BAB III
LANDASAN TEORI

3.1 Pembentukan Undang-Undang

Undang-Undang adalah satu jenis peraturan prundang-undangan sebagaimana dimaksud oleh Pasal 7 ayat (1) Undang-Undang Nomor 12 Tahun 2011. Sementara yang dimaksud dengan Peraturan perundang-undangan sebagaimana dimaksud Pasal angka 1 adalah peraturan tertulis yang memuat norma hukum yang mengikat secara umum dan dibentuk atau ditetapkan oleh lembaga Negara atau pejabat yang berwenang melalui prosedur yang ditetapkan dalam peraturan perundang-undangan.

Pembentukan undang-undang meliputi perencanaan, penyusunan, pembahasan dan pengesahan, dan pengundangan. Tahap perencaaan adalah tahap penyusunan dilakukan dalam Program Legislasi Nasional (Prolegnas). Tahap penyusunan adalah pembahasan yang dilakukan oleh Dewan Perwakilan rakyat (DPR), Pemerintah, dan Dewan Perwakilan Daerah (DPD) jika terkait dengan Rancangan Undang-Undang yang menjadi Kewenangan DPD (Pasal 22D UUD 1945). Pada berikutnya adalah tahap pembahasan dan pengesahan. Pada tahap pembahasan dilakuka oleh DPR, Presiden, dan DPD, sedangkan pengesahannya dilakukan oleh DPR (Pasal 66 sampai dengan Pasal 74 UU Nomor 12 Tahun 2011). Sementara tahap akhir pembentukan Undang-Undang adalah tahap pengundangan yaitu penempatannnya dalam Lembaran Negara dan Tambahan Lembaran Negara agar semua orang mengetahuinya. (Pasal 81 dan Pasal 82 UU Nomor 12 Tahun 2011).

3.2 *Twitter*

Twitter adalah situs media sosial yang banyak dipakai dari berbagai kalangan. Pada tanggal 21 Maret 2006 di San Fransisco, California Jack Dorsey mendirikan *twitter*. Pengguna *twitter* dapat berinteraksi dengan pengguna lainnya dimanapun dan kapanpun namun pengguna yang tidak terdaftar hanya bisa membaca *tweet* pengguna lain. Pengguna *twitter* dapat membuat *tweet* dengan kapasitas 140

karakter termasuk spasi dan tanda baca, tetapi *twitter* telah memperbanyak jumlah *tweet* menjadi 280 karakter pada bulan September 2017 (Maulana, 2017).

Twitter menyediakan API, API (*Application Programming Interface*) merupakan suatu program atau aplikasi yang diciptakan oleh suatu perusahaan tertentu dengan tujuan mempermudah pihak aplikasi lain dalam mengakses aplikasi tersebut. *Twitter* API diciptakan untuk mempermudah pihak-pihak lain yang ingin mengambil informasi atau data dari *twitter*. Sebelum menggunakan *twitter* API, pengguna harus memiliki *customer key* dan *customer secret* dengan tujuan agar aplikasi *web* yang dibentuk dapat diketahui oleh pihak *twitter* (Rustiana & N, 2017).

Berikut merupakan beberapa fitur yang terdapat dalam *twitter* (Sugiharto, 2018) :

1. *Following*

   Merupakan akun pengguna *twitter* yang diikuti akun pengguna *twitter* lainnya.
2. *Followers*

   Merupakan akun pengguna *twitter* yang mengikuti akun pengguna *twitter* lainnya.
3. *Tweet*

   Pesan yang terdapat pada *twitter* dengan kapasitas 280 karakter.
4. *Retweet*

   Tweet yang telah dibagikan dan dibagikan ulang oleh pengguna lainnya
5. *Mention*

   Melibatkan beberapa pengguna pada pesan yang dibagikan dengan awalan “@” pada username mereka.
6. *Trending Topics*

   Merupakan sepuluh topik yang sedang ramai dibicarakan di *twitter* pada waktu tertentu.
7. *Twitter* Search

   Fasilitas yang diberikan agar pengguna dapat lebih mudah mencari subjek, nama, tempat atau kata tertentu.

