari tiga kata. Contoh Unigram, Bigram, dan Trigram adalah sebagai berikut:

1. Unigram

   {“belum”,”keluar”,”kamar”, ”sudah”,”dibersihkan”,”tanpa”,”permisi”}

2. Bigram
   {”belum_keluar”,”keluar_sudah”,”sudah_dibersihkan”,”dibersihkan_tanpa”,”tanpa_permisi”}
3. Trigram
   {“belum_keluar_kamar”,”keluar_kamar_sudah”,”kamar_sudah_dibersihkan”,”sudah_dibersihkan_tanpa”,”dibersihkan_tanpa_permisi”}

2.5. Pembobotan TF

Pembobotan TF adalah (*Term Frequency)* merupakan salah satu pembobotan yang paling sering digunakan [1]. *Term Frequency* adalah banyaknya jumlah kata atau *term* tertentu yang ada pada suatu dokumen. Semakin besar jumlah kemunculan suatu kata atau *term* dalam dokumen, akan semakin besar pula bobot atau akan memberikan nilai kesesuaian yang besar. Rumus TF adalah sebagai berikut:

$$TF(t_k, d_j) = f(t_k, d_j) \tag{2.1}$$

Keterangan:

$f(t_k, d_j)$ = jumlah kemunculan *term* k pada sebuah dokumen j

2.6. Query Expansion Ranking

Query Expansion Ranking merupakan sebuah seleksi fitur yang bertujuan untuk mengurangi kompleksitas komputasi tanpa mengurangi kualitas dari analisis sentimen [5]. Query Expansion Ranking mulanya terinspirasi dari metode Query Expansion yang bekerja dengan cara probabilistic weighting model untuk memberi skor pada setiap fitur yang berfungsi untuk meningkatkan kualitas query yang dimasukkan oleh pengguna. Persamaan dari QER adalah sebagai berikut:

$$Score = \frac{|p_f + q_f|}{|p_f - q_f|} \tag{2.2}$$

Score$_f$ adalah skor atau nilai QER, p_f adalah peluang fitur *f* dalam dokumen kelas positif, q_f adalah peluang fitur dalam dokumen kelas negatif [16]. Nilai-nilai di atas dihitung berdasarkan dua persamaan di bawah ini.

$$p_f = \frac{df_+^f + 0.5}{n^+ + 1.0} \tag{2.3}$$

$$q_f = \frac{df_-^f + 0.5}{n^- + 0.5} \tag{2.4}$$

df_+^f adalah jumlah dokumen positif yang mengandung fitur *f*, df_-^f adalah jumlah dokumen negatif yang mengandung fitur *f*, n^+ adalah jumlah dokumen positif, n^- adalah jumlah dokumen negatif.

2.7. Naïve Bayes

Naïve Bayes merupakan metode *supervised learning* yang sederhana dan banyak digunakan. Naïve Bayes juga merupakan salah satu algoritma dengan pembelajaran tercepat dan dapat menangani banyak *feature* atau *class*. Terlepas dari model yang sederhana, Naïve Bayes cocok untuk menyelesaikan setiap masalah.

Bayes adalah teknik prediksi berbasis probabilistik sederhana berdasarkan pada penerapan teorema Bayes dengan asumsi independensi (ketidaktergantungan) yang kuat [17]. Pada Teorema Bayes, independensi yang dimaksud adalah fitur pada sebuah data tidak ada kaitannya dengan ada atau tidaknya fitur lain dalam data yang sama.

Prediksi Bayes berdasarkan pada teorema Bayes dengan persamaan umum sebagai berikut.

$$\boldsymbol{P(Y|X) = \frac{P(X|Y) \times P(Y)}{P(X)}} \tag{2.5}$$

Keterangan:

X : Data dengan *class* yang belum diketahui

Y : Hipotesis data X merupakan suatu kelas spesifik

P(Y|X) : Probabilitas akhir bersyarat (*conditional probability*) atau suatu hipotesis Y terjadi jika diberikan bukti X (*evidence*) terjadi.

