Hak cipta dan penggunaan kembali:

Lisensi ini mengizinkan setiap orang untuk menggubah, memperbaiki, dan membuat ciptaan turunan bukan untuk kepentingan komersial, selama anda mencantumkan nama penulis dan melisensikan ciptaan turunan dengan syarat yang serupa dengan ciptaan asli.

Copyright and reuse:

This license lets you remix, tweak, and build upon work non-commercially, as long as you credit the origin creator and license it on your new creations under the identical terms.

Perpustakaan Univeristas Multimedia Nusantara

BAB II
LANDASAN TEORI

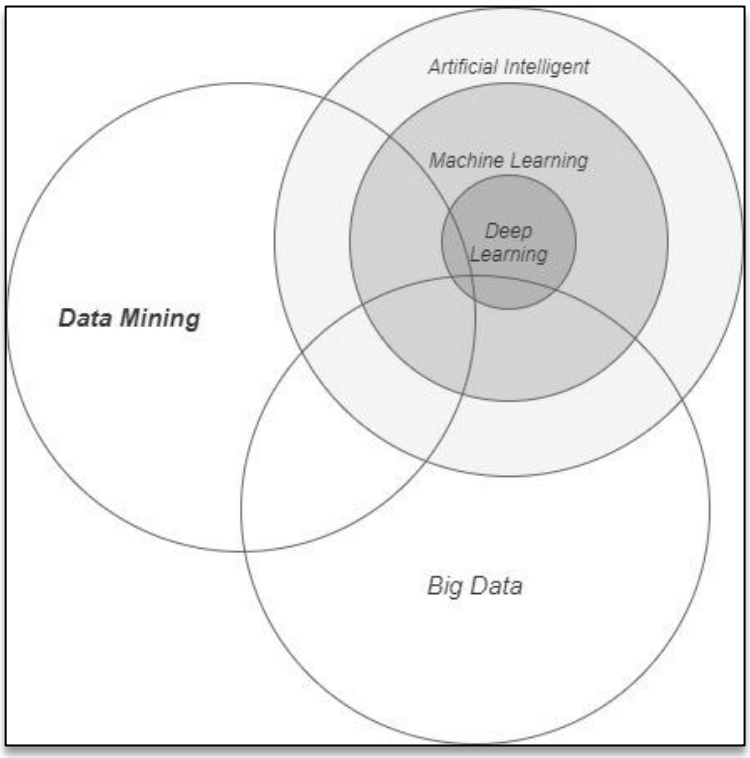
2.1. Big Data

Seiring dengan banyaknya perkembangan teknologi dibeberapa tahun terakhir data semakin beragam dan kompleks (Suyanto, 2019). Menurut International Data Corporation (IDC), volume data pada tahun 2011 mencapai 1,8 *zettabyte* atau 1,8 trilyun *gigabytes*. Pada tahun 2013, volume data sudah menjadi 2,8 *zettabytes* dan diperkirakan akan terus meningkat pesat hingga mencapai 44 *zettabytes* di tahun 2020 (Turner, 2014). Oleh karena itu, sekarang ini muncullah sebuah istilah *big data* yang menggambarkan volume data yang sangat besar yang terstruktur ataupun tidak (Suyanto, 2019).

Big Data terdiri dari empat karakteristik yaitu berukuran sangat besar (*high-volume*), sangat bervariasi (*high-variety*), kecepatan pertumbuhannya tinggi (*high-velocity*), dan sangat tidak jelas (*high-veracity*). Jika melihat *big data* secara keseluruhan tentu manusia akan bingung untuk mengerti dan memahaminya, oleh karena itu *big data* butuh untuk dianalisis sehinga manusia dapat memahami informasi, mendapatkan pengetahuan serta mengambil keputusan dengan baik (Suyanto, 2019).