8. *Direct Message*

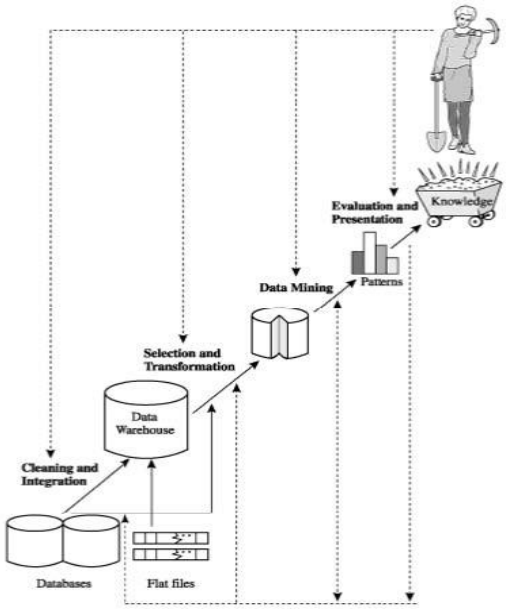
   Digunakan untuk mengirimkan pesan pribadi ke pengguna lainnya yang bersifat privasi.

9. *Hashtag*

   Hashtag atau tanda tagar (#) merupakan tanda yang digunakan untuk menglompokkan konten.

3.3 Data *Mining*

Data *mining* merupakan sebuah proses untuk mendapatkan informasi yang berguna dari kumpulan basis data (Tan, 2006). Data *mining* umumnya digunakan untuk menemukan pengetahuan atau informasi yang tersembunyi dalam basis data. Secara umum, data *mining* merupakan proses yang menggunakan teknik matematik, statistik, dan *machine learning* untuk mengidentifikasi dan pengekstrakan informasi dan berguna yang tersimpan dalam basis data (Aronson, Liang dan McCarthy, 2006). *Knowledge Discovery in Database* (KDD) merupakan sebutan lain dari data mining yang merupakan suatu proses pengambilan informasi yang bermanfaat dan tidak diketahui sebelumnya dari kumpulan sebuah data (Bramer, 2007).



Gambar 3. 2 Tahapan dalam *Knowledge Discovery From Database* (KDD)

(**Sumber :** Han dan Kamber.,2012)

Berdasarkan KDD pada Gambar 3.2 terdapat urutan proses berikut:

a. Pembersihan Data *(Data Cleaning)*

   Pembersihan data digunakan untuk menghilangkan *noise* dan data yang tidak konsisten. Penghapusan data yang tidak memiliki kelengkapan atribut sesuai yang dibutuhkan dilakukan pada tahap ini.

b. Integrasi Data (*Data Integration*)

   Merupakan proses penggabungan data dari beberapa sumber data untuk diolah.

c. Seleksi Data (*Data Selection*)

   Proses pengambilan data yang berkaitan dengan analisis yang akan digunakan dalam proses data *mining.*

d. Transformasi Data (*Data Transformation*)

   Suatu proses transformasi data kedalam bentuk yang diinginkan dalam *mining.*

e. Penambangan Data (*Data Mining*)

   Merupakan proses yang penting dalam KDD yang melibatkan teknik tertentu untuk memperoleh suatu pola dari data yang digunakan.

f. Evaluasi Pola (*Pattern Evaluation*)

   Merupakan proses untuk mengkaji kebenaran dari suatu pola data yang mewakili *knowledge* pada data.

g. Represntasi Pengetahuan (*Knowledge Representation*)

   Suatu proses representasi secara visual kepada pengguna dalam mempermudah pemahaman mengenai hasil dari data *mining*.

