BAB II

LANDASAN TEORI

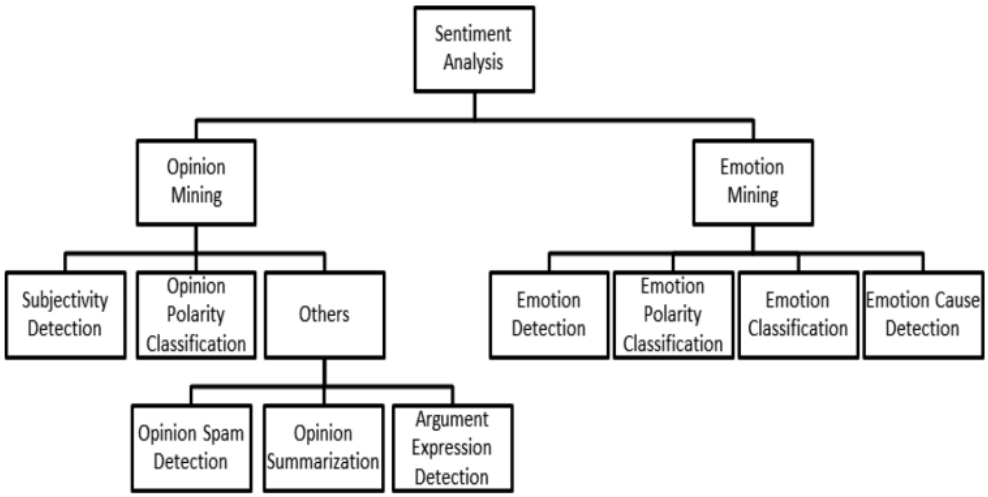
2.1 Analisa Sentimen

Ketertarikan pada bidang analisis sentimen belakangan ini meningkat baik untuk akademisi maupun industri karena banyaknya data yang tersedia di sosial media (Cieliebak et al., 2017). *Sentiment analysis* atau analisis sentimen dalam bahasa indonesia adalah sebuah teknik atau cara yang digunakan untuk mengidentifikasi bagaimana sebuah sentimen diekspresikan menggunakan teks dan bagaimana sentimen tersebut bisa dikategorikan sebagai sentimen positif maupun sentimen negatif (Muchammad Shiddieqy Hadna et al., 2016). Pendapat serupa dikemukakan oleh (Medhat et al., 2014) dimana *Sentiment Analysis* atau *Opinion Mining* merupakan sebuah studi yang mempelajari mengenai pendapat, sikap, dan emosi seseorang terhadap suatu entitas yang dapat mewakili suatu individu, acara atau suatu topik tertentu.

Analisis sentimen mencakup deteksi, analisis, dan evaluasi keadaan pikiran manusia terhadap berbagai peristiwa, masalah, layanan atau minat lainnya (Yadollahi et al., 2017). Lebih tepatnya, bidang ini bertujuan untuk menggali pendapat, sentimen dan emosi berdasarkan pengamatan orang-orang yang bisa didapatkan melalui tulisan, ekspresi wajah, ucapan, musik, gerakan, dan lain sebagainya (Yadollahi et al., 2017). Tujuan dari analisis sentimen sendiri adalah untuk menemukan pendapat, mengidentifikasi sentimen yang mereka ungkapkan, dan kemudian mengklasifikasikan polaritasnya (Medhat et al., 2014). Dengan kata

lain, analisis sentimen berfungsi untuk mengklasifikasikan teks kedalam kelas positif, negatif atau netral (Cieliebak et al., 2017). Beberapa pendapat mengenai analisis sentimen dapat diambil kesimpulan bahwa analisis sentimen adalah sebuah proses untuk menentukan sentimen atau opini dari seseorang yang biasanya di wujudkan dalam bentuk teks dan bisa dikategorikan sebagai sentimen positif atau negatif (Muchammad Shiddieqy Hadna et al., 2016).

Analisis sentimen sendiri dapat dibagi kedalam 2 bagian, yaitu *opinion mining* yang berkaitan dengan ekspresi dan pendapat; dan *emotional mining* yang berkaitan dengan emosi seseorang dalam pengucapan atau artikulasi (Yadollahi et al., 2017).