P(X|Y) : Probabilitas sebuah bukti X akan mempengaruhi hipotesis Y

P(Y) : Probabilitas awal (priori) hipotesis y terjadi tanpa memandang bukti apapun.

P(X) : Probabilitas awal (priori) bukti X terjadi tanpa memandang bukti atau hipotesis yang lain.

Ide dasar dari aturan Bayes merupakan hasil dari hipotesis atau peristiwa (H) yang dapat diperkirakan berdasarkan beberapa bukti (E) yang diamati. Ada beberapa hal penting yang harus diperhatikan dari aturan Bayes, yaitu sebagai berikut:

1. Sebuah probabilitas awal atau priosi H atau P(H) adalah probabilitas dari suatu hipotesis sebelum bukti diamati.
2. Sebuah probabilitas akhir H atau P(H|E) adalah probabilitas dari suatu hipotesis setelah bukti diamati.

Bergantung pada situasi yang tepat dari model probabilitas, Naïve Bayes dapat dilatih dengan sangat efektif dalam *supervised learning*. Dalam aplikasi praktis, estimasi parameter untuk model Naïve Bayes menggunakan metode *likelihood* maksimum, artinya seseorang dapat bekerja dengan model Naïve Bayes tanpa mempercayai probabilitas Bayesian atau menggunakan metode Bayesian yang lain.

Kelebihan dari Naïve Bayes adalah metode ini hanya membutuhkan sedikit data latih untuk menentukan estimasi parameter yang diperlukan untuk proses klasifikasi. Karena diasumsikan sebagai variabel independen, hanya varian dari variabel dalam suatu kelas yang diperlukan untuk menentukan pengklasifikasian, bukan seluruh dari matriks kovarians [12].

2.8. Naïve Bayes untuk Klasifikasi

Hubungan antara Naïve Bayes dan klasifikasi, korelasi hipotesis dan bukti dengan klasifikasi adalah bahwa hipotesis dalam teorema Bayes adalah label kelas yang menjadi target pemetaan pada klasifikasi, sedangkan bukti merupakan fitur-fitur yang menjadi masukkan dalam model klasifikasi [18]. Jika X adalah vektor masukkan yang berisi fitur dan Y adalah label kelas, Naïve Bayes dituliskan P(Y|X). Notasi tersebut berarti probabilitas label kelas Y didapatkan setelah fitur-fitur X diamati. Notasi ini disebut juga probabilitas akhir (*posterior probability*) untuk Y, sedangkan P(Y) disebut probabilitas awal (*prior probability*) Y.

Formulasi Naïve Bayes untuk klasifikasi adalah:

$$\boldsymbol{P(Y|X)} = \frac{\boldsymbol{P(Y)\prod_{i=1}^{q} P(X_i|Y)}}{\boldsymbol{P(X)}} \tag{2.6}$$

Keterangan:

P(X|Y) = Probabilitas data dengan vector X pada kelas Y.

P(Y) = Probabilitas kelas awal Y.

$\prod_{i=1}^{q} P(X_i|Y)$ = Probabilitas independen kelas Y dari semua fitur dalam vektor X.

Nilai P(X) selalu tetap sehingga dalam perhitungan prediksi nantinya hanya tinggal menghitung bagian $P(Y)\prod_{i=1}^{q} P(X_i|Y)$ dengan memilih yang terbesar sebagai kelas yang dipilih sebagai hasil prediksi. Sementara probabilitas independen $\prod_{i=1}^{q} P(X_i|Y)$ tersebut merupakan pengaruh semua fitur dari data terhadap setiap kelas Y, yang dinotasikan dengan :

$$\boldsymbol{P(Y|X = y) = \prod_{i=1}^{q} P(X_i|Y) = y} \tag{2.7}$$

Setiap set fitur $X = \{X_1, X_2, X_3, \ldots, X_q\}$ terdiri atas q atribut (q dimensi).