3.4 *Machine Learning*

Machine Learning atau disebut juga pembelajaran mesin yang merupakan pendekatan dari kecerdasan buatan (*artificial intelegent*) yang banyak digunakan untuk menirukan dan menggantikan perilaku manusia untuk menyelesaikan masalah (Ahmad, 2017). *Machine learning* adalah suatu proses dalam kecerdasan buatan yang memiliki hubungan terhadap proses pembelajaram dan premrograman dengan menggunakan data masa lampau.

3.5 *Text Mining*

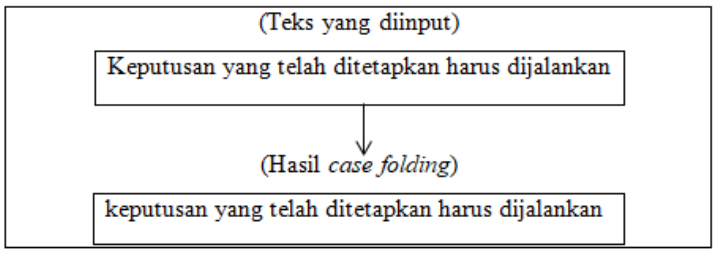
Text mining merupakan salah satu teknik yang dapat digunakan untuk klasifikasi, yang dapat berupa variasi dari data *mining* untuk menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar. *Text mining* merupakan suatu proses menambang suatu data yang berupa teks yang bersumber dari data tersebut (Maria, 2014). Data yang biasanya diperoleh dari dokumen dan digunakan untuk mencari kata-kata yang dapat mewakili isi dari dokumen tersebut, maka dapat di analisa hubungan antar dokumen.

3.6 *Pre-Procesing*

Pada tahap *preprocessing* dilakukan agar teks dapat menjadi data yang dapat diolah lebih lanjut (Falahah & Nur, 2015). Tujuan dilakukannya *preprocessing* yaitu 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). Adapun tahapan *preprocessing* yaitu:

a. *Case Folding*

Case folding merupakan tahapan untuk mengubah semua huruf yang ada dalam dokumen menjadi huruf kecil (*lowercase*). Huruf yang diubah mulai dari ‘a’ hingga ‘z’ (Damanik, 2014).



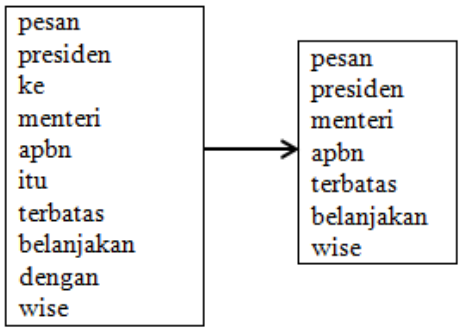
Gambar 3. 3 Proses *Case folding*

b. *Cleaning*

Merupakan suatu proses pembersihan kata pada dokumen dengan menghilangkan tanda baca seperti koma (,), titik (.), titik koma (;), titik dua (:), *mention*, RT, *hashtag*, dan lainnya yang kurang penting untuk mengurangi *noise* (Luqyana, Choslissodin, & Perdana, 2018).

c. *Filtering*

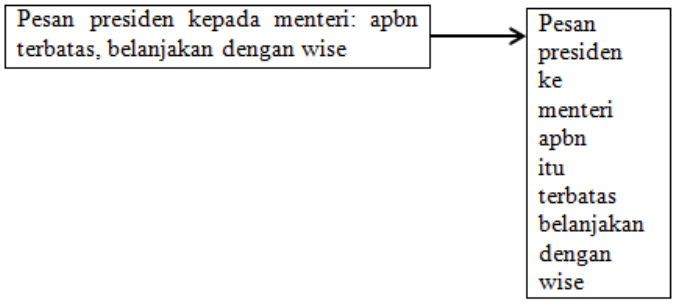
Dalam tahap *filtering* dapat menghilangkan kata yang kurang penting seperti kata ganti menggunakan *stopword*.



Gambar 3. 4 Proses *Filtering*

d. *Tokenizing*

Tokenizing merupakan proses yang digunakan untuk memotong dokumen menjadi pecahan kecil yang dapat berupa kalimat, bab, dan kata pada proses ini akan menghilangkan *whitespace* (Luqyana, Choslissodin, & Perdana, 2018).