Gambar 2. 1 Taksonomi tugas analisis sentimen

Opinion Mining lebih mengarah kepada konsep opini yang diungkapkan dalam teks yang dapat dikategorikan menjadi ekspresi positif, negatif atau netral, sementara *emotion mining* lebih mengarah kepada emosi seseorang (senang, sedih, marah) yang dituangkan kedalam sebuah teks (Yadollahi et al., 2017).

2.2 *Naïve Bayes*

Algoritma *naïve bayes classifier* merupakan algoritma yang digunakan untuk mencari nilai probabilitas tertinggi untuk mengklasifikasikan data uji pada kategori yang paling tepat, algoritma ini merupakan salah satu metode *machine learning* yang menggunakan perhitungan probabilitas (Romadloni et al., 2019). Keuntungan dari penggunaan algoritma *naïve bayes* adalah bahwa algoritma ini hanya membutuhkan sejumlah data kecil pelatihan untuk memperkirakan parameter yang diperlukan untuk klasifikasi (Romadloni et al., 2019). Algoritma ini bekerja dengan cara mengklasifikasikan kelas berdasarkan pada probabilitas sederhana dimana dalam hal ini diasumsikan setiap attribut yang ada sifatnya saling terpisah, adapun persamaan rumus metode tersebut adalah seperti persamaan (1) dan (2) berikut (Somantri & Dairoh, 2019).

$$P = (Y_k | x_1, x_2, \dots x_a)$$

$$P = (Y_k | x_a) = \frac{P(Y_k) + P(X_a | Y_k)}{P(X_a)}$$

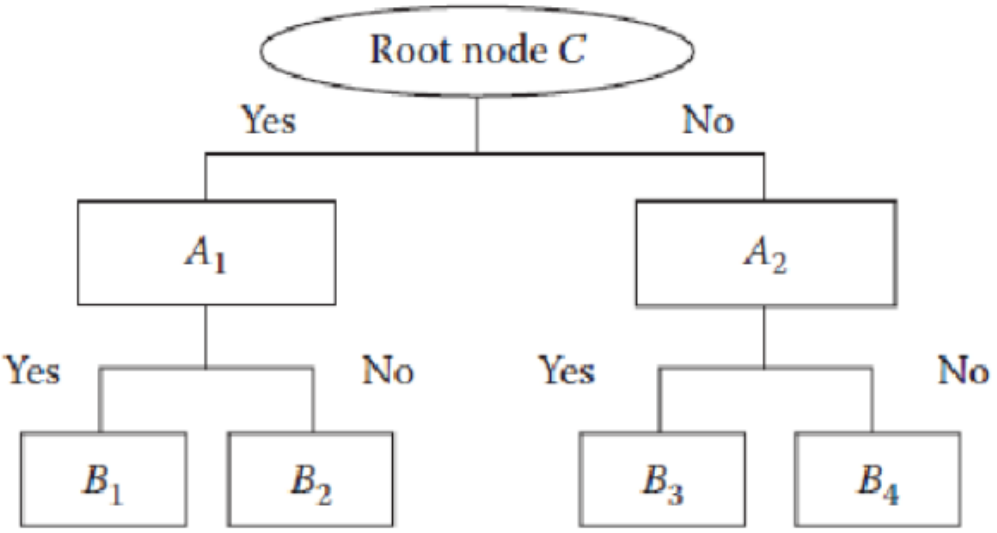
Dimana :

- $Y_k P(X_a|Y_k)$ = kategori kelas
- $P(Y_k)$ = probabilitas kelas
- $P(X_a)$ = probabilitas kemunculan dokumen

Berdasarkan hasil yang didapatkan, kemudian dilakukan proses pemilihan kelas yang optimal sehingga dipilih suatu nilai peluang terbesar dari setiap probabilitas kelas yang ada (Somantri & Dairoh, 2019).

2.3 *Decision Tree*

Metode *decision tree* adalah sebuah metode klasifikasi representasi dari sebuah *tree* atau pohon keputusan, dimana sebuah attribut direpresentasikan sebagai *node*, dan sebagai nilainya adalah cabang pohon tersebut, serta kelas dipresesentasikan sebagai kelas (Somantri & Dairoh, 2019). Dalam hal ini *node* yang paling atas dari sebuah *decision tree* dinamakan sebagai *root*, seperti yang ada pada gambar (Somantri & Dairoh, 2019).