Secara umum, Bayes mudah untuk menghitung fitur bertipe kategoris seperti klasifikasi. Akan tetapi untuk fitur numerik (kontinyu) ada perlakuan khusus sebelum dimasukkan pada Naïve Bayes [19], yaitu:

1. Melakukan diskritisasi pada setiap fitur kontinyu dengan mengganti nilai fitur kontinyu tersebut dengan nilai interval diskret. Pendekatan ini dilakukan dengan mentransformasikan fitur kontinu ke dalam fitur ordinal.
2. Mengasumsikan bentuk tertentu dari distribusi probabilitas untuk fitur kontinyu dan memperkirakan parameter distribusi dengan data pelatihan. Distribusi Gaussian biasanya dipilih untuk mempresentasikan probabilitas bersyarat dari fitur kontinyu pada sebuah kelas, sedangkan distribusi Gaussian dikarakteristikan dengan dua parameter mean, μ, varian, σ^2. Untuk setiap kelas *yj*, probabilitas bersyarat kelas y_j untuk fitur X_i adalah:

$$\boldsymbol{P(X_i = x_i \mid Y = y_j = \frac{1}{\sqrt{2\pi\sigma_{ij}}}\, exp^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}}} \qquad (2.8)$$

Parameter μ_{ij} bisa diestimasi berdasarkan sampel mean *Xi* ($\bar{x}$) untuk semua data latih yang menjadi kelas y_j, sedangkan σ_{ij}^2 dapat diperkirakan dari varian sampel (s^2) dari data latih [18].

Sementara itu rumus yang digunakan dalam Naïve Bayes klasifikasi teks dengan pembobotan kata TF untuk menghitung prababilitas kata yang akan muncul pada salah satu kelas adalah sebagai berikut:

$$P(w_k \mid c_i) = \frac{(n_{wk})c_i+1}{n_{ci}+n_k} \qquad (2.9)$$

Keterangan :

$P(w_k \mid c_i)$ = Peluang kemunculan w_k pada kategori c_i

w_k = Kata pada data latih yang dicari peluang kemunculannya

c_i = Kategori yang ada pada data

$(n_{wk})c_i$ = Jumlah kemunculan kata w_k pada kelas/kategori c_i

n_{ci} = Total kata yang ada pada kelas c_i

n_k = Total kata yang ada pada data latih

Pada saat tahap pengujian apabila terdapat kata yang belum muncul pada data latih maka nilai $P(c_i)$ = 1.

2.9. Confusion Matrix

Confusion Matrix berisi informasi tentang performa dari suatu sistem classifier yang dievaluasi menggunakan data atau matrix yang terdapat dalam Confusion Matrix. Confusion Matrix menganalisis kualitas klasifikasi yang telah dilakukan pada kelas aktual maupun kelas hasil prediksi [1]. Pengukuran yang diterapkan pada Confusion Matrix adalah menghitung *accuracy*, *precision*, *recall*, *f-measure* yang mengacu pada nilai *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) yang merupakan nilai keluaran dari Confusion Matrix. Parameter performa klasifikasi mengacu pada hasil accuracy apabila selihih tipis antara nilai FP dan FN [20].

Tabel 2. 1 Confusion Matrix

Confusion Matrix		Prediksi	
		Positif	Negatif
Aktual	Positif	TP	FN
	Negatif	FP	TN

Keterangan :

TP (*True Positive*) : Hasil prediksi bernilai True dan pengujian menghasilkan True

FP (*False Positive*) : Hasil prediksi bernilai False dan pengujian menghasilkan True

FN (*False Negative*) : Hasil prediksi bernilai True dan pengujian menghasilkan False

TN (*True Negative*) : Hasil prediksi bernilai False dan pengujian menghasilkan False

Confusion Matrix menunjukan tingkat akurasi dari proses klasifikasi yang telah dilakukan. Tingkat akurasi menunjukan proporsi jumlah prediksi benar.