2.2. *Data Mining*

Data Mining adalah langkah analisis terhadap proses penemuan pengetahuan di dalam basis data atau *knowledge discovery in database* yang disingkat KDD (Usama Fayyad, 1996). Pengetahuan tersebut dapat diartikan sebagai pola data atau relasi antar data valid yang belum diketahui sebelumnya (Suyanto, 2019). *Data Mining* merupakan gabungan dari beberapa ilmu komputer yang diartikan sebagai proses penemuan bentuk atau pola-pola baru dari data yang sangat besar (*Big Data*). Ruang lingkup *data mining* meliputi dari irisan metode *artificial intelligence*, *machine learning*, statistika, dan sistem basis data (Clifton, 2010). Dalam banyak literatur, sebagian ahli mendefinisikan *data mining* untuk mengekstrak pola-pola penting dari kumpulan-kumpulan data yang sangat banyak atau bervolume besar (*big data*) dapat dipelajari atau ditemukan dalam bidang *machine learning* sehingga *machine learning* dapat disebut sebagai teknik-teknik yang mendukung dan digunakan dalam *data mining*. *Machine learning* juga sering berhubungan dengan *deep learning* dan *artificial intelligence* (Suyanto, 2019).



Gambar 2. 1 Relasi data mining, big data, artificial intelligence, machine learning, dan deep learning.

Pada Gambar 2. 1, *deep learning* merupakan sub bagian dari *machine learning*, sedangkan *machine learning* sendiri merupakan sub bagian dari *artificial intelligence*. Relasi yang ada pada Gambar 2. 1 yaitu untuk melakukan pengolahan *big data* dengan *data mining* maka dibutuhkan teknik atau sistem *artificial intelligence* yang mengandung subbagian dan sub-subbagian yaitu *machine learning*, dan *deep learning*. Dengan kata lain *artificial intelligence* digunakan untuk mempermudah penambangan data atau *data mining* (Suyanto, 2019).

Tujuan dari *data mining* adalah untuk mengekstrak atau mengambil intisari pengetahuan dari kumpulan data untuk mendapatkan struktur yang dapat dipahami

oleh manusia. Proses *data mining* meliputi beberapa faktor seperti basis data, manajemen data, prapemrosesan data, pertimbangan model dan inferensi, ukuran ketertarikan, pertimbangan kompleksitas, pasca pemrosesan terhadap struktur yang ditemukan, visualisasi, dan *online updating* (Suyanto, 2019).

Umumnya fungsi data mining dibagi menjadi dua, yaitu deskriptif dan prediktif. Deskriptif berarti dapat dipakai untuk mencari pola-pola dan menjelaskan karakteristik data, sedangkan prediktif artinya untuk membuat sebuah model pengetahuan untuk melakukan prediksi (Suyanto, 2019).

Berdasarkan fungsionalitas data mining, ada beberapa proses yang dapat dilakukan (Usama Fayyad, 1996):

1. Klasifikasi (*Classification*), yaitu mengeneralisasi struktur yang sudah diketahui untuk diaplikasikan pada data yang baru. Contohnya seperti klasifikasi penyakit ke dalam beberapa jenis, klasifikasi surat elektronik apakah termasuk spam atau bukan.
2. Klasterisasi (*clustering*), yaitu mengelompokkan data yang tidak diketahui kelas-kelasnya ke dalam beberapa klaster tertentu berdasarkan kemiripan atributnya.
3. Regresi (*regression*), yaitu fungsi yang membuat model data dengan kesalahan prediksi yang sedikit mungkin.
4. Deteksi Anomali (*anomaly detection*), yaitu fungsi untuk meneliti atau mengidentifikasi data yang tidak biasa, bisa berupa pencilan (*outlier*),

perubahan (deviasi) yang sangat penting dan butuh investigasi yang lebih lanjut.

5. Pembelajaran aturan asosiasi (*asociation rule learning*) atau pemodelan ketergantungan (*dependency modelling*) yaitu fungsi mencari relasi antar variabel.

6. Perangkuman (*summarization*), yaitu fungsi yang menyediakan data yang sudah direpresentasikan sehingga jadi lebih sederhana, termasuk visualisasi dan membuat laporan.

2.3. *Twitter Crawling*

Crawling data adalah sebuah tahap di dalam penelitian yang bertujuan untuk mengumpulkan, mengambil atau mengunduh kumpulan data dari suatu sumber atau database (Jaka Eka Sembodo et al, 2016). Jadi *Twitter Crawling* merupakan suatu proses untuk mengambil atau mengunduh data dari server twitter dengan bantuan Application Programming Integration (API) twitter baik berupa data user maupun data *tweet*. Pengumpulan data dalam penelitian ini menggunakan data dari server twitter berupa kicauan atau *tweet* beserta atribut-atributnya.