Gambar 2. 2 Gambaran struktur decision tree

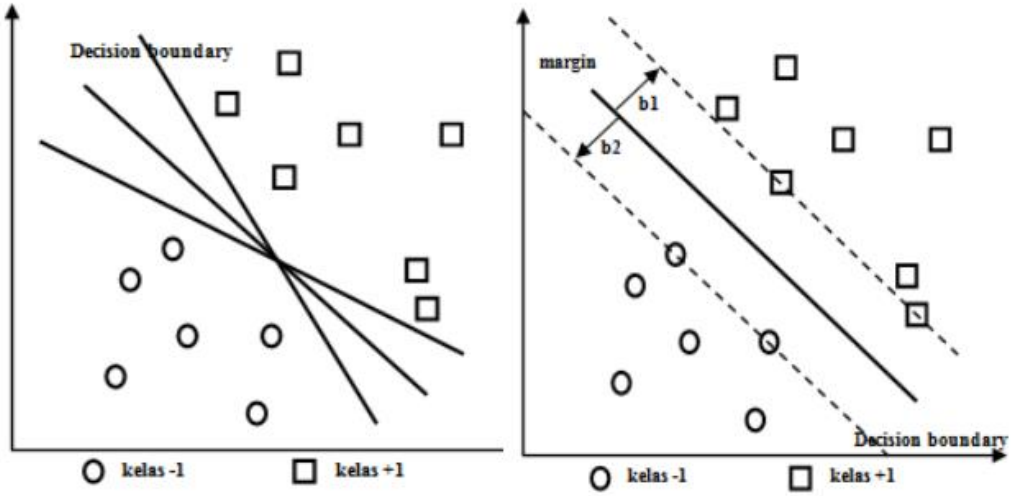
Salah satu algoritma decision tree adalah C4.5, algoritma ini yang digunakan untuk menghasilkan pohon keputusan dan biasa digunakan untuk klasifikasi, oleh karena hal tersebut C4.5 sering disebut sebagai pengklasifikasi statistik (Kusrorong et al., 2019).

2.4 K-*Nearest Neighbors*

Algoritma *K-Nearest Neighbors (k-NN)* merupakan salah satu algoritma paling populer dalam machine learning hal ini karena prosesnya mudah dan sederhana, selain itu *k-NN* juga salah satu dari algoritma *supervised learning* dengan proses belajar berdasarkan nilai dari variabel target yang terasosiasi dengan nilai variabel prediktor (Tempola et al., 2018). Prinsip sederhana yang diadopsi oleh algoritma NN adalah "jika suatu hewan berjalan seperti bebek, maka hewan itu mingkin bebek", semakin dekat lokasi data uji, maka bisa dikatakan bahwa data latih tersebut yang lebih dipandang mirip oleh data uji (Aulianita, 2016). Dalam algoritma *k-NN* semua data yang dimiliki harus memiliki label, sehingga ketika ada data baru yang diberikan kemudian dibandingkan dengan data yang telah ada dan diambil data yang paling mirip dan melihat label dari data tersebut (Tempola et al., 2018). Data pembelajaran diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data, ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi data pembelajaran, nilai k yang terbaik untuk algoritma ini tergantung pada data, secara umumnya, nilai k yang tinggi akan mengurangi efek noise pada klasifikasi, tetapi membuat batasan antarasetiap klasifikasi menjadi lebih kabur (Romadloni et al., 2019).

2.5 *Support Vector Machine*

Support vector machine adalah seperangkat metode pembelajaran terbimbing (*supervised learning*) yang menganalisis data dan mengenali pola, digunakan untuk klasifikasi dan analisis regresi (Muchammad Shiddieqy Hadna et al., 2016). *Support Vector Machine* merupakan salah satu metode terbaik yang bisa dipakai dalam permasalahan klasifikasi, konsep SVM bermula dari masalah klasifikasi dua kelas sehingga membutuhkan *training set* positif dan negatif (Pratama et al., 2018). Konsep klasifikasi ini dilakukan dengan memaksimalkan batas *hyperplane* yang memisahkan suatu set data atau class (Nasution & Hayaty, 2019). Kemampuan *Support Vector Machine* dalam menemukan *hyperplane* terbaik menjadikan algoritma ini memiliki tingkat generalitas yang tinggi serta menjadikannya algoritma dengan tingkat akurasi yang terbaik dibandingkan dengan algoritma lainnya (Nasution & Hayaty, 2019).



