BAB III
LANDASAN TEORI

3.1 *Twitter*

Twitter adalah sebuah layanan jejaring sosial (media sosial) dan juga mikroblog yang memungkinkan penggunanya berkirim dan membaca pesan yang tidak lebih dari 280 karakter yang disebut sebagai tweet. Sebelumnya, pesan di *Twitter* hanya sampai 140 karakter tetapi pada tanggal 7 November 2017 ditambah menjadi 280 karakter. *Twitter* didirikan pada 21 Maret 2006 oleh Jack Dorsey, Noah Glass, Biz Stone, dan Evan Williams. Sosial Media *Twitter* sendiri dirilis ke publik pada 15 Juli 2006. Pusatnya berada di San Francisco, California, Amerika Serikat (Leslie, 2009).

Twitter API (*Application Programming Interface*) merupakan sejumlah fungsi yang dapat digunakan pengembang perangkat lunak untuk mengolah data saat membangun perangkat lunak. *Twitter* API menyediakan beberapa fungsi untuk melakukan suatu tugas tertentu, sehingga pengembang perangkat lunak hanya memanggil fungsi tersebut di dalam perangkat lunak yang dibangun. *Twitter* API menggunakan arsitektur REST (*Representational State Transfer*) sehingga *Twitter* API dapat digunakan pada format data yang beragam seperti XML maupun JSON. *Twitter* API terdiri atas *Twitter Search* API dan *Twitter Streaming* API. Perbedaan keduanya yaitu, *Twitter Search* API menitikberatkan fungsi pencarian ke masa lampau sedangkan *Twitter Streaming* API menitikberatkan fungsi pencarian ke masa yang akan datang (Rustiana & Nina, 2017).

Berdasarkan Kusuma (2009) terdapat beberapa istilah-istilah yang sering ditemui pada *Twitter*, yaitu:

a. *Timeline* adalah daftar *tweet* terbaru dari pengguna *Twitter* yang diikuti pemilik akun, termasuk *tweet* yang dibuat pemilik akun.
b. *Direct Message* (DM) yaitu fasilitas berkirim pesan antar pengguna secara lebih *private*. DM hanya bisa dilakukan oleh pihak yang diikuti (di-*follow*).

c. *Trending topics* adalah daftar tema yang tengah hangat diperbincangkan di kalangan pengguna *Twitter*.
d. *Tweet* merupakan informasi yang terdiri dari pesan 140 karakter. *Tweet* berisi berita terbaru atau ter*update* yang berkaitan dengan hal - hal yang pemilik akun sukai.
e. *Reply tweet* atau sering disebut *response tweet* (RT) adalah komentar atau balasan atas *tweet*.
f. *Retweet* adalah menyalin seluruh isi *tweet* dari akun lain.
g. *Follow* adalah mengikuti akun dan informasi yang disampaikan oleh seorang pengguna.
h. *Follower* adalah pengikut atau yang mengikuti akun seseorang.
i. *Mention* (@) digunakan untuk menyebut username pihak yang akan diajak berkomunikasi. Penggunaan simbol ini berada di awal sebelum menuliskan username pihak yang dituju.
j. *Hastags* atau tanda pagar atau tagar (#) adalah tanda yang digunakan untuk menandai kata kunci untuk topik diskusi atau informasi yang dibagikan agar mudah dicari.

3.2 Kebakaran Hutan

Kebakaran hutan dibedakan dengan kebakaran lahan. Kebakaran hutan yaitu kebakaran yang terjadi di dalam kawasan hutan, sedangkan kebakaran lahan adalah kebakaran yang terjadi di luar kawasan hutan dan keduanya bisa terjadi baik disengaja maupun tanpa sengaja (Hatta, 2008).

Kebakaran hutan ialah terbakarnya sesuatu yang menimbulkan bahaya atau mendatangkan bencana. Kebakaran dapat terjadi karena pembakaran yang tidak dikendalikan, karena proses spontan alami, atau karena kesengajaan. Proses alami sebagai contohnya kilat yang menyambar pohon atau bangunan, letusan gunung api yang menebarkan bongkahan bara api, dan gesekan antara ranting tumbuhan kering yang mengandung minyak karena goyangan angin yang menimbulkan panas atau percikan api (Notohadinegoro, 2006). Kebakaran yang terjadinya akibat kesengajaan manusia dikarenakan oleh beberapa kegiatan,

seperti kegiatan ladang, perkebunan (PIR), Hutan Tanaman Industri (HTI), penyiapan lahan untuk ternak sapi, dan sebagainya (Hatta, 2008).