Tabel 2. 1. Atribut-atribut yang didapatkan setelah *crawling*.

Atribut	Keterangan
screen_name	Nama akun unik yang mengidentifikasikan pengguna twitter
username	Nama unik yang mengidentifikasikan pengguna twitter
user_id	Nomor unik dari akun twitter
tweet_id	Nomor unik dari *tweet* akun pengguna
tweet_url	Sumber lokasi *tweet*
timestamp	Penanda tanggal dan jam *tweet* dibuat
timestamp_epochs	Penanda waktu *tweet* dibuat dalam bentuk sistem Unix
text	Isi teks atau kalimat dari *tweet*
text_html	Teks *tweet* dalam bahasa untuk membuat halaman web
links	Berisi link yang terdapat dalam *tweet*
hashtags	Berisi tagar yang ada dalam tweet untuk mengelompokkan tweet dengan tujuan memudahkan pencarian
has_media	Penanda sebuah *tweet* memiliki gambar, video atau media lain atau tidak
img_urls	Berisi url dari gambar yang ada dalam *tweet*
video_url	Berisi url dari video yang ada dalam *tweet*
likes	Jumlah like dalam *tweet*
retweets	Jumlah berapa kali *tweet* di*retweet* (diposting ulang)
replies	Jumlah balasan *tweet*
is_replied	Penanda apakah *tweet* ada yang membalas
is_reply_to	Penanda apakah *tweet* merupakan sebuah balasan
parent_tweet_id	Nomor unik *tweet* yang dibalas
reply_to_users	Data user yang dibalas *tweet*nya

2.4. Python

Python adalah *sebuah bahasa pemrograman* yang berorientasi pada objek. Logo Python digambarkan pada Gambar 2. 2. Python dapat digunakan dalam pembangunan dan pengembangan perangkat lunak. Python juga dapat dijalankan di berbagai macam sistem operasi. Saat ini, Python merupakan bahasa pemrograman yang cukup populer bagi para penggunanya di bidang data science dan analisis, karena terdapat cukup banyak library yang menyediakan fungsi untuk analisis data

dan fungsi dalam machine learning, data preprocessing tools, serta visualisas. Secara umum, python memiliki ciri-ciri seperti (Timothy Orvin Edwardo, 2020):

a. Cukup banyak mendukung *library*.

b. Bahasanya cukup mudah dipahami.

b. Bahasa yang *interpreted* karena kode pemrograman dijalankan atau dieksekusi satu per satu dan melakukan *debugging* dengan lebih mudah dibandingkan dengan bahasa yang di*compile*.

c. Bahasa yang mudah dipakai di mana-mana karena hanya membuat kode satu kali untuk menjalankan di *platform* lain.

d. Bahasa yang open-source atau bisa dikembangkan oleh siapa pun.

Bahasa pemrograman python ini dibuat oleh Guido van Rossum. Awalnya, pembuatan bahasa pemrograman ini adalah untuk bahasa *script* tingkat tinggi pada sistem operasi terdistribusi Amoeba. Bahasa pemrograman ini menjadi umum digunakan untuk kalangan engineer seluruh dunia dalam pembuatan perangkat lunaknya, bahkan beberapa perusahaan menggunakan python sebagai pembuat perangkat lunak komersial (Tyan Suhesti, 2014).

Gambar 2. 2. Logo Python.

Sebagai bahasa yang sering digunakan, bahasa pemrograman python memiliki alasan untuk bisa menjadi populer (Timothy Orvin Edwardo, 2020):

a. Menyediakan *open-source library*, *frameworks*, dan *tools* untuk data mining,

b. Bahasa yang cukup mudah dipahami. Penulisan kode program di python relatif lebih singkat.

c. Multifungsi, tidak hanya untuk data processing, namun python juga bisa untuk mengerjakan tugas lain seperti membuat website dan tampilan GUI (*Graphical User Interface*).

Dalam pengerjaan data mining dengan python, dapat menggunakan Integrated Development Enviroment (IDE) seperti Jupyter Notebook atau dengan *pengolah teks* seperti Sublime, Notepad, Notepad++ dengan extension .py.