Gambar 2. 3 Ilustrasi Support Vector Machine

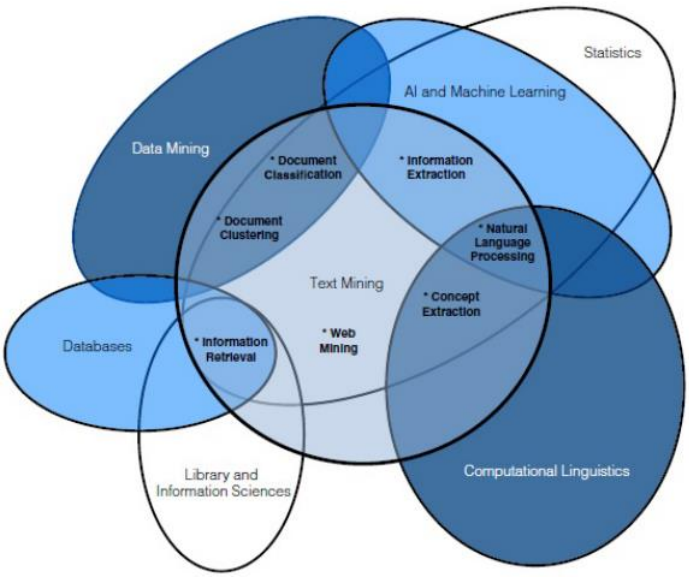
Gambar diatas menjelaskan konsep klasifikasi SVM, pada gambar (a) terdapat sejumlah data dengan lingkaran sebagai kelas -1 dan kotak sebagai kelas +1, pada gambar tersebut juga terdapat *hyperplane* yang mungkin untuk set data (Neneng et al., 2016). Gambar (b) adalah *hyperplane* yang paling maksimal, perhitungan *hyperplane* dilakukan dengan cara menghitung jarak margin dengan data terdekat dari masing-masing kelas, data terdekat ini disebut *Support Vector Machine* (Neneng et al., 2016). Beberapa kelebihan dari metode SVM adalah metode ini bekerja baik untuk sekumpulan data yang tidak dapat dipisahkan secara linear. Pada penggunaan kernel Gaussian, pemilihan parameter C dan r yang tepat dapat membuat SVM bekerja dengan baik walaupun data yang dilatih memiliki nilai bias (Widiastuti et al., 2017).

2.6 R *Programming*

R merupakan bahasa pemrograman yang bersifat *open source* dan merupakan sebuah perangkat lunak yang pada dasarnya digunakan oleh ahli statistik dan penambang data untuk mempelajari berbagai macam data statistik (Singh & Choudhary, 2017). R sendiri secara khusus untuk analisis statistik yang membuatnya sangat cocok digunakan untuk aplikasi ilmu *data science* (Van Atteveldt & Benoit, 2017). Meskipun pemahaman untuk orang yang ingin mempelajari R *Programming* bisa menjadi sangat sulit, terutama untuk orang tanpa pengalaman pemrograman sebelumnya, R memberikan beberapa *tools* yang memudahkan penggunanya dalam melakukan *text analysis* hanya dengan menggunakan beberapa perintah sederhana (Van Atteveldt & Benoit, 2017).

2.7 *Text Mining*

Sebuah studi menyatakan bahwa *text mining* telah menjadi salah satu bidang yang terkenal dan tergabung dalam beberapa bidang penelitian seperti *computational linguistics, Information Retrieval (IR)* dan *Data Mining* (Eroglu et al., 2018). *Text Mining* berbeda dari *Data Mining* yang lebih fokus kepada penemuan pola yang menarik dari basis data yang besar dibandingkan informasi tekstual (Eroglu et al., 2018). *Text Mining* adalah adalah sebuah proses mengekstrasi pola yang menarik dan signifikan untuk mengeksplorasi pengetahuan dari suatu sumber data tekstual (Talib et al., 2016). *Text Mining* menjadi area baru yang mencari dan meng-ekstrak data tekstual yang alami, hal ini dilakukan untuk dapat menganalisa teks dan memisahkan informasi yang dibutuhkan untuk tujuan yang spesifik (Preethi & Radha, 2016). Gambar 2.1 menunjukkan diagram venn untuk hubungan antar teknik *text mining* dan fungsionalitas mereka (Talib et al., 2016)



