BAB II
TINJAUAN PUSTAKA

Bab ini berisi tinjauan terhadap penelitian-penelitan yang terkait dan landasan teori yang mendukung penelitian ini. Lebih lanjut hal tersebut diuraikan pada sub-sub bab berikut ini.

2.1 Penelitian Terkait

Dalam menyusun penelitian ini, peneliti menggunakan beberapa penelitian sebelumnya yang ada dalam bentuk jurnal. Jurnal-jurnal yang dipilih tentunya berkaitan serta akan digunakan sebagai referensi bagi peneliti nantinya. Adapun penelitian yang dilakukan oleh Ahmad Fathan Hidayatullah, 2016 dengan judul penerapan *text mining* dalam klasifikasi judul skripsi, salah satu masalah yang berkaitan dengan *text classification* yang ditemukan di perguruan tinggi yaitu proses pengelompokkan judul skripsi secara otomatis. Penelitian ini bertujuan untuk membuat model data judul skripsi di bidang informatika menggunakan *Support Vector Machine* (SVM) dan *Naïve Bayes*. Berdasarkan hasil eksperimen, model SVM memiliki akurasi yang lebih rendah dengan perbedaan yang cukup signifikan jika dibandingkan dengan model yang dihasilkan dari algoritma *Naive Bayes*. Pada perhitungan *precision, recall, dan f-score* diketahui bahwa hasil perhitungan ketiganya memiliki pola yang sama dengan perhitungan akurasi. Secara keseluruhan, hasil perolehan *f-score* dengan algoritma *Naïve Bayes* memberikan hasil yang lebih tinggi dibandingkan dengan algoritma SVM

Penelitian tentang Analisis Sentimen juga dilakukan oleh Hermanto, 2016 dengan judul implementasi *text mining* menggunakan *Naïve Bayes* untuk penentuan kategori tugas akhir mahasiswa berdasarkan abstraksinya. Dengan semakin meningkatnya penggunaan aplikasi sistem informasi di berbagai bidang, turut berdampak pada kebutuhan untuk peningkatan kecepatan pemrosesan data. Pemrosesan data yang menggunakan proses semi manual, mempunyai beberapa kendala, di antaranya: waktu proses lebih lama dan besaran data yang diproses menjadi relatif sedikit.

Oleh karena itu, dalam penelitian ini dikembangkan penggunaan *Naïve Bayes* untuk membantu bagian koordinator tugas akhir dalam melakukan pengelompokan proposal tugas akhir. Metode *Naïve Bayes* yang akan diimplementasi ke dalam sistem informasi proposal tugas akhir dapat memberikan sebuah solusi baru untuk menentukan kategori proposal tugas akhir berdasarkan abstraksi yang dibuat mahasiswa. Dalam hasil uji coba metode ini, dapat disimpulkan cukup berhasil dan secara garis besar dapat dijadikan sebagai perangkat bantu dalam melakukan klasifikasi dokumen tugas akhir. Tingkat akurasi berdasarkan pengujian untuk kategori *hardware* dan *networking* mencapai 86%, kategori sistem informasi tingkat akurasi mencapai 80% dan kategori sistem informasi akuntansi mencapai 89%. Secara keseluruhan, berdasarkan jumlah *dataset* yang diujikan dan tingkat keberhasilan yang dicapai, maka sistem ini mempunyai tingkat akurasi 87%.

Metode *Naïve Bayes Classifer* juga dipakai pembelajaran daring seperti yang dilakukan oleh peneliti Samsir, Ambiyar, Unung Verawardina, pada tahun 2021 dengan judul analisis sentimen pembelajaran daring pada *twitter* di masa pandemi COVID-19 menggunakan metode *Naïve Bayes*. Strategi melawan pandemi dengan pembatasan sosial memaksa semua institusi pendidikan menerapkan pembelajaran daring. Namun pembelajaran daring yang awalnya sebagai strategi menjadi kontroversi karena singkatnya proses adaptasi. Perubahan mendadak dari pembelajaran tatap muka ke pembelajaran daring pada skala besar menyebabkan berbagai *respons* di masyarakat. Penelitian ini bertujuan menganalisis opini publik terhadap pembelajaran daring pada masa pandemi COVID-19 di Indonesia pada awal November 2020. Penelitian dilakukan dengan penambangan teks berbasis dokumen pada *twitter* yang dianalisis menggunakan algoritma *Naïve Bayes*. Temuan menunjukkan bahwa pembelajaran daring memiliki 30% sentimen positif, 69% sentimen negatif, dan 1% netral pada periode tersebut. Tingginya sentimen negatif dihasilkan karena ketidak puasaan masyarakat terhadap pembelajaran daring. Beberapa *tweet* menunjukkan kekecewaan dengan kata ‘stres’ dan ‘malas’ merupakan kata yang memiliki frekuensi tinggi dalam percakapan.