Contoh proses yang dilakukan oleh python untuk proyek *data mining* sederhana (Timothy Orvin Edwardo, 2020):

a. Memasukkan data mentah. Data dapat diimpor ke dalam python dengan menggunakan *library* yang memiliki fungsi untuk mengimpor data dengan format csv ke python.

b. Eksplorasi data dan praproses data. Proses ini dapat dilakukan dengan mudah, karena di dalam python sudah terdapat fungsi untuk melihat persebaran data dan melakukan manipulasi data untuk mengatasi data yang tidak sesuai.

c. Pemodelan *data mining*. Proses ini dapat dilakukan dengan tersedianya fitur-fitur dan metode-metode *machine learning* yang cukup lengkap. Contoh metodenya seperti *neural network* (NN), *decision tree*, *support vector machine* (SVM), *random forest*, *regression tree*, *logistic regression*.

d. Evaluasi model. Model yang telah dibuat dari pemodelan *data mining*, seperti menghitung akurasi (*accuracy*), spesifisitas, sensitivitas, presisi (*precision*), *error rate* dapat ditampilkan dengan menggunakan fungsi *classification_score*.

e. Merepresentasikan hasil dari model. Fungsi ini dapat memvisualisasikan data yang telah diolah dalam bentuk plot ataupun hasil dalam bentuk web.

2.5. Rapidminer Studio

Rapidminer merupakan salah satu software yang dapat digunakan untuk melakukan proses *data mining*. Logo Rapidminer digambarkan pada Gambar 2. 3.

Proses yang dapat dilakukan oleh Rapidminer dalam *text mining* adalah seputar analisis teks, mengekstrak pola-pola dari dataset yang besar dan menggabungkan atau mengkombinasikannya dengan berbagai macam metode statistika, kecerdasan buatan (*artificeial intelligence*), dan *database* (M. Faid et al, 2019).

Dalam memproses *data mining*, Rapidminer menggunakan metode CRISP-DM (*the cross-industry standard process for data mining*). Ada enam tahapan dalam *data mining* yang dikerjakan oleh Rapidminer (rapidminer.com, 2020) yaitu *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*.

Gambar 2. 3. Logo Rapidminer.

2.6. Analisis Sentimen

Analisis sentimen atau biasa disebut *opinion mining* adalah sebuah pembelajaran perhitungan atas pendapat atau opini, penilaian, sikap, dan emosi orang terhadap suatu entitas, individu, masalah, peristiwa, topik, maupun atribut-atribut lainnya (Bing Liu, 2012). Analisis sentimen biasanya dilakukan untuk mencari opini publik atau pelanggan terhadap suatu produk atau layanan yang dimiliki suatu perusahaan, organisasi atau entitas. Analisis sentimen juga dapat diartikan sebagai pembelajaran sebuah opini, masalah, perasaan maupun emosi dari

seseorang atau publik dalam menanggapi sesuatu dalam bentuk teks atau tulisan di mana dalam menentukan sebuah sentimen, dilakukan dengan melakukan perhitungan dari beberapa kata yang terdapat dalam kalimat, dokumen, ataupun teks (Calvin Alexander et al, 2014). Sebagai contoh, analisis sentimen biasanya dilakukan oleh calon pelanggan yang juga ingin tahu tentang pendapat pengguna yang pernah menggunakan atau mengkonsumsi suatu produk atau layanan sebelum mereka menggunakan layanan atau membeli produk. Contoh lainnya adalah analisis sentimen dapat digunakan oleh politikus untuk mengetahui bagaimana pandangan publik terhadap dirinya dan terhadap lawan politiknya, dengan begitu politikus dapat menentukan strategi yang tepat untukmengalahkan lawannya. Analisis sentimen juga dapat digunakan untuk melihat kenapa suatu produk menjadi tidak laku terjual atau tidak mendapatkan perhatian lebih dari masyarakat.

2.7. *Cleansing*

Cleansing adalah tahap pembersihan karakter yang tidak diperlukan. Penghapusan karakter yang dilakukan adalah seperti simbol, tanda baca, maupun URL. Fungsi dari *cleansing* adalah untuk mengurangi sampah atau kotoran pada data (Achmad et al, 2018).

2.8. *Case Folding*

Case Folding adalah tahap untuk meratakan jenis huruf, bisa menjadi huruf besar semua atau huruf kecil semua. *Case Folding* berguna agar semua jenis huruf menjadi seragam (Achmad et al, 2018).