Gambar 3. 5 Proses *Tokenizing*

3.7 *Word Cloud*

Word cloud merupakan suatu bentuk analisis data yang mengekstrak model untuk menggambarkan kelas data yang penting, seperti model yang disebut klasifikasi. Dalam proses w*ord cloud*, diketahui bahwa apabila frekuensi kemunculan suatu kata tersebut sering maka hasil yang ditampilkan smakin besar.

Gambar 3. 6 Tampilan *Word Cloud*

(Sumber : Setiabudi, 2015)

3.8 Asosiasi Kata

Asosiasi kata dapat digunakan untuk mengetahui kata apa saja yang sering muncul pada sebuah dokumen. Asosiasi kata juga dapat mengetahui keterkaitan dan hubungan antar kata, misalnya antar dua kata atau lebih digunakan secara bersamaan dalam sebuah dokumen. Dalam asosiasi kata dapat juga dilihat dari nilai korelasi antar kata, dimana nilai korelasi berkisar antara -1 sampai 1. Jika nilai mendekati 1 atau -1 maka hubungan antar kata tersebut semakin kuat, sedangkan jika nilai mendekati 0 maka hubungan antar kata semakin lemah (Pratiwi & Widodo, 2017).

3.9 Pembobotan Kata *Term Frequency - Inverse Document Frequency*

Pemberian bobot hubungan pada suatu kata (*term*) terhadap dokumen sering dikenal dengan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Dengan menggunakan TF-IDF dapat menjadi ukuran statistik yang digunakan untuk mengetahui seberapa penting sebuah kata dalam sebuah dokumen.

Perhitungan pembobotan menggunakan TF-IDF terdapat beberapa tahap yaitu (Luqyana, Cholissodin, & Perdana, 2018):

1. Perhitungan *Term Frequency* (TF)

$$tf_{i,j} = \frac{n_{i,j}}{\sum n_{i,j}} \quad (3.1)$$

Keterangan :

$tf_{i,j}$ = Frekuensi *term*

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

2. Perhitungan *Document Frequency* (IDF)

$$P(X \mid Z) = \frac{P(Z \mid X)P(X)}{P(Z)} \quad (3.2)$$

Keterangan :

idf = *Inverse Document Frequency*

N = Jumlah dokumen

df_i = Jumlah frekuensi dokumen yang mengandung *term*

3. Perhitungan bobot TF-IDF

$$W_{i,j} = tf_{i,j} \times idf_i \quad (3.3)$$

Keterangan :

$W_{i,j}$ = Bobot TF-IDF

idf_i = *Inverse Document Frequency*

$tf_{i,j}$ = Frekuensi suatu kata

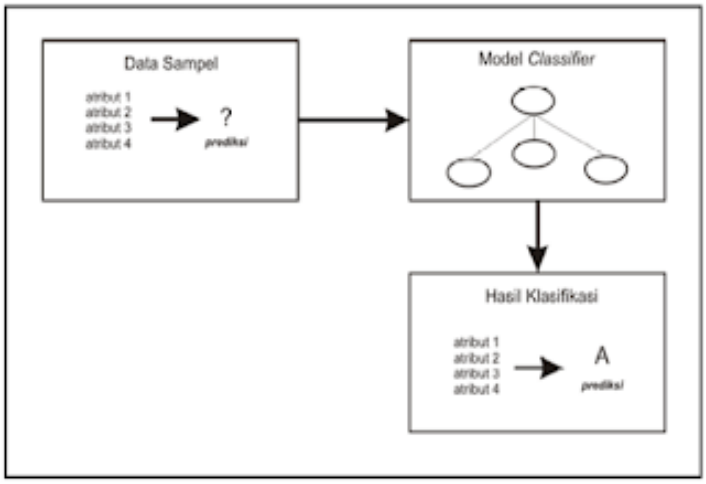
3.10 Analisis Sentimen

Analisis sentimen merupakan proses memahami, mengekstrak, dan mengolah data tekstual secara otomatis untuk mendapatkan informasi yang terkandung dalam suatu kalimat sehingga menjadi informasi yang bermanfaat (Akbari, Novianty, & Setianingsih, 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*. *Sentiment classification*

(negatif atau positif) digunakan untuk memprediksi *sentiment polarity* berdasarkan data sentimen dari pengguna (Dang, Zhang, & Chen, 2010).