Kemudian penelitian yang menggunakan algoritma *Naïve Bayes* juga dilakukan oleh Fajar Ratnawati, 2018 dengan judul implementasi algoritma *Naïve Bayes* terhadap analisis sentimen opini film pada *twitter*. Penelitian ini bertujuan untuk ketika seseorang menulis opini suatu film, maka semua unsur yang ada di dalam film tersebut akan dituliskan. Data opini film pada penelitian ini diambil dari komentar film yang ditulis di *twitter*. Banyaknya opini yang dituliskan di twitter membutuhkan pengklasifikasian sesuai sentimen yang dimiliki agar mudah untuk mendapatkan kecenderungan opini tersebut terhadap film apakah cenderung beropini positif atau negatif. Algoritma yang akan digunakan pada penelitian ini adalah algoritma *Naïve Bayes*. Berdasarkan hasil eksperimen, analisis sentimen yang dapat dilakukan oleh sistem dengan akurasi yang didapat adalah 90 % dengan rincian nilai *precission* 92%, *recall* 90% dan *f-measure* 90%.

Tabel 2.1 Penelitian Terkait

NO	**Peneliti**	**Judul**	**Tahun**	**Masalah**	**Metode**	**Hasil**
1	Ahmad Fathan Hidayatullah	Penerapan Text Mining dalam Klasifikasi Judul Skripsi	2016	Masalah yang berkaitan dengan text classification yang ditemukan di perguruan tinggi yaitu proses pengelompokkan judul skripsi secara otomatis.	Support Vector Machine (SVM).	SVM memiliki akurasi yang lebih rendah dengan perbedaan yang cukup signifikan jika dibandingkan dengan model yang dihasilkan dari algoritma Naive Bayes.

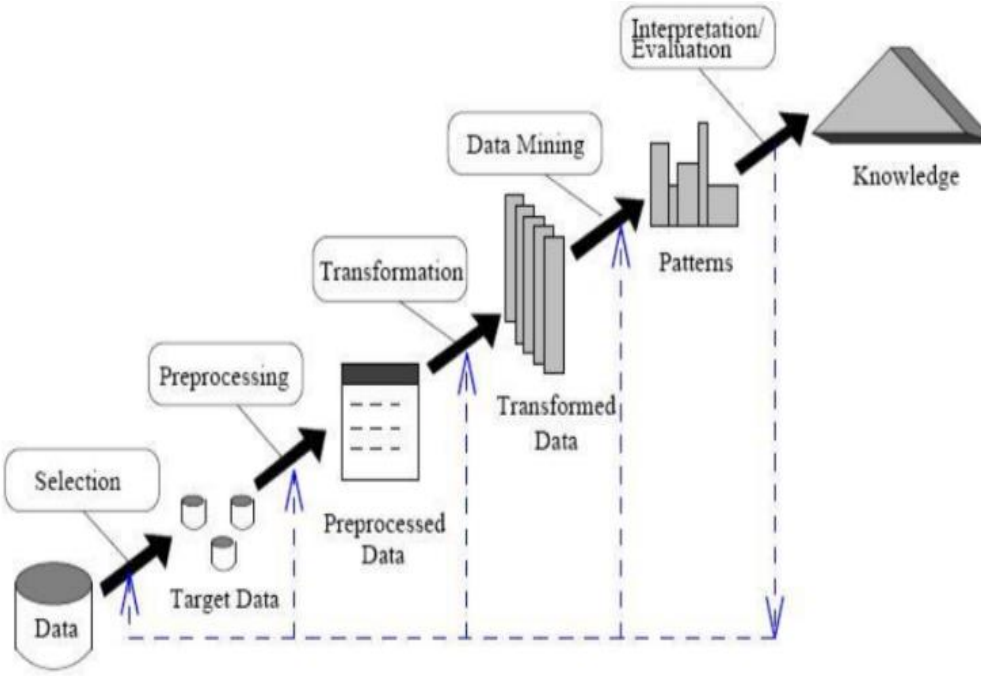
NO	Peneliti	Judul	Tahun	Masalah	Metode	Hasil
2	Agus Hermanto	Implementasi Text Mining Menggunakan Naive Bayes Untuk Penentuan Kategori Tugas Akhir Mahasiswa Berdasarkan Abstraksinya	2016	Pemrosesan data yang menggunakan proses semi manual, mempunyai beberapa kendala, diantaranya: waktu proses lebih lama dan besaran data yang diproses menjadi relatif sedikit sejumlah tempat yang masuk jalur, tetapi kurang ditemui adanya shelter.	Naive Bayes Classifer	Hasil pengujian dapat disimpulkan bahwa pengelompokan tugas akhir dengan menggunakan naive bayes ini berhasil dengan cukup baik dengan tingkat akurasi yang berhasil dicapai dari 105 kali pengujian, adalah 91 berhasil dan 14 kali gagal atau sekitar 87 %
3	Samsir, Ambiyar, Unung Verawardina.	Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes	2018	Pembelajaran daring dianggap sebagai strategi kemudian menjadi kontroversi karena perlu adaptasi. Perbedaan infrastruktur, kualitas koneksi, perangkat yang	Naive Bayes Classifer.	Penelitian pada periode tersebut menunjukkan 30% sentimen positif, 69% sentimen negatif, dan 1% netral. Persepsi negatif