2.9. *Tokenization*

Tokenization adalah untuk memecahkan suatu kalimat ataupun dokumen menjadi beberapa bagian atau kata-kata. Hasil dari pemecahan tersebut disebut sebagai token. Ada beberapa model atau jenis token yang dapat digunakan, yaitu unigram, bigram dan trigram (Achmad et al, 2018). Contoh untuk masing-masing unigram, bigram, dan trigram adalah "analisis", "analisis_sentimen", dan "analisis_sentimen_positif".

2.10. *Filtering*

Filtering adalah proses untuk menyaring sesuatu dalam dokumen. Penyaringan yang dilakukan dapat berupa penyaringan kata melalui *stopwords* dan juga penyaringan panjang huruf pada suatu kata (Achmad et al, 2018).

2.11. *Stemming*

Stemming adalah proses mengubah kata yang berimbuhan atau berakhiran ataupun keduanya menjadi kata dasar (Achmad et al, 2018). Contohnya kata "menganalisa" menjadi "analisa".

2.12. *K-Fold Cross Validation*

K-Fold Cross Validation merupakan salah satu jenis pengujian yang berfungsi untuk menilai kinerja proses dari sebuah metode algoritma dengan cara membagi beberapa data secara acak dan mengelompokkan data tersebut sebanyak nilai k pada *k-fold*. Kemudian salah satu kelompok dari *k-fold* tersebut akan menjadi *data test* dan sisanya akan menjadi *data train* (Abbas et al, 2017).

2.13. TF-IDF

TF-IDF adalah suatu metode pemberian bobot yang menyatukan dua buah konsep, yaitu *Term Frequency* dan *Document Frequency*. *Term Frequency* adalah konsep dimana pembobotan dilakukan dengan mencari seberapa sering atau frekuensi dari munculnya suatu istilah dalam suatu dokumen atau teks. Setiap dokumen atau teks biasanya memiliki panjang yang berbeda-beda, sehingga dapat terjadi sebuah kata muncul lebih banyak di dalam dokumen atau teks yang panjang dibandingkan dengan dokumen atau teks yang pendek. Jadi, *Term Frequency* biasanya sering dibagi dengan panjang dokumen atau total kata dalam dokumen (Achmad et al, 2018).

Document Frequency adalah jumlah dokumen atau teks di mana sebuah kata muncul. Semakin kecil frekuensi kemunculannya, maka semakin kecil nilainya. Di dalam perhitungan *Term Frequency*, semua kata yang ada di dalamnya biasanya dianggap penting atau bermakna. Oleh karena itu, sangat perlu untuk menghitung TF-IDF, di mana skor atau nilai dapat didapatkan dengan menggunakan sebuah persamaan (Achmad et al, 2018).

$$W_{t,d} = W_{tf_{t,d}} . idf_t$$

Rumus 2. 1. Rumus *weighting word* pada umumnya.

idf_t = frekuensi *term*

$W_{tf_{t,d}}$ = bobot frekuensi *term*

$W_{t,d}$ = bobot TF-IDF

Contoh proses perhitungan TF-IDF dapat dilihat pada **Tabel 2. 2**, **Tabel 2. 3, Tabel 2. 4, Tabel 2. 5, Tabel 2. 6, Tabel 2. 7.** Pada **Tabel 2. 2**, merupakan tiga kalimat yang digunakan sebagai contoh pembobotan. **Tabel 2. 3** adalah contoh kalimat yang telah diproses siap digunakan dalam pembobotan. **Tabel 2. 4** merupakan data munculnya kata pada sebuah kalimat. **Tabel 2. 5** adalah tabel yang berisi langkah-langkah menghitung *Term Frequency* (TF). **Tabel 2. 6** adalah tabel yang berisi langkah-langkah menghitung *Inverse Document Frequency* (IDF). **Tabel 2. 7** adalah hasil perkalian dari TF-IDF. Semua hasil dari TF-IDF digunakan sebagai model untuk melakukan klasifikasi dengan menggunakkan metode algoritma.

Tabel 2. 2. Contoh kalimat sebelum diproses.

Contoh Kalimat sebelum diproses	
Kalimat 1	Sebuah Program televisi 'Kriminal', tidak dicocokan untuk anak!
Kalimat 2	Tayangan edukasi di televisi itu sangat baik karena mengedukasi
Kalimat 3	Program edukasi itu sangat mendidik

Tabel 2. 3. Contoh kalimat yang sudah melewati proses.