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2.10)$$

Precision adalah proporsi dari pelabelan yang teridentifikasi dengan benar, rumus untuk mencari *precision* adalah:

$$Precision = \frac{TP}{TP+FP} \quad (2.11)$$

Recall merupakan proporsi dari informasi yang dapat ditemukan dari label, rumus yang dapat digunakan untuk mencari *recall* adalah:

$$Recall = \frac{TP}{TP+FN} \quad (2.12)$$

Precision dan *Recall* dapat digunakan untuk mendapatkan proporsi pengukuran lain yaitu *F1-Score*. *F1-Score* merupakan *harmonic mean* dari perhitungan *Precision* dan *Recall*, rumus untuk mencari *F1-Score* adalah:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.13)$$

2.10. Web Scraping

Web scraping adalah cara yang digunakan untuk mendapatkan data atau informasi dari suatu website yang dilakukan secara otomatis. Tujuan dari web scraping adalah untuk menggali informasi dari situs web yang berbeda dan tidak terstruktur kemudian mengubahnya menjadi bentuk yang lebih rapi dan terstruktur dalam bentuk basis data, spreadsheets, atau Comma Separated Values (CSV) [2].

2.11. Penelitian-Penelitian Terkait

Tabel 2.2 Penelitian-Penelitian Terkait

Review Literatur Pertama [1]	
Judul Artikel	Analisis Sentimen Berbasis Aspek Ulasan Pelanggan Terhadap Kertanegara Premium Guest House Menggunakan Support Vector Machine
Penulis	Wirdhayanti Paulina, Fitra Abdurrachman Bachtiar, Alfi Nur Rusydi
Judul Jurnal/Proceeding	Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer
Tahun Penerbitan	2020
Masalah Utama yang diangkat	Menganalisis sentimen terhadap ulasan Kertanegara di OTA karena Kertanegara hanya berfokus pada ulasan *Guest Review*. Dengan menggunakan metode *Support Verctor Machine* (SVM) dan *Term Weighting* (TF-IDF).
Metode Ekstraksi	Menggunakan TF-IDF
Metode Klasifikasi	Support Vector Machine
Hasil Penelitian dan Kesimpulan	Hasil : Visualisasi dari hasil klasifikasi sentimen ditampilkan melalui dashboard dengan menggunakan tool Metabase yang menyajikan 6 komponen informasi diantaranya, informasi mengenai jumlah ulasan yang digunakan untuk analisis sentimen, bar chart yang menunjukkan jumlah ulasan pelanggan berdasarkan terhadap semua aspek yang telah ditentukan, line chart untuk menunjukkan tren jumlah ulasan pelanggan terhadap semua aspek yang telah ditentukan berdasarkan kelas sentimen pada kurun waktu tertentu, pie chart menunjukkan jumlah sentimen pada setiap aspek yang dipilih, yaitu line chart menunjukkan tren jumlah sentimen untuk setiap aspek yang dipilih pada

	kurun waktu tertentu, tabel yang menampilkan nama, tanggal, dan isi ulasan berdasarkan aspek dan sentimen yang dipilih. Kesimpulan : - Penggunaan metode Support Vector Machine yang digunakan dengan pembobotan TF-IDF dapat menjadi salah satu cara untuk menyelesaikan permasalahan klasifikasi dalam analisis sentimen. - Pengujian hasil klasifikasi sentimen memiliki rata-rata yang baik dengan nilai Accuracy, Precision. Recall, dan F1-Score diatas 70%. - Hasil visualisasi dashboard dengan tools Metabase menghasilkan nilai 67,5 pada Usability Testing menggunakan metode System Usability Scale (SUS). Dashboard memiliki tingkat Acceptable dengan rating Good yang berarti dashboard dapat diterima dengan baik oleh pihak Kertanegara Premium Guest House.
Review Literatur Kedua [2]	
Judul Artikel	Analisis Sentimen Berbasis Aspek pada Ulasan Pelanggan Restoran Bakso President Malang dengan Metode Naïve Bayes Classifier
Penulis	Whita Parasati, Fitra Abdurrachman Bachtiar, Nanang Yudi Setiawan
Judul Jurnal/Proceeding	Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer
Tahun Penerbitan	2020
Masalah Utama yang diangkat	Pihak Bakso President Malang tidak memiliki data opini pelanggan, maupun penerapan teknologi dalam