NO	Peneliti	Judul	Tahun	Masalah	Metode	Hasil
				digunakan, dan masih mahalnya kuota internet menjadi hambatan utama. Perubahan mendadak dari pembelajaran tatap muka ke pembelajaran daring pada skala besar menyebabkan berbagai respons di masyarakat.		dihasilkan karena ketidakpuasaan masyarakat terhadap pembelajaran daring. Beberapa twit menunjukkan kekecewaan dengan kata 'stres' dan 'malas' merupakan kata yang memiliki frekuensi tinggi dalam percakapan pada periode tersebut
4	Fajar Ratnawati	Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitte	2018	Banyak pengguna twitter yang memberikan tanggapan tentang film yang pernah mereka tonton dalam bentuk opini berupa pengalaman baik maupun buruk. Hal tersebut menjadikan film	Naive Bayes Classifer.	Hasil pengujian dan analisis yang telah dilakukan dapat disimpulkan pengklasifikasian data opini film berbahasa Indonesia berdasarkan sentimennya

NO	Peneliti	Judul	Tahun	Masalah	Metode	Hasil
				memiliki berbagai macam topik yang informasinya bisa digali kembali.		dapat dilakukan dengan algoritme Naive Bayes Classifier dengan pembagian datasetnya menggunakan 5-fold cross validation. Akurasi tertinggi didapat pada fold kedua yaitu 90%, precision 92%, Recall 90% dan fmeasure 90%

2.2 *Data mining*

Data Mining merupakan sebuah proses penggalian data dengan menemukan pola unik dari beberapa *dataset* dengan jumlah data yang besar, dapat disimpan pada *database*, data *warehouse*, maupun penyimpanan informasi lainnya (Hikmawan et al., 2020). Pada proses implementasi *data mining* juga menjadi satu bagian dari proses dari KDD *Knowledge Discovery in Database*. KDD merupakan proses untuk mengekstrak sebuah pola data dengan menggunakan suatu metode maupun algoritma tertentu.



Gambar 2.1 Alur Proses Knowledge Discovery Database (Mustafa et al., 2018).

Ada pula tahapan *Knowledge Discovery Database* selaku berikut.

1. Informasi *selection*: pemilihan informasi dari sekumpulan informasi operasional.
2. *Preprocessing*: informasi mining butuh dicoba proses cleaning dengan tujuan buat membuang duplikasi informasi.
3. *Transformation*: ialah proses coding pada informasi yang sudah diseleksi.
4. *Interpretation/ Evaluation*: Sesi ini mencakup pengecekan apakah pola ataupun data yang ditemui berlawanan dengan kenyataan ataupun hipotesa yang terdapat tadinya ataupun tidak.
5. *Knowledge*: proses melaksanakan metode visualisasi serta representasi hasil dari pengolahan data.

2.2.1 Pengelompokan Data Mining

Ada beberapa teknik yang dimiliki *data mining* berdasarkan tugas yang bisa dilakukan, yaitu (Mustafa et al., 2018).

1. Deskripsi

   Para peneliti biasanya mencoba menemukan cara untuk mendeskripsikan pola

dan trend yang tersembunyi dalam data.

2. Estimasi

   Estimasi mirip dengan klasifikasi, kecuali variabel tujuan yang lebih kearah numerik dari pada kategori.

3. Prediksi

   Prediksi memiliki kemiripan dengan estimasi dan klasifikasi. Hanya saja, prediksi hasilnya menunjukkan sesuatu yang belum terjadi (mungkin terjadi di masa depan).

4. Klasifikasi

   Dalam klasifikasi variabel, tujuan bersifat kategorik. Misalnya, kita akan mengklasifikasikan pendapatan dalam tiga kelas, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

2.3 Analisis Sentimen

Analisis sentimen dikatakan sebagai *opinion mining* dapat digunakan dalam berbagai kemungkinan domain/entitas dari produk dan jasa, peristiwa sosial dan politik serta kegiatan tertentu lainnya. *Opinion* atau pendapat adalah pusat dari semua aktifitas manusia karena merupakan pemberi pengaruh utama perilaku kita. *Opinion* dan konsep sejenisnya seperti sentimen, evaluasi, sikap, dan emosi adalah subjek studi tentang analisis sentimen.

Analisis sentimen merupakan topik penelitian yang aktif dibawah *Natural Language Proccessing* yang bertujuam untuk membangun sebuah metode yang dapat diimplementasikan menjadi sebuah *tools* yang dapat digunakan untuk mengekstraksi informasi subjektif berupa sentimen atau opini dalam sebuah data *text*. Kecenderungan penelitian tentang analisis sentimen berfokus pada pendapat yang menyatakan suatu sentimen memiliki nilai positif atau negatif (Blidex & Wibowo, 2021).