Contoh kalimat yang sudah melewati proses dari *Cleansing* hingga *Stemming*	
Kalimat 1	program televisi kriminal tidak cocok anak
Kalimat 2	tayangan edukasi televisi baik edukasi
Kalimat 3	program edukasi didik

Tabel 2. 4. *Term Occurences.*

Term Occurences									
	program	televisi	kriminal	tidak	cocok	anak	edukasi	baik	didik
Kalimat 1	1	1	1	1	1	1	0	0	0
Kalimat 2	0	1	0	0	0	0	2	1	0
Kalimat 3	1	0	0	0	0	0	1	0	1

Tabel 2. 5. *Term Frequency.*

Term Frequency (TF)										
Langkah 1 - Hitung total token dalam kalimat (*token_number*)										
	program	televisi	kriminal	tidak	cocok	anak	edukasi	baik	didik	*token_number*
Kalimat 1	1	1	1	1	1	1	0	0	0	6
Kalimat 2	0	1	0	0	0	0	2	1	0	4
Kalimat 3	1	0	0	0	0	0	1	0	1	3
Langkah 2 - Rasio kemunculan token / *token_number*										
	program	televisi	kriminal	tidak	cocok	anak	edukasi	baik	didik	*token_number*
Kalimat 1	0,167	0,167	0,167	0,167	0,167	0,167	0,000	0,000	0,000	6
Kalimat 2	0,000	0,250	0,000	0,000	0,000	0,000	0,500	0,250	0,000	3
Kalimat 3	0,333	0,000	0,000	0,000	0,000	0,000	0,333	0,000	0,333	3
Langkah 3 - Cari *normalized term frequency vector*										
	Formula									Hasil
Kalimat 1	SQRT [(0.167)^2 + (0.167)^2 + (0.167)^2 + (0.167)^2 + (0.167)^2 + (0.167)^2 + (0)^2 + (0)^2 + (0)^2]									0,408
Kalimat 2	SQRT [(0)^2 + (0.250)^2 + (0)^2 + (0)^2 + (0)^2 + (0)^2 + (0.500)^2 + (0.250)^2 + (0)^2]									0,612
Kalimat 3	SQRT [(0.333)^2 + (0)^2 + (0)^2 + (0)^2 + (0)^2 + (0)^2 + (0.333)^2 + (0)^2 + (0.333)^2]									0,577
Langkah 4 - Hasil dari Langkah 2 / Hasil dari Langkah 3										
	program	televisi	kriminal	tidak	cocok	anak	edukasi	baik	didik	
Kalimat 1	0,408	0,408	0,408	0,408	0,408	0,408	0,000	0,000	0,000	
Kalimat 2	0,000	0,408	0,000	0,000	0,000	0,000	0,816	0,408	0,000	
Kalimat 3	0,577	0,000	0,000	0,000	0,000	0,000	0,577	0,000	0,577	

Tabel 2. 6. *Inverse Document Frequency.*

Inverse Document Frequency (IDF)									
Langkah 1 - Hitung total kalimat yang memiliki kata "…"									
	program	televisi	kriminal	tidak	cocok	anak	edukasi	baik	didik
Kalimat 1	1	1	1	1	1	1	0	0	0
Kalimat 2	0	1	0	0	0	0	2	1	0
Kalimat 3	1	0	0	0	0	0	1	0	1
	2	2	1	1	1	1	2	1	1
Langkah 2 - ln(Total kalimat yang memiliki kata "…" / Total kalimat)									
Total Kalimat	program	televisi	kriminal	tidak	cocok	anak	edukasi	baik	didik
3	0,405	0,405	1,099	1,099	1,099	1,099	0,405	1,099	1,099

Tabel 2. 7. Term Frequency - Inverse Document Frequency.