Menurut Darwiati dan Tuheteru (2010) di Indonesia, kebakaran hutan dan lahan hampir 99% diakibatkan oleh kegiatan manusia baik disengaja maupun tidak (unsur kelalaian). Diantara angka persentase tersebut, kegiatan konversi lahan menyumbang 34%, peladangan liar 25%, pertanian 17%, kecemburuan sosial 14%, proyek transmigrasi 8%; sedangkan hanya 1% yang disebabkan oleh alam. Faktor lain yang menjadi penyebab semakin hebatnya kebakaran hutan dan lahan sehingga menjadi pemicu kebakaran adalah iklim yang ekstrim, sumber energi berupa kayu, deposit batubara dan gambut.

3.3 Banjir

Banjir adalah debit aliran aair sungai dalam jumlah yang tinggi, atau debit air di sungai secara relative lebih dari kondisi normal akibat hujan yang turun di hulu atau di suatu tempat tertentu terhadi secara terus menerus, sehingga air tersebit tidak dapat ditampung oleh alur sungai yang ada, maka air melimpah keluar dan menggenangi daerah sekitarnya (Febriani, 2014). Banjir terjadi akibat naiknya permukaan air akibat curah hujan yang diatas normal, perbuhan suhu, tanggul yang jebol, pencairan salju yang cepat, terhambatnya aliran air di tempat lain (Ligal, 2008).

Ada dua peristiwa banjir, pertama peristiwa banjir yang terjadi pada daerah yang biasanya tidak terjadi banjir dan kedua peristiwa banjir terjadi karena limpahan air banjir dari sungai karena debit air yang tidak mampu dialirkan oleh sungai atau debit air lebih besar lebih besar dari kapasitas pengaliran sungai yang ada (Kodoatie dan Sugiyanto, 2002)

3.4 Data *Mining*

Data *Mining* adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar. Data *Mining* juga disebut sebagai serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data (Pramudiono, 2007). Data *Mining*, sering juga disebut sebagai

knowledge discovery in database (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar (Santoso, 2007). Secara umum Data *Mining* adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basisdata dengan melakukan penggalian pola-pola dari data dengan tujuan untuk mengubah data menjadi informasi yang lebih berharga yang diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat dalam suatu basis data.

3.5 *Machine Learning*

Machine Learning adalah suatu area dalam *artificial intelligence* atau kecerdasan buatan yang berhubungan dengan pengembangan teknik-teknik yang bisa diprogramkan dan belajar dari data masa lalu. Pengenalan pola, data *mining* dan *machine learning* sering dipakai untuk menyebut sesuatu yang sama. Bidang ini bersinggungan dengan ilmu probabilitas dan statistik kadang juga optimasi. *Machine learning* menjadi alat analisis dalam data *mining* (Santoso, 2007). Dengan kata lain *Machine Learning* merupakan ilmu yang mempelajari bagaimana memberikan kemampuan terhadap komputer untuk melakukan aktifitas belajar guna menyelesaikan masalah secara mandiri.

Pada proses pembelajaran menggunakan *machine learning* terbagi menjadi tiga cara yakni:

- *Model Supervised Learning / Predictive*

Model ini digunakan untuk memprediksi hasil masa depan berdasarkan data historis. Model prediktif biasanya diberi instruktsi yang jelas sejak awal seperti apa yang perlu dipelajari dan bagaimana itu perlu dipelajari. Algoritma pembelajaran ini disebut *Supervised Learning.*

Sebagai contoh, *Supervised Learning* digunakan saat perusahaan pemasaran mencoba untuk mengetahui pelanggan mana yang cenderung berpindah atau mencari *supplier* lain. Algoritma ini juga bisa digunakan untuk memprediksi kemungkinan terjadinya bahaya seperti gempa bumi, tornado dan lain-lain, dengan tujuan untuk mengetahui Total Nilai Asuransi. Beberapa conntoh

algoritma yang digunakan adalah *Nearest Neighbour, Naïve Bayes, Decision Tree, Regression*, dan lain-lain.