3.11 Klasifikasi

Klasifikasi dilakukan untuk menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi terdapat dua pekerjaan yang dilakukan yaitu membangun model sebagai *prototipe* untuk disimpan sebagai memori dan penggunaan model tersebut untuk melakukan prediksi pada suatu objek data lain agar diketahu dikelas mana data tersebut dalam model yang telah disimpan (Prasetyo, 2012).



Gambar 3. 7 Proses Klasifikasi

(**Sumber :** Han dan Kamber, 2006)

Menurut Han dan Kamber (2006) terdapat beberapa persiapan yang dilakukan untuk mendapatkan hasil klasfikasi yaitu:

1. Pembersihan Data

   Pembersihan data dilakukan untuk mengurangi data yang cacat dalam data *training*, terdapat beberapa metode yang digunakan seperti *smoothing* untuk menghilangkan *noise* data dan melengkapi jika terdapat data yang hilang.

2. Analisis Relevansi

   Atribut-atribut yang telah digunakan dalam proses klasifikasi memungkinkan terdapat atribut yang sangat berhubungan antara satu sama

lain, kedua atribut ini memiliki kemiripan sehingga proses klasifikasi menjadi kurang optimal sehingga lebih baik jika salah satu atribut tersebut dibuang.

3.12 Teorema Bayes

Bayes merupakan teknik prediksi yang mengacu konsep probabilistik sederhana berdasarkan penerapan teorema bayes dengan asumsi independensi yang kuat (Prasetyo, 2012). Dalam Bayes (terutama *Naive bayes*) independensi yang kuat yaitu bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama. Misalkan X dan Z merupakan kejadian dalam ruang sampel, berikut merupakan probabilitas bersyarat dalam persamaan (3.4) (Larose, 2006).

$$P(X \mid Z) = \frac{P(X \cap Z)}{P(Z)} \tag{3.4}$$

Dimana P(X∩Z) adalah probabilitas interaksi X dan Z, dan P(Z) adalah probabilitas Z. Untuk $P(Z \mid X) = \frac{P(Z \cap X)}{P(X)}$ sehingga $P(X \cap Z)$ = P (X|Z) P(Z). Nilai $P(X \cap Z)$ kemudian disubtitusikan ke dalam persamaan (3.4), maka diperoleh persamaan 3.5:

$$P(X \mid Z) = \frac{P(Z \mid X)P(X)}{P(Z)} \tag{3.5}$$

3.13 *Naïve Bayes Classifier*

Klasifikasi *bayesian* merupakan pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas dalam suatu kelompok kelas. Penjelasan klasifikasi *naive bayes* merupakan proses klasifiksai yang diperlukan beberapa petunjuk untuk menetukan kelas apa yang sesuai untuk sampel yang akan dianalisis. Metode klasifikasi *naive bayes* diuraikan menjadi persamaan 3.6:

$$P(C \mid F_1, F_2, ..., F_n) = \frac{P(C)P(F_1, F_2, ..., F_n \mid C)}{P(F_1, F_2, ..., F_n)} \tag{3.6}$$