Term Frequency - Inverse Document Frequency (TF-IDF)									
TF * IDF									
Kalimat 1	0,166	0,166	0,449	0,449	0,449	0,449	0,000	0,000	0,000
Kalimat 2	0,000	0,166	0,000	0,000	0,000	0,000	0,331	0,449	0,000
Kalimat 3	0,234	0,000	0,000	0,000	0,000	0,000	0,234	0,000	0,634

2.14. Algoritma Klasifikasi

Klasifikasi merupakan sebuah bagian teknik dari *data mining* yang mempelajari sekelompok data untuk menghasilkan aturan-aturan yang dapat mengklasifikasi atau mengenali data-data baru yang belum pernah dipelajari (Suyanto, 2019). Klasifikasi dapat didefinisikan sebagai suatu proses yang digunakan untuk menyatakan suatu objek data sebagai salah satu kategori atau kelas yang sudah didefinisikan sebelumnya (Zaki et al. 2013), jadi untuk mengklasifikasikan suatu objek (kalimat, teks, maupun dokumen) dapat menggunakan beberapa data atau kata yang sudah didefinisikan sebelumnya. Proses pengklasifikasian menggunakan model dari pembobotan dan menggunakan metode perhitungan algoritma yang berbeda-beda. Ada beberapa contoh algoritma klasifikasi yang dapat digunakan yaitu *K-Nearest Neighbor, Naïve Bayes Classifier* dan *Decision Tree*

2.14.1. *K-Nearest Neighbor* (K-NN)

K-NN merupakan bagian salah satu metode klasifikasi *Nearest Neightbor Rule* (NNR). NNR merupakan metode klasifikasi yang dapat memproses di mana data memiliki banyak pencilan-pencilan yang membuat proses klasifikasi menjadi sulit yang tidak memerlukan parameter yang rumit. Istilah "*rule*" menyatakan aturan untuk menentukan berdasarkan tetangga yang terdekat (Suyanto, 2019).

Metode K-NN beroperasi dengan menemukan sejumlah k objek data ataupun pola masukan yang memilih kelas dengan jumlah data terbanyak di antara k data. Untuk menghindari hasil k yang sama, maka sebisa mungkin jumlah k adalah ganjil (Suyanto, 2019).

2.14.2. *Naïve Bayes Classifier*

Metode *Naïve Bayes Classifier* atau yang dikenal dengan teorema Bayes, ditemukan oleh seorang yang bernama Thomas Bayes. Algoritma Naive Bayes juga dapat memprediksi peluang di masa depan berdasarkan kejadian atau data sebelumnya. Ciri utama dari metode *Naïve Bayes Classifier* ini adalah asumsi yang sangat kuat dan naif terhadap suatu independensi dari masing-masing kejadian atau data (Suyanto, 2019).

Keuntungan utama dari penggunan metode *Naïve Bayes Classifier* adalah metode ini hanya membutuhkan sedikit saja jumlah *data train* (data pelatihan) yang digunakan untuk menentukan nilai yg diperlukan dalam proses pengklasifikasian.

2.14.3. *Decision Tree*

Decision tree merupakan salah satu metode klasifikasi yang sangat populer dan cukup banyak digunakan karena praktis. Metode dirancang untuk menemukan model klasifikasi yang tahan terhadap derau. *Decision tree* membangun model klasifikasi berupa sebuah pohon keputusan secara *top-*

down (Suyanto, 2019). Cara kerjanya adalah dengan mengevaluasi semua atribut-atribut menggunakan ukuran-ukuran statistik yang biasanya berupa *information gain*, yang digunakan untuk mengukur efektifitas sebuah atribut saat sedang mengklasifikasikan himpunan sampel data.

Decision tree memiliki beberapa algoritma, namun yang paling populer adalah *Iterative Dychotomizer version* 3 (ID3). ID3 merupakan *decision tree learning* yang melakukan pencarian atau proses secara banyak atau rakus, sehingga belum bisa dipastikan optimal atau tidak (Suyanto, 2019).

2.15. Sosial Media

Media sosial merupakan media yang dapat digunakan oleh setiap masyarakat untuk hidup bersosial secara daring dengan saling berbagi cerita, tulisan, berita, gambar, video, pesan suara dan lain-lain. Disebut sebagai media sosial karena menggambarkan berbagai macam teknologi yang digunakan untuk saling berkolaborasi, bertukar informasi, dan berinteraksi (Taprial, 2012). Dalam sebuah media sosial, setiap individu dapat menggunggah yang diinginkan seperti cerita kehidupan, karya seni dalam bentuk apapun, pendapat publik, keresahan dan lain-lain. Dengan kebebasan dalam bersosialisasi dalam media sosial, memungkinkan untuk menganalisa setiap aktivitas yang terjadi di dalamnya. Media sosial juga dapat dijadikan media analisa berbagai opini yang disalurkan masyarakat untuk suatu objek yang sedang viral.