- *Model UnSupervised Learning/Descriptive*

Model ini digunakan untuk melatih dimana tidak ada target yang ditetapkan dan tidak ada faktor yang penting dari yang lainnya. Sebagai contoh penggunaan *model unspervised learning* ini, bila seorang penjual pengecer ingin mengetahui kombinasi produk apa yang cenderung lebih sering dibeli konsumen. Di industri farmasi, digunakan untuk memprediksi penyakit mana yang mungkin terjadi bersamaan dengan diabetes. Contoh algoritma yang digunakan di model ini adalah *K-Means Clustering Algorithm.*

- *Reinforcement Learning (RL)*

Model ini adalah contoh pembelajaran mesin dimana mesin dilatih untuk mengambil keputusan spesifik berdasarkan kebutuhan bisnis dengan tujuan utama untuk memaksimalkan efisiensi (kinerja). Ide dari *Reinforcement learning* ini adalah mesin/perangkat lunak melatih dirinya secara terus menerus berdasarkan lingkungan yang dipengaruhinya, dan menerapkan pengetahuan yang diperkaya unntuk memecahkan masalah bisnis. Proses belajar yang terus-menerus ini memastikan lebih sedikit keterlibatan manusia sehingga akan banyak menghemat waktu. Contoh algoritma yang digunakan dalam RL adalah *Markov Decision Process.*

3.6 *Text Mining*

Text mining merupakan proses dalam mencari data guna untuk menemukan beberapa informasi atau menemukan intisari dari informasi tersebut yang selanjutnya data yang telah didapat tersebut dapat diolah. *Text mining* juga dikenal sebagai data *mining* teks atau penemuan pengetahuan dari *database* tekstual. Sesuai dengan buku *The Text mining Handbook, text mining* dapat didefinisikan sebagai suatu proses menggali informasi dimana seorang *user* berinteraksi dengan sekumpulan dokumen menggunakan *tools* analisis yang merupakan komponen-komponen dalam data mining. Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen.

Jadi, sumber data yang digunakan dalam *text mining* adalah sekumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Adapun tugas khusus dari *text mining* antara lain yaitu pengkategorisasian teks dan pengelompokkan teks (Triawati, 2009).

Langkah-langkah yang dapat dilakukan dalam melakukan *text mining* adalah sebagai berikut:

3.6.1 *Text Preprocessing*

Text Preprocessing merupakan tahapan awal yang dilakukan untuk mempersiapkan teks agar dapat diolah lebih lanjut. *Preprocessing* secara umum bertujuan untuk mengubah informasi dari tiap-tiap sumber data ke dalam bentuk atau format yang baku sebelum menerapkan berbagai metode-metode pengambilan data terhadap dokumen yang akan diproses (Feldman & Sanger, 2006). Ada beberapa tahapan *pre-processing*, antara lain:

a. *Cleaning Data*

Merupakan proses pembersihan kata dengan menghilangkan tanda baca yang bertujuan untuk mengurangi *noise.*

b. Case Folding

Tahap ini merupakan pengubahan semua karakter huruf diubah menjadi dokumen yang berisi kalimat menjadi huruf kecil atau sebaliknya. Pada tahap *case folding* juga menghilangkan yang dianggap tidak valid atau menghilangkan karakter selain bentuk huruf seperti angka dan lain-lain.

c. Tokenizing

Tokenizing adalah proses memotong suatu kalimat menjadi beberapa bagian berdasarkan kata perkata. Potongan kata perkata tersebut disebut dengan token.

d. Filtering

Filtering merupakan tahap pembersihan kalimat dengan membuang kata-kata yang tidak signifikan, seperti kata ganti, kata hubung, kata keterangan, dan lain-lain dengan menggunakan *stopword* yakni daftar kata

yang akan dihapus pada dokumen, pada *filtering* juga dilakukan penghapusan terhadap spasi yang berlebih akibat dari penghapusan beberapa kata.

3.6.2 *Features Selection*

Pada tahap *feature selection* merupakan tahap pada proses *text mining* dengan menggunakan proses *stopword removal* yaitu berfungsi untuk menghilangkan beberapa kata dari kalimat. Kata yang dihilangkan berupa kata-kata yang tidak memiliki arti seperti kata penghubung atau, yang dan kata-kata lainnya (Nata dan Yudiastra, 2017).