Gambar 2. 4 Text Mining interaction with other field

2.8 *Twitter API*

Twitter API, merupakan API (*Application Programming Interface)* yang disediakan oleh twitter untuk memfasilitasi pengguna agar dapat berinteraksi dengan data-data yang ada pada aplikasi twitter, data yang dimaksud seperti *tweet*, id pengguna, lokasi, waktu pembuatan tweet dan lain lain (Saputra, 2017). Twitter API menggunakan arsitektur REST (Representational State Transfer) sehingga twitter API dapat digunakan pada format data yang beragam seperti xml ataupun JSON (Rustiana & Rahayu, 2017). Untuk memanfaatkan Twitter API, pengguna harus menggunakan bahasa server side scripting seperti php, python, R dan lain-lain, dengan menggunakan bahasa-bahasa tersebut pengguna dapat melakukan request kepada Twitter API, dan respon hasilnya dirupakan dalam format JSON (Saputra, 2017). Twitter API dapat digunakan oleh pengembang aplikasi dengan cara mendaftarkan aplikasi yang ingin dibuat beserta alasan mengapa membutuhkan Twitter API untuk mendapatkan Consumer Key, Consumer Secret, Access Token, dan Access Token Secret yang nantinya akan dimasukkan kedalam script untuk proses crawling data dari twitter (Putri & Hendrowati, 2018). Agar komunikasi pengguna dengan Twitter API aman, maka Twitter menerapkan Oauth atau OpenAuthorization. OAuth merupakan protokol terbuka yang memungkinkan pengguna untuk berbagi resource pribadi seperti foto, video, data pengguna dan lain-lain yang tersimpan di suatu situs web, dengan situs lain tanpa memberikan nama pengguna dan kata sandi pengguna tersebut.

OAuth mengizinkan pengguna untuk memberikan akses kepada situs pihak ketiga untuk mengakses informasi mereka yang disimpan di penyedia layanan lain tanpa harus membagi izin akses atau keseluruhan data mereka (Saputra, 2017).

2.9 *Data Pre-Processing*

Data *Pre-Processing* adalah tugas penting bagi banyak peneliti, administrator, organisasi dan perusahaan untuk mengumpulkan data dan menganalisis sejumlah data atau informasi yang spesifik (Amalanathan, 2019). Data pre-processing merupakan tahapan yang sangat penting dalam melakukan analisis sentimen, karena tahapan ini dapat menentukan data dapat diklasifikasikan dengan benar (Krouska et al., 2016). Tahapan pre-processing atau praproses data merupakan proses untuk mempersiapkan data mentah sebelum dilakukan proses lain, pada umumnya tahapan ini dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah bentuk data menjadi bentuk yang lebih mudah diproses oleh sistem (Mujilahwati, 2016). Persiapan data sangat direkomendasikan karena berbagai alasan seperti kualitas database dan persiapan analisis data (Amalanathan, 2019). Dalam melakukan praproses data yang bersumber dari twitter dapat dilakukan dengan beberapa proses ekstrasi data antara lain *case folding, remove punctuation, remove username, remove hashtag, clean number, clean one char, remove URL dan remove RT* (Mujilahwati, 2016).

2.10 *Tf-Idf*

Term frecuency-Inverse document frecuency atau lebih dikenal dengan *Tf-Idf* digunakan untuk mengekstrak kalimat dengan cara memberikan nilai atau bobot pada kalimat (Widiastuti et al., 2017). Metode ini menggabungkan dua konsep untuk perhitungan bobot, yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen tertentu dan inverse frekuensi dokumen yang mengandung kata tersebut (Nurjannah & Fitri Astuti, 2013). Nilai *Tf-Idf* menentukan besar bobot kalimat, penentuan nilai bobot dilakukan dengan cara menghitung frekuensi kemunculan kata dalam dokumen (Widiastuti et al., 2017). Rumus untuk *Tf-Idf* adalah sebagai berikut (Nurjannah & Fitri Astuti, 2013).

$$tf = 0{,}5 + 0{,}5 \times \frac{tf}{\max(tf)}$$

$$idft = \log\left(\frac{D}{dft}\right)$$

$$Wd.t = tfd.t \times IDFd.t$$

Keterangan:

tf : banyaknya kata yang dicari pada sebuah dokumen

max tf : jumlah kemunculan terbanyak term pada dokumen yang sama

Nilai D = total dokumen

dft = jumlah dokumen yang mengandung term t.