Dari persamaan 3.6, variabel F menjelaskan karakterikstik petunjuk yang digunakan dalam proses klasifikasi dan variabel C menjelaskan kelas. Pada

persamaan 3.6 menjelaskan peluang masuknya sampel karakteristik tertentu dalam kelas C (*posterior*) merupakan peluang munculnya kelas C (sebelum masuknya sampel di sebut *prior*) yang dikalikan dengan peluang munculnya karakteristik sampel pada kelas C (*likelihood*), dibagi dengan peluang munculnya karakteristik-karakteristik sampel secara global (*evidence*). Rumus pada NBC dapat diuraikan pada persamaan 3.7:

$$Posterior = \frac{Prior \times Likelihood}{Evidence} \tag{3.7}$$

Pada persamaan di atas, nilai *evidence* tetap untuk tiap kelas dalam satu sample. *Posterior* merupakan nilai yang akan dibandingkan dengan nilai *posterior* kelas lainnya dalam menentukan suatu sampel kedalam kelas yang akan diklasifikasikan. Uraian dalam persamaan *naive bayes* dijelaskan dengan menguraikan ($C|F,....,F_n$) menggunakan aturan perkalian seperti persamaan 3.8:

$$P(C|F_1,F_2,....F_n) = P(C)P(F_1,...F_n|C) \tag{3.8}$$
$$= P(C)P(F_1|C)P(F_2,...,F_n|C,F_1)$$
$$= P(C)P(F_1|C)P(F_2|C,F_1)P(F_3,...,F_n|C,F_1,F_2,F_3)$$
$$= P(C)P((F_1|C)P(F_2|C,F_1)P(F_3|C,F_1,F_2)P(F_4,...,F_n|C,F_1,F_2,F_3)$$
$$= P(C)P((F_1|C)P(F_2|C,F_1)P(F_3|C,F_1,F_2)...P(F_n|C,F_1,F_2,F_3,....,F_{n-1})$$

Dari penjabaran di atas dalam persamaan 3.8 dapat diketahui bahwa penjabaran tersebut mengakibatkan semakin banyak dan kompleknya suatu faktor-faktor yang dapat berpengaruh terhadap nilai probabilitas, dimana hampir tidak mungkin apabila dianalisa satu persatu. Sehingga perhitungan pada persamaan 3.8 menjadi semakin sulit dilakukan. Pada metode NBC inilah digunakan asumsi bahwa independensi yang sangat tinggi (*naif*), dimana setiap masing-masing petunjuk untuk $F_1,F_2,F_3,....,F_n$ saling bebas (*independence*) antara satu sama lainnya. Maka berdasarkan asumsi tersebut berlaku persamaan 3.9 sebagai berikut:

$$P\{F_i \mid F_j\} = \frac{P\{F_i \cap F_j\}}{P\{F_j\}} = \frac{P\{F_i\}P\{F_j\}}{P\{F_j\}} = P\{F_i\} \tag{3.9}$$

dimana i≠j, maka :

$$P\left\{F_i \mid C, F_j\right\} = P\left\{F_i \mid C\right\} \quad (3.10)$$

Pada persamaan 3.10 merupakan persamaan dari model teorema *naive bayes* yang kemudian digunakan untuk proses klasifikasi dalam metode klasifikasi NBC (Saleh, 2015).

Contoh :

Dalam contoh ini dokumen yang di ambil memiliki kategori/*class* yaitu olahraga, teknologi dan otomotif. Pada dokumen yang ke 7 berisikan "Madrid Anti Barcelona" belum memiliki kategori sehingga dilakukan analisis.

Tabel 3. 1 Dokumen Teks

Dokumen	**Teks**	**Kategori**
1.	Barcelona Tumbangkan Madrid	Olah Raga
2.	4G LTE Indosat Sudah Aktif	Teknologi
Dokumen	**Teks**	**Kategori**
3.	Ancelolti : Barcelona "Bantu" Madrid	Olah Raga
4.	Oli Mesin Anti Panas	Otomotif
5.	Barcelona VS Madrid	Olah Raga
6.	Jaringan Baru 4G LTE	Teknologi
7.	Madrid Anti Barcelona	?