3.6.3 *Text Representation*

Text representation merupakan tahapan merubah data tekstual menjadi sebuah data dari jumlah kalimat yang ada kemudian dilihat dari jumlah kata-kata yang berbeda pada kalimat.

3.6.4 *Application of Text mining Techniques*

Pada tahap ini merupakan tahap penentuan dari proses *text mining* yang telah dilakukan. Pada tahapan ini digunakan sebagai pengambilan informasi yang ingin diketahui dari data yang ada dengan berbagai teknik yang dapat digunakan seperti *classification, Clustering, information extraction, trend analysis, distribution analysis* dan *association rules.*

3.7 *Wordcloud*

Wordcloud merupakan sebuah sistem yang memunculkan visualisasi kata-kata dengan memberikan penekanan pada frekuensi kemunculan kata terkait dalam wacana tertulis (Qeis, 2017). Secara umum *Wordcloud* adalah representasi visual dari data teks, biasa digunakan untuk menggambarkan data pada sebuah situs.

Gambar 3.1 Tampilan *Wordcloud*

3.8 Asosiasi Kata

Asosiasi kata biasanya digunakan untuk mengetahui kata yang sering mucul pada sebuah data teks. Selain itu juga dapat digunakan untuk melihat kata yang memiliki hubungan atau keterkaitan dalam sekumpulan data teks.

Asosiasi kata atau hubungan kata juga dapat dilihat dari nilai korelasi kata, nilai korelasi bervariasi bekisar antara -1 sampai 1. Dapat diketahui apabila nilai mendekati 1 atau -1 maka hubungan antar kata tersebut makin erat, sedangkan jika nilai mendekati 0 maka hubungan semakin lemah.

3.9 *Sentiment Analysis*

Sentiment Analysis dilakukan untuk melihat pendapat atau kecenderungan pendapat terhadap suatu masalah atau objek oleh seseorang, apakah cenderung berpendapat negatif atau positif. *Sentiment Analysis* biasa dilakukan untuk memantau perkembangan pasar atau menanggapi suatu permasalahan, salah satu contoh penggunaannya di dunia nyata adalah indentifikasi kecenderungan pasar atau pendapat terhadap suatu objek (Fahrur Rozi, 2012). Analisis sentimen juga menganalisis sebagian data untuk mengetahui emosi manusia. Analisis sentimen dapat dikategorikan kedalam tiga *task*, yaitu *informative text detection*, *information extraction* dan *sentiment interestingness classification* (*emotional*, *polarity indentification*). *Sentiment classification* (negatif atau positif) digunakan untuk memprediksi *sentiment polarity* berdasarkan data sentimen dari pengguna. (Dang, Zhang, & Chen, 2010).

Analisis Sentimen terbagi menjadi 3 kategori yakni sentimen negatif, sentimen positif, atau sentimen netral. Pada umumnya berupa pendapat atau tanggapan terhadap subjek tertentu, baik menggunakan media sosial maupun media lainnya. Dalam analisis sentimen terdapat tiga klasifikasi utama, yaitu : (Medhat, Hassan, & Korashy, 2014).

1. Level dokumen

   Tujuan dari analisis sentimen level dokumen adalah untuk mengklasifikasikan sebuah dokumen opini yang menunjukkan bahwa dokumen itu berisi opini positif, netral, atau negatif dan menganggap bahwa seluruh isi dokumen tersebut membahas satu topik saja.

2. Level kalimat

   Pada level kalimat, analisis sentimen bertujuan untuk mengklasifikasikan sentimen yang ada pada tiap-tiap kalimat. Analisis sentimen pada level kalimat akan menentukan apakah sebuah kalimat tersebut menyatakan ekspresi opini negatif, netral, atau positif.

3. Level aspek

   Pada level ini, analisis sentimen bertujuan untuk mengklasifikasikan sentimen berdasarkan aspek khusus dari sebuah entitas. Pemberi opini dapat memberikan opini yang berbeda untuk aspek-aspek yang berbeda dari entitas yang sama seperti ”Kualitas suara dari telepon ini jelek, tapi baterainya tahan lama”.