IDF = Inversed Document Frequency (log2 (D/df))

d = dokumen ke-d

t = kata ke-t dari kata kunci

W = bobot dokumen ke-d terhadap kata ke-t

2.11 *Rapidminer*

RapidMiner merupakan platform perangkat lunak yang menyediakan *integrated environment* untuk *machine learning, text mining, predictive analytics,* dan *bussiness analytics*(Dr.J.Arunadevi et al., 2018). RapidMiner biasa digunakan untuk bisnis, aplikasi komersial, serta untuk kebutuhan penelitian, pendidikan, dan pelatihan karena kecepatannya dalam membuat *prototype* dan mendukung semua proses *machine learning* termasuk persiapan data, visualisasi data, validasi dan optimasi (Kori, 2017). Selain menyediakan *tools* untuk pemrosesan data (ETL), pemodelan data, dan visualisasi data, RapidMiner juga memungkinkan koneksi ke beberapa sumber data seperti Oracle, Microsoft SQL Server, MySQL, Excel, dan berbagai format data lainnya (Krstevski et al., 2011). RapidMiner dikembangkan pada *open core model*, dengan menggunakan RapidMiner *Basic Edition* pengguna dapat mengunduhnya dibawah lisensi AGPL (Dr.J.Arunadevi et al., 2018).

2.12 K-fold Cross Validation

Cross Validation atau dapat disebut estimasi rotasi adalah sebuah teknik validasi model untuk menilai bagaimana hasil statistik analisis akan menggeneralisasi kumpulan data independen, teknik ini utamanya digunakna untuk melakukan prediksi model dan memperkirakan seberapa akurat sebuah model prediktif ketika dijalankan dalam prakteknya (Tempola et al., 2018). Dalam sebuah masalah prediksi, sebuah model biasanya diberikan kumpulan data (dataset) yang diketahui untuk digunakan dalam menjalankan pelatihan (*data training*), serta kumpulan data yang tidak diketahui (*data testing*) terhadap model yang diuji (Kusrorong et al., 2019). Tujuan dari Cross Validation adalah untuk mendefinisikan dataset untuk "menguji" model dalam tahap pelatihan (validasi data), dalam rangka untuk membatasi masalah seperti terjadinya overfitting (Kusrorong et al., 2019). Penggunaan k-fold cross validation untuk menghilangkan bias pada data, pelatihan dan pengujian dilakukan sebanyak k kali, pada percobaan pertama, subset S1 diperlakukan sebagai data pengujian dan subset lainnya diperlakukan sebagai data pelatihan, pada percobaan kedua subset S1, S3,…Sk menjadi data pelatihan dan S2 menjadi data pengujian, dan setererusnya (Tempola et al., 2018). Tingkat kesalahan pada iterasi yang berbeda akan dihitung rata-ratanya untuk menghasilkan error rate secara keseluruhan, model yang memberikan rata-rata kesalahan terkecil dapat dikatakan sebagai model yang terbaik (Menarianti, 2015).