2.16. Twitter

Twitter adalah sebuah jejaring sosial dan jaringan informasi yang terdiri atas pesan-pesan sepanjang 280 karakter yang disebut cuitan (*tweet*) atau kicauan (help.twitter.com, 2020). Twitter didirikan pada tahun 2006 oleh Jack Dorsey, Biz dan Evan Williams. Dengan menggunakan Ruby on Rails Twitter dirancang pada 2006 oleh Evan Williams dan Biz Stone sedangkan Jack Dorsey bergabung dengan manajemen (britannica.com/topic/Twitter, 2010). Pada awalnya Twitter dibuat seperti SMS gratis dengan elemen jejaring sosial yang akhirnya berkembang dapat saling mengirimkan pesan, gambar maupun video.

Sebagai sebuah jejaring sosial Twitter yang aktif di dunia, twitter memiliki jumlah pengguna harian. Dilansir dari Kompas.com, pada laporan finansial Twitter kuartal ke 3 tahun 2019, pengguna aktif harian Twitter berjumlah 145 juta pengguna. Pada tahun 2019 *Country Industry Head* Twitter Indonesia Dwi Adriansah, menyatakan Indonesia diklaim sebagai salau satu negara yang memiliki pertumbuhan pengguna aktif hariannya paling besar.

Twitter dapat digunakan di gawai pintar maupun di browser. Twitter memiliki berbagai macam fitur yang dapat dijumpai:

2.16.1. Beranda

Beranda merupakan halaman utama sebagai tempat terkumpulnya berbagai macam kicauan (*tweet*) dari akun-akun yang diikuti oleh pengguna.

Biasanya kicauan diperbaharui setiap detiknya. Di halaman ini, pengguna juga dapat menggunggah kicauannya berupa tulisan, gambar, maupun video.

2.16.2. Profil

Profil adalah halaman yang berisi tentang informasi pribadi seperti kicauan pengguna akun, jumlah pengikut, jumlah yang diikuti dan juga biografi atau deskripsi singkat tentang pengguna. Di halaman ini juga pengguna dapat menghapus kicauan yang pernah diunggahnya.

2.16.3. Following

Following berisi nama-nama akun yang pernah diikuti oleh pengguna. Di halaman ini pengguna dapat membatalkan mengikuti akun-akun yang pernah diikuti sebelumnya.

2.16.4. Followers

Follower adalah halaman berisi akun-akun yang mengikuti pengguna. Jika tidak diinginkan untuk diikuti oleh sebuah, pengguna dapat menghapus dengan memblokir akun tersebut.

2.16.5. Mentions

Mentions berisi informasi atau pemberitahuan jika sebuah akun sedang memanggil akun pengguna dalam cuitannya. Pengguna dapat memilih untuk membalas, membiarkan atau melaporkan cuitan akun yang telah memanggilnya.

2.16.6. Favorite

Favorite merupakan kumpulan dari kicauan yang disukai oleh pengguna. Jika pengguna menyukai suatu kicauan dan ingin melihatnya atau mencari kicauan yang disukainya, pengguna tinggal masuk ke halaman ini.

2.16.7. Pesan Langsung

Pesan langsung ini bersifat privasi antara dua akun, sehingga pembicaraan tidak dapat dilihat oleh publik. Pesan langsung biasa digunakan jika ingin membicarakan hal yang cukup penting yang tidak ingin disebarkan di publik.

2.16.8. Tagar (*Hashtag*)

Tagar biasanya ditandai dengan simbol pagar (#). Penggunaan tagar adalah untuk mengelompokkan cuitan sesuai topik pembicaraan, sehingga jika ingin mencari cuitan yang berhubungan dengan sesuatu bisa lebih mudah.

2.16.9. Trending Topic

Trending topic berisi *hashtag* tentang topik-topik pembicaraan yang sedang ramai dibicarakan oleh publik. *Trending topic* juga sebagai acuan untuk melakukan analisis melalui kicauan-kicauan yang berhubungan dengan *trending topic*.