2.4 *Covid-19*

Covid-19 (Corona Virus Disease/COVID-19) sebuah nama baru yang diberikan oleh *Wolrd Health Organization* (WHO) bagi pasien dengan infeksi *Covid-19* yang pertama kali dilaporkan dari kota Wuhan, Cina pada akhir 2019. Penyebaran terjadi secara cepat dan membuat ancaman pandemi baru. *Covid-19* adalah kumpulan virus yang bisa menginfeksi sistem pernapasan.

Pada banyak kasus, gejala awal infeksi *Covid-19* bisa menyerupai gejala flu, yaitu demam, pilek, batuk kering, sakit tenggorokan, dan sakit kepala. Setelah itu, gejala dapat hilang dan sembuh atau malah memberat. Penderita dengan gejala yang berat bisa mengalami demam tinggi, batuk berdahak bahkan berdarah, sesak napas, dan nyeri pada bagian dada hingga kematian (Hidayani, 2020).

2.5 Vaksin

Vaksin adalah produk biologi yang berisi antigen berupa *mikroorganisme* yang sudah mati atau masih hidup yang dilemahkan, masih utuh atau bagiannya, atau berupa *toksin mikroorganisme* yang telah diolah menjadi *toksoid* atau protein *rekombinan*, yang ditambahkan dengan zat lainnya, yang bila diberikan kepada seseorang akan menimbulkan kekebalan spesifik secara aktif terhadap penyakit tertentu.

Vaksinasi adalah pemberian vaksin yang khusus diberikan dalam rangka menimbulkan atau meningkatkan kekebalan seseorang secara aktif terhadap suatu penyakit, sehingga apabila suatu saat terpajan dengan penyakit tersebut tidak akan sakit atau hanya mengalami sakit ringan dan tidak menjadi sumberpenularan. Vaksinasi program adalah pelaksanaan vaksinasi kepada masyarakat yang pendanaannya ditanggung atau dibebankan pada pemerintah. Vaksinasi gotong royong adalah pelaksanaan vaksinasi kepada karyawan/karyawati, keluarga dan individu lain terkait dalam keluarga yang pendanaannya ditanggung atau dibebankan pada badan hukum/badan usaha (Fitriani Pramita Gurning et al., 2021).

2.6 *Text Mining*

Text mining merupakan bagian dari data mining dimana proses yang dilakukan utamanya adalah melakukan ekstraksi pengetahuan dan informasi dari pola-pola yang terdapat dalam sekumpulan dokumen teks menggunakan alat analisis tertentu. *Text mining* dapat diolah untuk berbagai macamkeperluan diantarannya adalah untuk *summarization*, pencarian dokumen teks dan *sentiment analisys* (Symeonidis et al., 2018).

Text mining merupakan penerapan konsep dan teknik data mining untuk mencari pola dalam teks, yaitu proses penganalisisan teks guna mendapatkan informasi yang bermanfaat untuk tujuan tertentu. Berdasarkan ketidakteraturan struktur data teks, maka proses *text mining* memerlukan beberapa tahap awal yang pada intinya adalah mempersiapkan agar teks dapat diubah menjadi lebih terstruktur. Pada saat ini, *text mining* sudah diterapkan di berbagai bidang, di antaranya.

2.7 *Information Extraction* (Ekstraksi Informasi)

Identifikasi terhadap hubungan dan frase-frase kunci dalam teks dengan mencari urutan yang sudah ditetapkan dalam *text* menggunakan pencocokan pola.

1. *Topic Tracking* (Pelacakan Topik)
   Berdasarkan pada profil *user* dan berbagai dokumen yang dilihat *user*, *text mining* bisa memprediksi dokumen-dokumen lain yang menjadi perhatian/minat *user* tersebut.
2. *Summarization* (Peringkasan)
   Meringkas suatu dokumen untuk menghemat waktu dari sisi pembaca.
3. *Clustering* (Penggugusan)
   Mengelompokkan dokumen-dokumen yang mirip tanpa memiliki kategori yang sudah ditetapkan sebelumnya.
4. *Concept Linking* (Penautan Konsep)
   Menghubungkan berbagai dokumen terkait dengan mengidentifikasi konsep yang digunakan berbsama dan dengan demikian membantu para user untuk menemukan informasi yang barangkali mereka tidak akan temukan dengan menggunakan metode-metode pencarian tradisional.