2.13 *ROC Curve*

ROC curve banyak digunakan dalam peneitian data mining dalam menilai hasil prediksi, secara teknis *ROC Curve* dibagi menjadi dua dimensi, dimana tingkat *True Positif* diletakkan pada sumbu Y dan tingkat *False Positif* diletakkan pada sumbu X (Menarianti, 2015). Grafik *Receiver Operating characteristic (ROC)* adalah teknik untuk menggambarkan, mengorganisasi dan memilih pengklasifikasi berdasarkan kinerja mereka, kurva ini digunakan untuk mengukur nilai *Area Under Curve (AUC)* (Indrayuni, 2016). *AUC* digunakan untuk mempresentasikan grafis yang menentukan klasifikasi mana yang lebih baik dengan menghitung luas daerah dibawah *ROC* (Menarianti, 2015). *AUC* mengukur kinerja diskriminatif dengan memperkirakan probabilitas *output* dari sampel yang dipilih secara acak dari populasi positif dan negatif, semakin besar AUC, semakin kuat klasifikasi yang digunakan (Menarianti, 2015). Kurva *ROC* menunjukkan akurasi dan membandingkan klasifikasi secara visual dengan mempresentasikan *confussion matrix*, sedangkan AUC dihitung untuk mengukur perbedaan performansi metode yang digunakan (Rosandy, 2016). Pedoman untuk mengklasifikasikan keakuratan pengujian menggunakan nilai AUC adalah sebagai berikut (Indrayuni, 2016).

a) 0.90 – 1.00 = *Excellent Classification*
b) 0.80 – 0.90 = *Good Classification*
c) 0.70 – 0.80 = *Fair Classification*
d) 0.60 – 0.70 = *Poor Classification*
e) 0.50 – 0.60 = *Failure*

2.14 *Confussion Matrix*

Confussion matrix melakukan pengujian untuk memperkirakan obyek yang benar dan salah, urutan pengujian ditabulasikan dalam *confussion matrix* dimana kelas yang diprediksi ditampilkan dibagian atas dan kelas yang diamati di sebelah kiri (Menarianti, 2015). *Confussion Matrix* digambarkan dengan tabel yang menyatakan jumlah data uji benar diklasifikasikan dan jumlah data uji yang salah diklasifikasikan (Rahman et al., 2017).

Correct Classification	*Classified as*	
	Predicted "+"	Predicted "-"
Actuall "+"	True Positives	False Negative
Actual "-"	False Positives	True Negatives

Tabel 2. 1 Tabel Confussion Matrix

Berdasarkan tabel diatas:

a. True Positives adalah jumlah record data positif yang diklasifikasikan sebagai nilai positif
b. False Positives adalah jumlah record data negatif yang diklasifikasikan sebagai nilai positif
c. False Negatives adalah record data positif yang diklasifikasikan sebagai nilai positif
d. True Negatives adalah record data negatif yang diklasifikasikan sebagai nilai negatif

Nilai dari *True Positive* dan *True-Negative* memberikan informasi ketika classifier dalam melakukan klasifikasi data bernilai benar, sedangkan *False Positive* dan *False-Negative* memberikan informasi ketika classifier salah dalam melakukan klasifikasi data (Fibrianda & Bhawiyuga, 2018). Perhitungan akurasi, presisi dan *recall* dengan tabel *confussion matrix* adalah sebagai berikut (Rosandy, 2016) :

$$\text{Akurasi} = \frac{(A+D)}{(A+B+C+D)}$$

$$\text{Presisi} = \frac{(A)}{(C+A)}$$

$$\text{Recall} = \frac{A}{(A+B)}$$

Keterangan :

A = *True Positives*

B = *False Negatives*

C = *False Positives*

D = *True Negatives*

Akurasi merupakan presentase jumlah record data yang diklasifikasikan (prediksi) secara benar oleh suatu algoritma (Rahman et al., 2017). Sedangkan presisi merupakan proporsi jumlah dokumen teks yang relevan terkendali diantara semua dokumen teks yang terpilih oleh sistem dan *recall* merupakan proporsi jumlah dokumen teks yang relevan terkendali diantara semua dokumen teks relevan yang ada pada koleksi (Andika et al., 2019). Akurasi, Presisi dan Recall dapat diberi nilai dalam bentuk angka dengan menggunakan perhitungan persentase (1-100%) atau dengan menggunakan bilangan antara 0-1, sebuah sistem akan dianggap baik jika nilai akurasi, presisi dan recallnya tinggi (Rosandy, 2016).