	mengolah dan menganalisis data yang dapat menghasilkan informasi tentang perspektif pelanggan terhadap aspek kepuasan pelanggan. Metode yang digunakan yaitu dengan algoritma Naïve Bayes untuk mengklasifikasikan ulasan pada level aspek.
Metode Ekstraksi	Web Scraping pada situs TripAdvisor dan Google Review.
Metode Klasifikasi	Naïve Bayes
Hasil Penelitian dan Kesimpulan	Hasil : Analisis sentimen dalam setiap aspek menghasilkan nilai akurasi sebesar 88% pada aspek Makanan, 76% pada aspek Layanan, dan 84% pada aspek Atmosfer. Kesimpulan : Penggunaan Naïve Bayes Classifier dan TF-IDF dapat dijadikan opsi dalam melakukan analisis sentimen level aspek untuk penilaian kepuasan pelanggan Bakso President Malang.
Review Literatur Ketiga [10]	
Judul Artikel	ANALISIS SENTIMEN MULTI-ASPEK BERBASIS KONVERSI IKON EMOSI DENGAN ALGORITME NAÏVE BAYES UNTUK ULASAN WISATA KULINER PADA WEB TRIPADVISOR
Penulis	Sitti Aliyah Azzahra, Arief Wibowo
Judul Jurnal/Proceeding	Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)
Tahun Penerbitan	2020
Masalah Utama yang diangkat	Metode Naïve Bayes yang digunakan untuk analisis sentimen pada komentar atau ulasan yang mengandung emoticon pada situs TripAdvisor untuk mengetahui nilai sentimen dari suatu objek wisata yang diulas.

Metode Ekstraksi	Term Frequency Murni (Raw-TF).
Metode Klasifikasi	NAÏVE BAYES
Hasil Penelitian dan Kesimpulan	Hasil : Hasil pengujian menunjukkan bahwa penggunaan seluruh kombinasi metode tersebut dalam proses klasifikasi sentimen mampu menghasilkan nilai akurasi sebesar 98,67%. Kesimpulan : Dari semua proses rangkaian kegiatan penelitian di atas yaitu analisis sentimen yang dilakukan menggunakan metode Naïve Bayes Classifier dengan penambahan fitur konversi emoticon dan Multi- Aspect Sentence Labeling mendapatkan nilai akurasi sebesar 98.67% yang sebelumnya hanya sebesar 88.78%, sehingga dapat disimpulkan bahwa penggunaan kombinasi metode tersebut telah terbukti memperbaiki hasil penelitian sebelumnya
Review Literatur Keempat [21]	
Judul Artikel	Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes
Penulis	Billy Gunawan, Helen Sasty Pratiwi, Enda Esyudha Pratama
Judul Jurnal/Proceeding	Jurnal Edukasi dan Penelitian Informatika (JEPIN)
Tahun Penerbitan	2018
Masalah Utama yang diangkat	Naïve Bayes yang digunakan sebagai metode untuk sistem analisis sentimen terhadap ulasan produk
Metode Ekstraksi	TF_IDF
Metode Klasifikasi	Naïve Bayes
Hasil Penelitian dan	Hasil : Metode Naïve Bayes dapat memprediksi kelas