5. *Question Answering* (Penjawaban Otomatis)

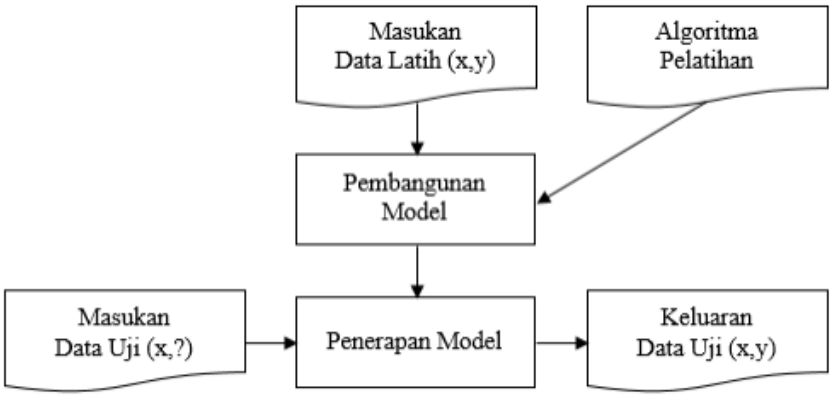
   Menemukan jawaban terbaik pada pertanyaan yang diberikan melalui pencocokan pola berbasis *knowledge*.

2.8 Klasifikasi

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui.

Classification adalah metode yang paling umum pada data mining. Persoalan bisnis seperti *Churn Analysis*, dan *Risk Management* biasanya melibatkan metode *classification*. *Classification* adalah tindakan untuk memberikan kelompok pada setiap keadaaan. Setiap keadaan berisi sekelompok atribut, salah satunya adalah *class attribute*. Metode ini butuh untuk menemukan sebuah model yang dapat menjelaskan *class attribute* itu sebagai fungsi dari *input attribute* yang termasuk dalam *Classification Algorithm* adalah *Decision Trees, Neural Network dan Naives Bayes* (Kurniawan et al., 2020).

Model dalam klasifikasi mempunyai arti yang sama dengan kotak hitam, di mana ada suatu model yang menerima masukan, kemudian mampu melakukan pemikiran terhadap masukan tersebut, dan memberikan jawaban sebagai keluaran dari hasil pemikirannya. Kerangka kerja (*framework*) klasifikasi ditunjukkan pada Gambar 2.2. Pada gambar tersebut disediakan sejumlah data latih (x,y) untuk digunakan sebagai data pembangun model. Model tersebut kemudian dipakai untuk memprediksi kelas dari data uji (x,?) sehingga diketahui kelas y yang sesungguhnya (Maricar et al., 2021).



Gambar 2.2 Proses Klasifikasi (Imron, 2019)

Kerangka kerja seperti yang ditunjukkan pada gambar 2.2 meliputi dua langkah proses, yaitu induksi dan deduksi. Induksi merupakan langkah untuk membangun model klasifikasi dari data latih yang diberikan, disebut juga proses pelatihan, sedangkan deduksi merupakan langkah untuk menerapkan model tersebut pada data uji sehingga kelas yang sesungguhnya dari data uji dapat diketahui, disebut juga proses prediksi.

2.9 *Naïve Bayes*

Naïve Bayes Classifier merupakan pengklasifikasi probabilitas sederhana berdasarkan pada *Teorema Bayes* dikombinasikan dengan "*Naïve*" yang berarti setiap atribut bersifat bebas (*independent*). *Naïve Bayes Classifier* dapat dilatih dengan efisien dalam pembelajaran terawasi (*supervised learning*).

Keuntungan dari klasifikasi adalah bahwa hanya membutuhkan sejumlah kecil data pelatihan untuk memperkirakan parameter (sarana dan varian dari variabel) yang diperlukan untuk klasifikasi. Karena variabel independen diasumsikan, hanya variasi dari variabel untuk masing-masing kelas harus ditentukan, bukan seluruh matriks kovarians. Secara garis besar algoritma Naïve Bayes dapat dijelaskan seperti Persamaan 1 (Hikmawan et al., 2020).

$$P(R|S) = \frac{P(R)P(S|R)}{\mathrm{P(S)}} \_\_\_\_\_ (1)$$

Berikut adalah penjelasan dari notasi-notasi yang digunakan pada Persamaan 1 tersebut.

R : Data yang belum diketahui kelasnya.

S : Hipotesis pada data R yang merupakan kelas khusus.

P(R|S) : Nilai probabilitas pada hipotesis R yang berdasarkan kondisi *S.*

P(R) : Nilai probabilitas pada hipotesis *R.*

P(S) : Nilai probabilitas *S.*

Dengan menggunakan persamaan diatas, data yang telah diperoleh dapat diproses dengan algoritma *Naive Bayes* untuk penilaian data yang akan diklasifikasikan.

2.10 *Twitter*

Twitter adalah sebuah situs jejaring sosial yang sedang berkembang pesat saat ini karena pengguna dapat berinteraksi dengan pengguna lainnya dari komputer ataupun perangkat mobile mereka dari manapun dan kapanpun. Setelah diluncurkan pada Juli 2006, jumlah pengguna *twitter* meningkat sangat pesat. Pada Januari 2021, diperkirakan jumlah pengguna *twitter* yang terdaftar sekitar 187 juta pengguna aktif (Fitriana et al., 2021).