3.10 *Confusion Matrix*

Confusion matrix adalah suatu metode yang biasanya digunakan untuk melakukan perhitungan akurasi pada konsep *data mining*. *Confusion matrix* digambarkan dengan tabel yang menyatakan jumlah data uji yang benar diklasifikasikan dan jumlah data uji yang salah diklasifikasikan (Rahman, Darmawidjadja, dan Alams, 2017)**.**

Pengukuran efektif dapat dilakukan dengan perhitungan perolehan atau *recall*, nilai ketepatan atau presisi, nilai akurasi, dan nilai *spesificity*. *Recall* merupakan proporsi jumlah yang dapat ditemukan kembali dalam proses pencarian. Presisi merupakan proporsi jumlah dokumen yang ditemukan dan

dianggap relevan untuk kebutuhan suatu informasi. Akurasi adalah nilai ketepatan suatu klasifikasi dalam bentu persen dan *spesificity* digunakan untuk mengukur proporsi negatif yang benar diidentifikasi (Sasongko, 2016).

Tabel 3.1 *Confusion Matrix*

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

1. *True Positive* (TP) merupakan kelas yang dihasilkan dari prediksi pada klasifikasi positif dan kelas sebenarnya positif.
2. *True Negative* (TN) merupakan kelas yang dihasilkan dari prediksi pada klasifikasi negatif padahal kelas sebenarnya positif.
3. *False Positive* (FP) merupakan kelas yang dihasilkan dari prediksi pada klasifikasi negatif padahal kelas sebenarnya positif.
4. *False Negative* (FN) meripakan kelas yang dihasilakan dari prediksi pada klasifikasi negatif dan kelas sebenarnya negatif.

Dari tabel diatas, didapatkan perhitungan *recall*, presisi, akurasi, dan perhitungan lainnya dalam rumus sebagai berikut:

$$Recall = \frac{TP}{FP+FN} x100 \quad (3.1)$$

$$\text{Presisi} = \frac{TP}{FP+TP} x100 \quad (3.2)$$

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} x100 \quad (3.3)$$

$$Spesificity = \frac{TN}{TN+FP} x100 \quad (3.4)$$

False Positive Rate (FPR) = 1-spesificity (3.5)

$$Area\ Under\ Curve = \frac{1+Recall-FPR}{2} \quad (3.6)$$

Nilai *Area Under Curve* (AUC) digunakan untuk mengukur kinerja deskriminatif menggunakan perkiraan probabilitas hasil dari sampel yang telah dipilih secara acak dari suatu populasi negatif dan positif. Nilai AUC berkisar antara 0 sampai 1, klasifikasi dikatakan baik jika nilai AUC semakin tinggi.

Tabel 3.2 Nilai *Area Under Curve* (AUC)

Nilai AUC	Keterangan
0.91 - 1.00	Klasifikasi sangat baik
0.81 – 0.90	Klasifikasi baik
0.71 – 0.80	Klasifikasi cukup
0.61 – 0.70	Klasifikasi buruk
≤ 0.60	Klasifikasi salah

(**Sumber :** Gorunescu, 2011)

3.11 *TF-IDF*

Menurut fitri dalam (Herwijayanti, Ratnawati & Muflikhah, 2018) setelah tahap *preprocessing*, selanjutnya data yang diperoleh dilakukan proses pembobotan data berupa kata menjadi numerik dengan menggunakan metode *Term Frequency Inverse Document Frequency* (TF-IDF). Metode TF-IDF digunakan untuk menentukan seberapa jauh keterhubungan kata (*term*) terhadap dokumen dengan memberikan bobot setiap kata.

Dalam perhitungan bobot menggunakan TF-IDF terdapat beberapa tahapan sebagai berikut (Luqyana, Cholissodin & Perdana, 2018):

1. *Term Frequency* (TF) merupakan perhitungan awal dengan menghitung pembobotan setiap kata dengan melihat jumlah kemunculan *term* atau kata pada sebuah dokumen dengan masing-masing bobot setiap kata berbeda bernilai 1. Dengan rumus sebagai berikut :

$$tf_{i,j} = \frac{n_{i,j}}{\sum n_{i,j}} \qquad (3.7)$$

   Ket :

   $tf_{i,j}$= *Term frequency* (kata ke-i dalam dokumen ke-j)