Kesimpulan	sentimen pada ulasan produk online sesuai dengan sistem yang disiapkan. Kesimpulan : Nilai akurasi pengujian pada 3 kelas menggunakan dataset 90% sebagai data latih dan 10% sebagai data uji mempunyai nilai 77,78%. Nilai tersebut lebih besar dibandingkan dengan pengujian 5 kelas menggunakan dataset 80% data latih dan 20% data uji menghasilkan nilai 52,66%.
Review Literatur Kelima [22]	
Judul Artikel	Pemilihan Fitur Pada Analisis Sentimen Review Travel Online Menggunakan Algoritma Naïve Bayes Dalam Penerapan Mutual Information Dan Particle Swarm Optimization (PSO)
Penulis	Lisda Widiastuti
Judul Jurnal/Proceeding	IJCIT (Indonesian Journal on Computer and Information Technology)
Tahun Penerbitan	2018
Masalah Utama yang diangkat	Penggunaan algoritma Naïve Bayes dengan mutual information dan *Particle Swarm Optimization* sebagai fitur pilihan dan algoritma Naïve Bayes tanpa fitur seleksi, dapat diterapkan pada prediksi travel review online.
Metode Ekstraksi	*Partikel Swarm Optimization*
Metode Klasifikasi	Naïve Bayes
Hasil Penelitian dan Kesimpulan	Hasil : Nilai akurasi terbaik didapat dengan jumlah 200 data review berdasarkan dari jumlah tersebut menghasilkan nilai akurasi 93.50% dengan nilai AUC 0.965. Peningkatan akurasi sebesar 12% dari nilai akurasi 81.50% yang hanya menggunakan algoritma

	naïve bayes saja. Kesimpulan : Prediksi review travel online dengan menggunakan algoritma naïve bayes dan fitur seleksi lebih unggul dibandingkan menggunakan algoritma naïve bayes tanpa fitur seleksi
Review Literatur Keenam [23]	
Judul Artikel	Analisis Sentimen dengan Naïve Bayes Terhadap Komentar Aplikasi Tokopedia
Penulis	Rita Aprian, Dudih Gustian
Judul Jurnal/Proceeding	Jurnal Rekayasa Teknologi Nusa Putra
Tahun Penerbitan	2019
Masalah Utama yang diangkat	Metode Naïve Bayes yang digunakan untuk menganalisis sentimen pada komentar di aplikasi Tokopedia.
Metode Ekstraksi	-
Metode Klasifikasi	Naïve Bayes
Hasil Penelitian dan Kesimpulan	Hasil : Performa yang dihasilkan dari hasil pengujian yang dilakukan Rapidminer terhadap 1.500 data testing dihasilkan nilai akurasi sebesar 97,13%, dengan nilai precision 1. Kesimpulan : Metode Naive Bayes terbukti dapat menganalisis sentimen secara otomatis
Review Literatur Ketujuh [24]	
Judul Artikel	Penerapan Analisis Sentimen untuk Menilai Suatu Produk pada Twitter Berbahasa Indonesia dengan Metode Naïve Bayes Classifier dan Information Gain
Penulis	Ahmad Wildan Attabi', Lailil Muflikhah, Mochammad

	Ali Fauzi
Judul Jurnal/Proceeding	Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer
Tahun Penerbitan	2018
Masalah Utama yang diangkat	Metode Naïve Bayes yang dioptimalkan dengan Information Gain untuk analisis sentimen terhadap penilaian suatu produk di Twitter.
Metode Ekstraksi	Information Gain
Metode Klasifikasi	Naïve Bayes
Hasil Penelitian dan Kesimpulan	Hasil : menghasilkan akurasi 74% apabila menambahkan Information Gain untuk mengoptimalkan hasil klasifikasi. Sedangkan, akurasi yang didapatkan tanpa Information Gain adalah 70% Kesimpulan : Information Gain berhasil meningkatkan akurasi pada klasifikasi dengan menggunakan metode Naïve Bayes
Review Literatur Kedelapan[6]	
Judul Artikel	Penerapan Word N-Gram untuk Sentiment Analysis Review Menggunakan Metode Support Vector Machine (Studi Kasus: Aplikasi Sambara)
Penulis	Fitriyani, Toni Arifin
Judul Jurnal/Proceeding	SISTEMASI : Jurnal Sistem Informasi
Tahun Penerbitan	2020
Masalah Utama yang diangkat	Aplikasi Sambara diharapkan memberikan efesiensi, efektifitas, dan perbaikan pelayanan. Keberhasilan aplikasi dapat diketahui dengan melakukan analysis sentiment review.
Metode Ekstraksi	Word N-Gram