2.11 *Twitter Api*

Twitter API (Application Programming Interface) adalah fasilitas pada *twitter* yang dapat digunakan oleh pengembang perangkat lunak untuk membangun suatu aplikasi yang terintegrasi dengan *twitter*, fasilitas ini juga memungkinkan pengembang untuk mengambil data yang ada pada *twitter*. Untuk menjadi pengembang aplikasi *twitter*, pengembang harus mendaftar pada situs *https://dev.twitter.com*. Setelah melakukan pendaftaran pengembang akan mendapatkan *consumer key, consumer access, access token* dan *access token secret* yang dapat digunakan sebagai syarat otentifikasi.

Twitter API menggunakan *arsitektur REST (Representational State Transfer)* sehingga *twitter* API dapat digunakan dengan berbagai macam format seperti XML ataupun JSON.

API digunakan untuk menggabungkan dua sumber yang berbeda untuk membuat suatu program aplikasi yang saling terintegrasi, *Application programming interface* adalah suatu program/aplikasi yang disediakan oleh pihak *developer* agar pihak pengembang aplikasi lainnya dapat lebih mudah mengakses aplikasi tersebut, *API* berfungsi sebagai jembatan antara aplikasi satu dengan aplikasi yang lain.

Twitter API yaitu sebuah aplikasi yang diciptakan oleh pihak *twitter* agar mempermudah pihak *developer* lain untuk mengakses informasi web *twitter* tersebut dengan ketentuan dan syarat yang berlaku seperti yang terdapat pada *https://dev.twitter.com/oauth*. Ada dua jenis *twitter* API yaitu (Anas, 2016).

1. *Twitter rest API*

   Terdiri dari *twitter rest* dan *twitter search*. *Twitter rest* memberi *core* data dan *core twitter* objek. *Twitter search* berfungsi untuk melakukan pencarian mengenai suatu *instance* objek *twitter* maupun mencari *trend*.
2. *Twitter streaming API.*

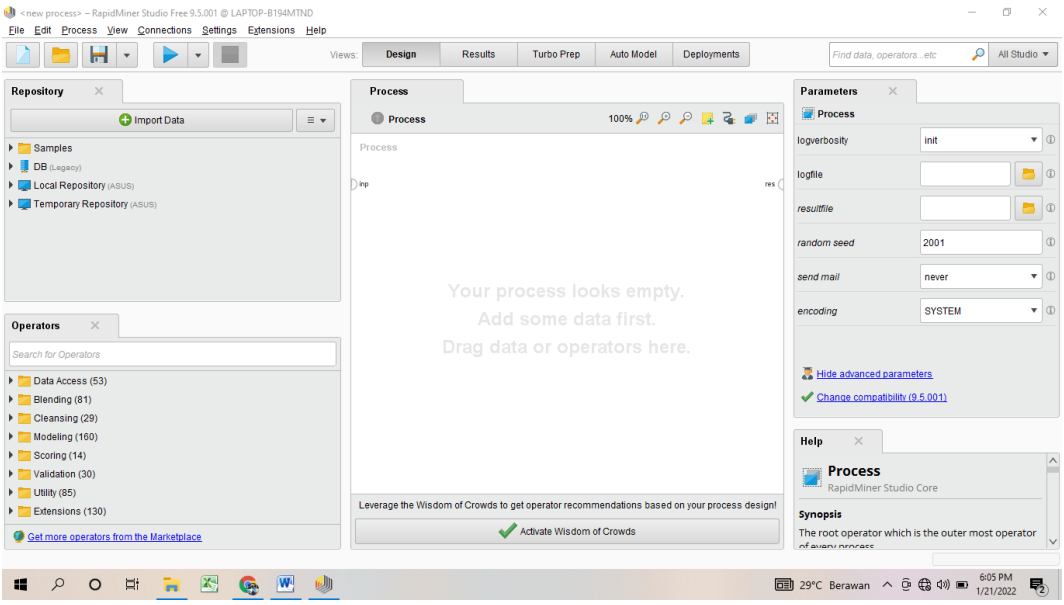
   API ini biasa digunakan untuk penggalian data karena melalui *API* ini informasi bkisa didapatkan secara *real time* dengan volume yang sangat tinggi.