Terdapat 17 kata yang terdapat pada dokumen tabel **3.1** diatas, yaitu 4G, Aktif, Ancelotti, Anti, Bantu, Barcelona, Baru, Indosat, Jaringan, LTE, Madrid, Mesin, Oli, Panas, Sudah, Tumbangkan, VS. Dari kumpulan dokumen pada tabel 3.1 akan berbentuk term document matrix sebagai berikut:

Tabel 3. 2 *Term Documen Matrix*

Doc	**4G**	**Aktif**	**Ancelotti**	**Anti**	**Bantu**	**Barcelona**	**Baru**	**Indosat**	**Jaringan**	**LTE**
1						1				
2	1	1						1		1
3			1		1	1				
4				1						

5						1				
6	1						1		1	1
7						1				

Doc	Madrid	Mesin	Oli	Panas	Sudah	Tumbangkan	VS	Class
1	1					1		OlahRaga
2					1			Teknologi
3	1							OlahRaga
4		1	1	1				Otomotif
5	1						1	OlahRaga
6								Teknologi
7	1							?

Kemudian dokumen berdasarkan kategori atau class olahraga

Tabel 3. 3 Dokumen Kategori Olahraga

Doc	4G	Aktif	Ancelotti	Anti	Bantu	Barcelona	Baru	Indosat	Jaringan	LTE
1						1				
3			1		1	1				
5						1				

Doc	Madrid	Mesin	Oli	Panas	Sudah	Tumbangkan	VS	Class
1	1					1		Olah Raga
2	1							Olah Raga
3	1						1	Olah Raga

Kemudian dihitung P(Ancelotti|Olahraga), P(Bantu|Olahraga), P(Barcelona|Olahraga), P(Madrid|Olahraga), P(Tumbangkan|Olahraga), dan P(VS|Olahraga). Pada dokumen class olah raga terdapat 10 kata .

$$P(OlahRaga) = \frac{3}{6} = 0,5$$

$$P(Ancelotti \mid OlahRaga) = \frac{1+1}{10+17} = 0,074$$

$$P(Bantu \mid OlahRaga) = \frac{1+1}{10+17} = 0,074$$

$$P(Barcelona \mid OlahRaga) = \frac{3+1}{10+17} = 0,148$$

$$P(Madrid \mid OlahRaga) = \frac{2+1}{10+17} = 0,111$$

$$P(Tumbangkan \mid OlahRaga) = \frac{1+1}{10+17} = 0,074$$

$$P(VS \mid OlahRaga) = \frac{1+1}{10+17} = 0,074$$

Kemudian dokumen berdasarkan kategori teknologi dihitung P(4G|Tekonologi), P(Aktif|Tekonologi), P(Baru|Tekonologi), P(4G|Tekonologi), P(Indosat|Olahraga), P(Jaringan|Tekonologi), dan P(Sudah|Tekonologi). Terdapat 9 kata pada dokumen yang memiliki class Teknologi.

Tabel 3. 4 Dokumen Kategori Teknologi

Doc	**4G**	**Aktif**	**Ancelotti**	**Anti**	**Bantu**	**Barcelona**	**Baru**	**Indosat**	**Jaringan**	**LTE**
2	1	1						1		1
6	1						1		1	1

Doc	**Madrid**	**Mesin**	**Oli**	**Panas**	**Sudah**	**Tumbangkan**	**VS**	**Class**
2					1			Teknologi
6								Teknologi

$$P(Teknologi) = \frac{2}{6} = 0,334$$

$$P(4G \mid Teknologi) = \frac{2+1}{9+17} = 0,1153$$

$$P(Aktif \mid Teknologi) = \frac{1+1}{9+17} = 0,0769$$

$$P(Baru \mid Teknologi) = \frac{1+1}{9+17} = 0,0769$$

$$P(Indosat \mid Teknologi) = \frac{1+1}{9+17} = 0,0769$$

$$P(Jaringan \mid Teknologi) = \frac{1+1}{9+17} = 0,0769$$

$$P(LTE \mid Teknologi) = \frac{2+1}{9+17} = 0,1153$$

$$P(Sudah \mid Teknologi) = \frac{1+1}{9+17} = 0,0769$$

Kemudian dokumen berdasarkan kategori otomotif dihitung P(Anti|Otomotif), P(Mesin|Otomotif), P(Oli|Otomotif), dan P(Panas|Otomotif). Terdapat 4 kata pada dokumen yang memiliki class Otomotif.