   $n_{i,j}$ = Banyaknya kata ke-i dalam dokumen ke-j

2. Selanjutnya adalah perhitungan *inverse document frequency* (IDF), yaitu menghitung term pada setiap dokumen. Sebelum mencari nilai IDF maka

terlebih dahulu mencari nilai DF yaitu merupakan perhitungan jumlah frekuensi dokumen yang mengandung term yang sama. IDF digunakan untuk mengurangi bobot suatu *term* jika banyak tersebar atau lebih banyak muncul di seluruh koleksi dokumen sehingga dianggap sebagai term umum yang dinilai tidak penting.

$$IDF = \log \frac{N}{df_i} \tag{3.8}$$

Ket :

IDF_i = *Inverse Document frequency* (kata -i)

N = Jumlah dokumen

df_i = Jumlah frekuensi dokumen yang mengandung term (kata ke-i)

3. Perhitungan bobot TF-IDF

$$W_{i,j} = tf_{i,j} \times idf_i \tag{3.9}$$

Ket :

IDF_i = *Inverse Document frequency* (kata ke-i)

$tf_{i,j}$ = *Term frequency* (kata ke-i dalam dokumen ke-i)

$W_{i,j}$ = bobot TF-IDF (kata ke-i dalam dokumen ke-j)

3.12 Teorema Bayes

Ide dasar dari Teorema Bayes adalah menangani masalah yang bersifat hipotesis yakni mendesain suatu klasifikasi untuk memisahkan objek (Santosa, 2007). Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Teorema bayes memiliki bentuk umum sebagai berikut (Sari, 2017) :

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)H} \tag{3.10}$$

Dengan keterangan :

X = Data dengan kelas yang belum diketahui

H = Hipotesis data merupakan suatu kelas spesifik

P(H|X) = Probabilitas hipotesis H berdasarkan X (posterior probabilitas)

P(X) = Probabilitas dari X

P(X|H) = Probabilitas X berdasarkan kondisi hipotesis H

P(H) = Probabilitas dari H (*prior* probabilitas).

3.13 *Naïve Bayes Classifier*

Klasifikasi Naïve Bayes adalah klasifikasi berdasar teorema Bayes dan digunakan untuk menghitung probabilitas tiap klas dengan asumsi bahwa antar satu kelas dengan kelas yang lain tidak saling tergantung (independen) (Ari dan Subanar, 2013). Maka dari asumsi didapatkan persamaan sebagai berikut:

$$P(C|F_1, F_2, ..., F_n) = \frac{P(C)P(F_1, F_2, ..., F_n|C)}{P(F_1, F_2, ..., F_n)} \qquad (3.11)$$

Pada persamaan diatas, variabel C menjelaskan kelas dan variabel F menjelaskan karakteristik petunjuk yang dibutuhkan untuk melakukan proses klasifikasi. Persamaan diatas menjelaskan peluang masuknya sampel karakteristik tertentu dalam kelas C (posterior) merupakan peluang munculnya kelas C (sebelum masuknya sampel, disebut *prior*) yang dikalikan dengan peluang kemunculan karakteristik sampel pada kelas C (*likelihood*), dibagi dengan peluang kemunculan karakteristik-karakteristik sampel secara global (*evidence*). Jadi, rumus pada NBC dapat ditulis secara sederhana sebagai berikut :

$$Posterior = \frac{Prior \times Likelihood}{Evidance} \qquad (3.12)$$

Pada persamaan diatas nilai *evidence* tetap untuk tiap kelas dalam satu sample. Posterior adalah nilai yang akan dibandingkan dengan nilai-nilai posterior kelas lainnya dalam menentukan suatu sampel kedalam kelas yang akan diklasifikasikan. Penjabaran dalam persmaan naive bayes dijelaskan dengan

menjabarkan (C|F,..., Fn) menggunakan aturan perkalian seperti persamaan dibawah ini:

$$P(C|F_1,F_2,...,F_n)=\frac{P(C)P(F_1,F_2,...,F_n|C)}{P(F_1,F_2,...,F_n)} \quad (3.13)$$

$$= \frac{P(C)P(F1|C)P(F2,\ldots,Fn|C,F1)}{P(F1,F2,\ldots,Fn)}$$

$$= \frac{P(C)P(F1|C)P(F2|C,F1)P(F3,\ldots,Fn|CF1,F2,F3)}{P(F1,F2,\ldots,Fn)}$$

$$= \frac{\mathrm{P(C)P(F1|C)P(F2|C,F1)P(F3|C,F1,F2)P(F4,\ldots,Fn|C,F1,F2,F3)}}{P(F1,F2,\ldots,Fn)}$$