2.12 *Rapid Miner*

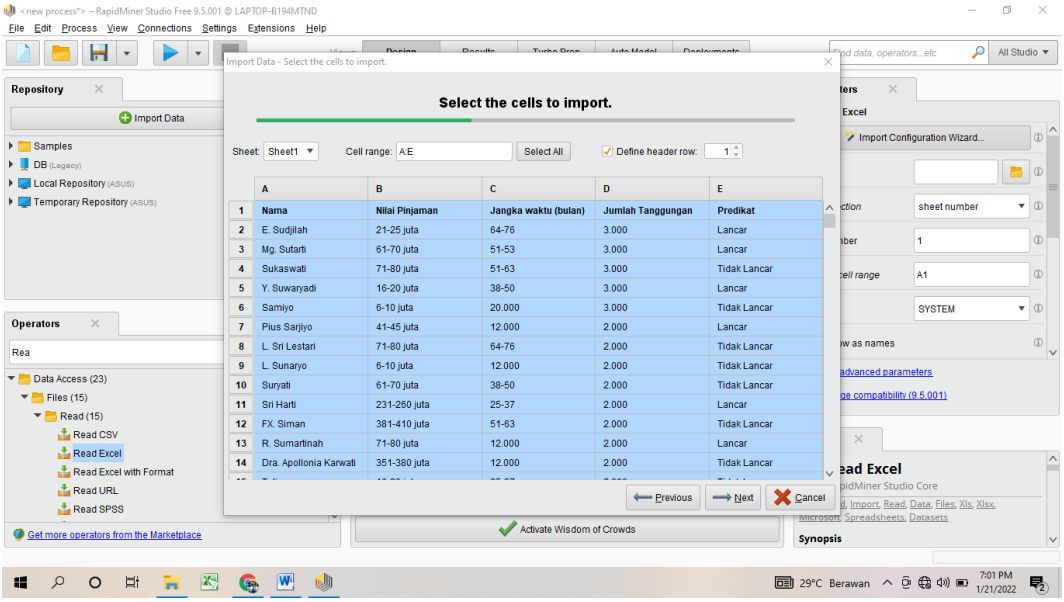
RapidMiner adalah *platform* perangkat lunak data ilmu pengetahuan yang dikembangkan oleh perusahaan dengan nama yang sama, yang menyediakan lingkungan terpadu untuk pembelajaran mesin (*machine learning*), pembelajaran mendalam (*deep learning*), penambangan teks (*text mining*), dan analisis prediktif (*predictive analytics*). Aplikasi ini digunakan untuk aplikasi bisnis dan komersial serta untuk penelitian, pendidikan, pelatihan, pembuatan *prototype* dengan cepat, dan pengembangan aplikasi serta mendukung semua langkah proses pembelajaran mesin termasuk persiapan data, visualisasi hasil, validasi dan pengoptimalan. *RapidMiner* dikembangkan dengan model *open core*. Berikut adalah pengenalan

dari perangkat lunak *rapid miner* yang dapat dilihat pada gambar dibawah ini (Nofitri & Irawati, 2019).