Tabel 3. 5 Dokumen Kategori Otomotif

Doc	4G	Aktif	Ancelotti	Anti	Bantu	Barcelona	Baru	Indosat	Jaringan	LTE
4				1						

Doc	Madrid	Mesin	Oli	Panas	Sudah	Tumbangkan	VS	Class
4		1	1	1				Teknologi

$$P(Otomotif) = \frac{1}{6} = 0{,}167$$

$$P(Anti \mid Otomotif) = \frac{1+1}{4+17} = 0,0952$$

$$P(Mesin \mid Otomotif) = \frac{1+1}{4+17} = 0,0952$$

$$P(Oli \mid Otomotif) = \frac{1+1}{4+17} = 0,0952$$

$$P(Panas \mid Otomotif) = \frac{1+1}{4+17} = 0,0952$$

Kemudian dihitung pada dokumen 7 berisikan˝"Madrid Anti Barcelona"

Hasil probabilitas pada class OlahRaga :

P(Olahraga) =P(OlahRaga) + P(Barcelona|Olahraga) + P(Anti|Olahraga) + P(Madrid|Olahraga)

= 0.5 + 0.1481 + 0.0370 + 0.1481

= 0.8332

Hasil probabilitas pada class Teknologi :

P(Teknologi) =P(Teknologi) + P(Barcelona|Tekonologi) + P(Anti|Teknologi) + P(Madrid|Teknologi)

= 0.334 + 0.0385 + 0.0385 + 0.385

= 0.4494

Hasil probabilitas pada class Otomotif :

P(Otomotif) =P(Otomotif) + P(Barcelona|Otomotif) + P(Anti|Otomotif) + P(Madrid|Otomotif)

= 0.167 + 0.0476 + 0.0952 + 0.0476

= 0.3574

Sehingga diperoleh P(Olah Raga) sebesar 0.8332, P(Teknologi) sebesar 0.4494, dan P(Otomotif) sebesar 0.3574. Karena P(Olahraga) > P(Teknologi) > P(Otomotif) maka dapat disimpulkan bahwa dokumen 7 tersebut dikategorikan sebagai dokumen Olah Raga.

3.14 *Confusion Matrix*

Untuk menggambarkan seberapa baik sistem klasifikasi yang digunakan menggunakan pengukuran kinerja dari suatu sistem. Salah satu metode yang

digunakan sebagai pengukuran kinerja klasifikasi yaitu *confusion matrix*. *Confusion matrix* menurut Han dan Kamber dalam penelitian (Fluorida, 2018) dapat diartikan sebagai alat yang berfungsi untuk melakukan analisis apakah *clasifier* tersebut dapat mengenali tuple dari kelas yang berbeda. *Confusion matrix* mengandung nilai *true* positif, *true* negatif, *false* positif, dan *false* negatif. Nilai dari *true* positif dan *true* negatif memberikan informsai bahwa ketika *clasifier* dalam melakukan klasifikasi data yang bernilai benar, dan sedangkan *false* negatif dan *false* positif memberikan informasi bahwa ketika *clasifier* salah dalam melakukan pengklasifikasian data.

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. 6 Confusion Matriks

	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:

$$Recall = \left[\frac{TP}{TP + FN} \right] \times 100 \quad (3.10)$$

$$Presisi = \left[\frac{TP}{FP + TP} \right] \times 100 \quad (3.11)$$

$$Akurasi = \left[\frac{TP + TN}{TP + TN + FP + FN} \right] \times 100 \quad (3.12)$$

$$Spesificity = \left[\frac{TN}{TN + FP} \right] \times 100 \quad (3.13)$$

False Positive Rate (FPR) = 1- spesificity (3.14)

$$AUC = \left[\frac{1 + Recall - FPR}{2} \right] \quad (3.15)$$

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. 7 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)