Persamaan diatas merupakan penjabaran yang menyebabkan semakin banyak dan semakin kompleksnya faktor yang dapat mempengaruhi nilai probabilitas, dimana hampir mstahil untuk dianalisis satu persatu. Pada metode NBC digunakan asumsi independensi yang sangat tinggi, dimana masing-masing petunjuk untuk $F1, F2, \ldots, Fn$ saling bebas (*independence*) antara satu sama lain. Berdasarkan asumsi tersebut maka berlaku suatu kesamaan berikut :

$$P(F_i|F_j)=\frac{P(F_i\cap F_j)}{P(F_j)}=\frac{P(F_i)P(F_j)}{P(F_j)}=P(F_i) \quad (3.14)$$

dimana i≠j, sehingga:

$$P(F_i|C,F_j) = P(F_i|C) \quad (3.15)$$

Persamaan berikut merupakan persamaan dari model teorema naive bayes yang selanjutnya akan digunakan sebagai proses klasifikasi dalam metode NBC (Saleh, 2015).

Contoh :

Jika ada seseorang (X) yang berumur "≤30" dengan pendapatan "sedang" bersetatus sebagai "siswa" dengan tingkatan kredit "cukup" . Apakah seseorang (X) harus membeli komputer atau tidak? Hitunglah dengan menggunakan peluang dengan metode klasifikasi *naive bayes* berdasarkan **Tabel 3.3.**

Hasil :

Berdasarkan hasil penyelesaian dibawah didapatkan bahwa seseorang tersebut mempunyai peluang membeli komputer sebesar 0.028 dan peluang untuk tidak membeli komputer sebesar 0.007 yang artinya bahwa orang tersebut masuk ke dalam kelas akan membeli komputer.

Tabel 3.3 Tabel Contoh Klasifikasi *Naïve Bayes*

Usia	Pendapatan	Siswa	Aset	Membeli Komputer
≤30	Tinggi	Bukan	Cukup	Tidak
≤30	Tinggi	Bukan	Sangat baik	Tidak
31…40	Tinggi	Bukan	Cukup	Ya
Usia	Pendapatan	Siswa	Aset	Membeli Komputer
>40	Sedang	Bukan	Cukup	Ya
>40	Rendah	Siswa	Cukup	Ya
>40	Rendah	Siswa	Sangat Baik	Tidak
31…40	Rendah	Siswa	Sangat Baik	Ya
≤30	Sedang	Bukan	Cukup	Tidak
≤30	Rendah	Siswa	Cukup	Ya
>40	Sedang	Siswa	Cukup	Ya
≤30	Sedang	Siswa	Sangat Baik	Ya
31…40	Sedang	Bukan	Sangat Baik	Ya
31…40	Tigggi	Siswa	Cukup	Ya
>40	Sedang	Bukan	Sangat Baik	Tidak

Penyelesaian :

Prior Probability (PH)

Membeli Komputer = “Ya” = 9/14 = 0.643

Membeli Komputer = “Tidak” = 5/14 = 0.357

Likelihood (P(X|H)

= P(≤30|Ya) = 2/9 = 0.222

= P(≤30|Tidak) = 3/5 = 0.6

= P(Sedang|Ya) = 4/9 = 0.444

= P(Sedang|Tidak) = 2/5 = 0.4

= P(Siswa|Ya) = 6/9 = 0.667

= P(Siswa|Tidak) = 1/5 = 0.2

= P(Cukup|Ya) = 6/9 = 0.667

= P(Cukup|Tidak) = 2/5 = 0.4

P(X|Ya Membeli Komputer) = 0.222 x 0.444 x 0.667 x 0.667 = 0.044

P(X|Tidak Membeli Komputer) = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

P(X|H)xP(H)

P(X|Ya Membeli Komputer) x P(Ya Membeli Komputer) = 0.044 x 0.643

= 0.28

P(X|Tidak Membeli Komputer) x P(Tidak Membeli Komputer) = 0.019 x

0.357 = 0.007