Metode Klasifikasi	Support Vector Machine (SVM)
Hasil Penelitian dan Kesimpulan	Hasil : Nilai akurasi Bi-Gram dengan jumlah data 1.332 menghasilkan nilai akurasi lebih tinggi dibandingan dengan nilai akurasi yang dihasilkan Unigram. Nilai akurasi Unigram sebesar 88.21% sedangankan Bigram 88.59%. Penerapan Trigram menghasilkan nilai akurasi terbesar dari jumlah data 1.200 dengan nilai akurasi 88.50% Kesimpulan : Nilai akurasi dengan kenaikan tertinggi yaitu pada penerapan Trigram dengan jumlah data 1.200. Kenaikan nilai akurasi sebesar 0.92% dibandingkan dengan Unigram menjadi 88.59% dengan nilai
Review Literatur Kesembilan [25]	
Judul Artikel	Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes
Penulis	Clarisa Hasya Yutika, Adiwijaya, Said Al Faraby Fakultas
Judul Jurnal/Proceeding	JURNAL MEDIA INFORMATIKA BUDIDARMA
Tahun Penerbitan	2021
Masalah Utama yang diangkat	Menggunakan TF-IDF dan Naïve Bayes pada analisis sentimen berbasis aspek untuk review Female Daily dan review menggunakan Bahasa multilingual.
Metode Ekstraksi	TF-IDF
Metode Klasifikasi	Naïve Bayes
Hasil Penelitian dan Kesimpulan	Hasil : Dari ketiga skenario yang telah dilakukan, didapat performansi tertinggi oleh dataset diterjemahkan ke dalam Bahasa Inggris kemudian

	diterjemahkan ke Bahasa Indonesia dan tidak menggunakan stopword removal dengan parameter alpha atau smoothing sebesar 1, min_df sebesar 0,01, max_df sebesar 0,7, dan max_features sebesar 2000 menghasilkan performansi terbaik dengan nilai F1-Score sebesar 62,81%. Kesimpulan : data yang tidak seimbang dapat mempengaruhi performansi.
Review Literatur Kesepuluh [14]	
Judul Artikel	Klasifikasi Sentimen pada Twitter Terhadap WHO Terkait Covid-19 Menggunakan SVM, N-Gram, PSO
Penulis	Noor Hafidz, Dewi Yanti Liliana
Judul Jurnal/Proceeding	JURNAL RESTI
Tahun Penerbitan	2021
Masalah Utama yang diangkat	Dari twitter dapat disarikan sentimen dan opini masyarakat dunia terhadap berbagai isu termasuk opini masyarakat dunia terhadap penanganan Covid-19 oleh WHO yang dikenal dengan analisis sentimen.
Metode Ekstraksi	TF-IDF
Metode Klasifikasi	SVM
Hasil Penelitian dan Kesimpulan	Hasil : Pada SVM, penambahan teknik N-gram memberikan efek yang merugikan dalam hal akurasi. Nilai akurasi berkurang dari 0,762 menjadi 0,752 untuk SVM dengan Bigram dan 0,709 untuk SVM dengan Trigram. Kesimpulan : Dengan model CRISP-DM dan membandingkan metode klasifikasi Naïve Bayes (NB),

	Support Vector Machine (SVM), dan k-Nearest Neighbor (k-NN), terbukti bahwa model klasifikasi SVM menunjukkan hasil terbaik. Rata-rata accuracy SVM sebesar 96.43% dilihat dari empat aspek, yaitu aspek desain sebesar 94.40%, aspek harga sebesar 97.44%, aspek spesifikasi sebesar 96.22%, dan aspek citra merk sebesar 97.63%.