2.15 Penelitian Terdahulu

Nama Peneliti	Nama Jurnal	Judul Penelitian	Hasil Penitian
(Hermanto et al., 2018)	Journal of Physics: Conference Series 1140 (2018)	Twitter Social Media Sentiment Analysis in Tourist Destinations Using Algorithms Naive Bayes Classifier (Hermanto et al., 2018)	Pada penelitian ini penulis menggunakan sosial media twitter untuk menganalisa destinasi wisata yang populer di Indonesia, hal ini dilakukan karena Indonesia berada dalam ranking 47 dunia sebagai destinasi wisata yang paling diminati di dunia. Metode yang digunakan untuk mengklasifikasikan sentimen adalah dengan menggunakan algoritma naïve bayes, namun diakhir karya ilmiah penulis menyarankan untuk menggunakan algoritma lain untuk mendapatkan hasil yang lebih akurat.
(Mardiana et al., 2019)	Jurnal PILAR Nusa Mandiri Vol.15, No.2 September 2019	Komparasi Metode Klasifikasi Pada Analisis Sentimen Usaha Waralaba Berdasarkan Data Twitter	penelitian ini, telah membandingkan lima metode klasifikasi dalam menentukan sentimen usaha waralaba berdasarkan opini dari Twitter. Pada

			pembahasan menunjukkan bahwa Naive Bayes, Neural Network, K-Nearest Neighbor, Support Vector Machines, dan Decision Tree memperoleh hasil yang berbeda dalam akurasi, *precision* dan juga *recall.* Namun, di antara kelima metode klasifikasi, dapat dilihat bahwa Support Vector Machine menghasilkan akurasi tertinggi sebesar 83%. Sedangkan *Decision Tree* memperoleh 81%, *Naive Bayes* sebesar 80%, dan *K-Nearest Neighbor* sebesar 52%
(Aulianita, 2016)	Journal Speed – Sentra Penelitian Engineering dan Edukasi – Volume 8 No 3 - 2016	Komparasi Metode K-Nearest Neighbors dan Support Vector Machine Pada Sentiment Analysis Review Kamera	Penelitian ini membandingkan dua metode, yaitu k-NN dan SVM yang diimplementasikan pada *sentiment analysis review* kamera untuk mendapatkan hasil klasifikasi teks terbaik. Kedua metode dipilih berdasarkan tinjauan terdahulu dan ingin

			membuktikan bahwa metode k-NN merupakan metode dengan akurasi terbaik dibandingkan dengan algoritma SVM. Hasil penelitian menunjukkan akurasi dan nilai AUC lebih tinggi didapatkan dengan menggunakan algoritma k-NN yang mendapatkan nilai akurasi = 79% dan nilai AUC = 0.929 dibandingkan menggunakan algoritma SVM yang mendapatkan nilai akurasi = 72% dan nilai AUC = 0.845 pada review kamera.

Mengacu pada penelitian terdahulu yang ada pada tabel 2.1, dapat ditemukan variabel-variabel yang sesuai digunakan pada penelitian ini. Pariwisata di Indonesia selalu mengalami peningkatan dari tahun ke tahun, sektor pariwisata di Indonesia sendiri memiliki berada di ranking 47 di dunia (Hermanto et al., 2018). Pertumbuhan sektor pariwisata di Indonesia didukung dengan adanya teknologi informasi seperti sosial media sebagai media promosi terhadap pariwisata di Indonesia, salah satu media sosial yang digunakan adalah *Twitter* (Hermanto et al., 2018). Beberapa algoritma seperti *Naive Bayes, K-Nearest Neighbor, Support Vector Machines,* dan *Decision Tree* biasa digunakan dalam melakukan sentimen analisis dan hasil akurasi terbaik didapat dengan menggunakan algoritma SVM (Mardiana et al., 2019). Namun dalam penelitian lain yang membandingkan kedua metode, yaitu SVM dan k-NN menyebutkan bahwa akurasi dan nilai *Area Under Curve* lebih tinggi dihasilkan dengan menggunakan metode *K-Nearest Neighbors* (Aulianita, 2016). Kedua penelitian tersebut membuktikan bahwa terdapat banyak faktor yang menentukkan hasil akurasi dan nilai AUC sehingga hasil akurasi dari setiap penelitian berbeda-beda tergantung cara pelatihan dan objek yang di teliti. Oleh karena hal tersbut, pada penelitian ini akan dilakukan perbandingan antara algoritma *Naive Bayes, K-Nearest Neighbor, Support Vector Machines,* dan *Decision Tree* dalam melakukan analisis sentimen masyarakat terhadap pariwisata di Indonesia, khususnya kota Bali.