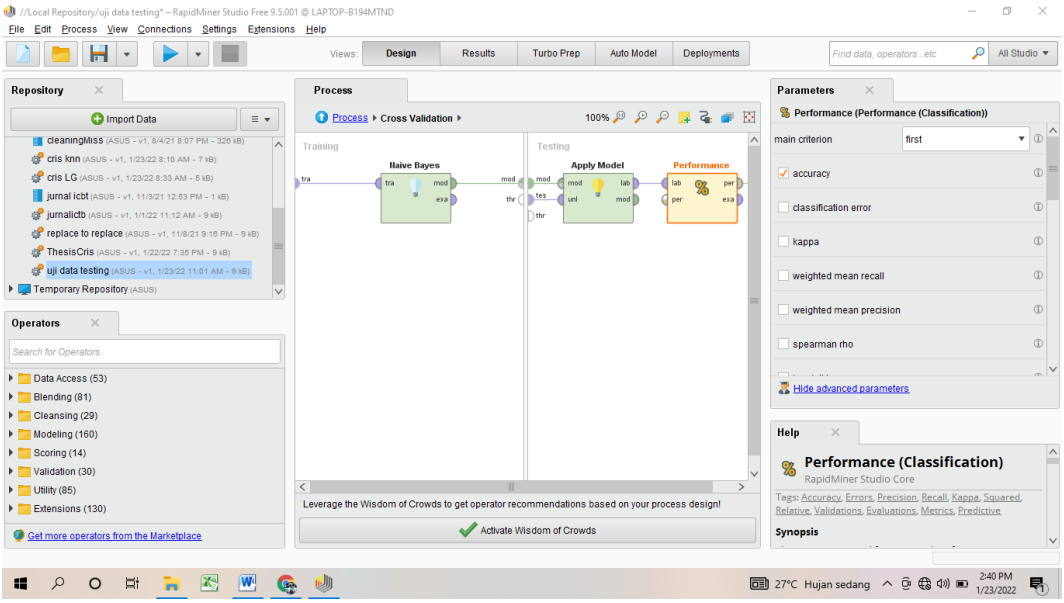
Gambar 2.3 Tampilan awal *rapid miner.*



Gambar 2.4 Halaman lembar kerja *rapid miner*



Gambar 2.5 Alur proses *import data*



Gambar 2.6 Pemodelan algoritma *Naïve Bayes*

2.13 *Crawling Data*

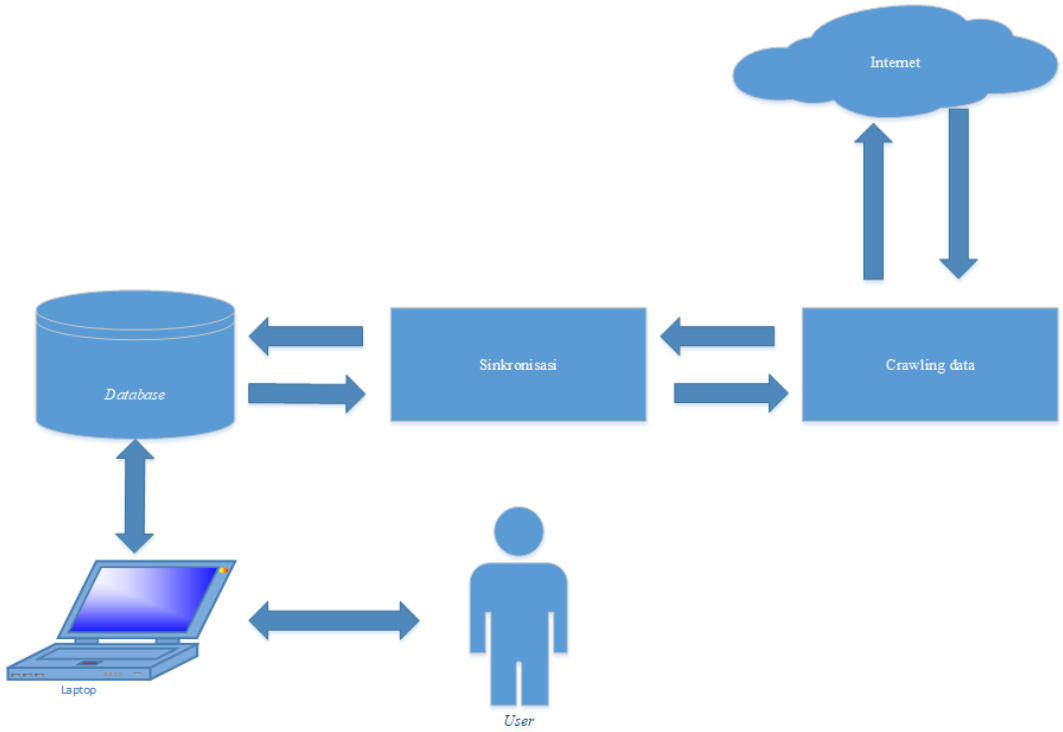
Crawling Data merupakan metode yang dilakukan menggunakan teknologi melalui *World Wide Web* (WWW). Biasanya bertujuan untuk mengambil data dengan kata kunci sesuai apa yang ingin di dapat. Artinya, crawling data secara otomatis

menemukan dan mengumpulkan sumber daya yang berbeda secara tertib dari internet sesuai dengan kebutuhan pengguna. Peneliti menerapkan metode *crawling* ini untuk mendapatkan informasi data sentimen mengenai komentar masyarakat tetang vasksin *booster*. Setelah nantinya file akan berbentuk dalam format .csv, kemudian dilanjutkan dengan pengolahan data.

Metode *crawling data* atau yang dikenal juga *web spider* atau *web robot* adalah suatu program yang yang dibangun dan dirancang dengan metode tertentu yang secara otomatis mengumpulkan semua data informasi yang diinginkan yang ada dalam bermacam sumber website. Sebuah program yang melintasi struktur *hypertext* dari web, dimulai dari sebuah alamat awal (*seed*) dan secara sekursif mengunjungi alamat web di dalam halaman web. Aplikasi *web crawling* mengambil informasi pada website yang diberikan kepadanya, kemudian menyerap dan menyimpan semua data informasi yang terkandung didalam website tersebut. Setiap kali aplikasi mengunjungi sebuah website, maka secara otomatis akan merekam semua link yang ada di halaman yang dikunjunginya itu untuk kemudian dikunjungi lagi satu persatu.

Selanjutnya aplikasi akan melakukan *data retrieval* dan menyimpannya ke dalam suatu media penyimpanan (*harddisk*) dengan kapasitas yang cukup besar. Data-data yang disimpan dalam *hardisk* ini kemudian, nantinya akan di ambil atau diakses pada saat dilakukan *query* (*mining data*). Dalam prosesnya *crawling data* yang berhasil dihimpun dapat mencapai milyaran sementara penyajiannya dapat dilakukan secara *real time*.

Aplikasi *crawling* menyajikan informasi dan memberikan tampilan antarmuka dengan berbagai bentuk seperti hubungan, mode pencarian berdasarkan sosial media, tampilan statistik, statistik bebasis *tag cloud*, yang dipadukan dengan informasi yang dinamis dan *relevan* melalui media sosial. Berikut adalah gambar alur *crawling* data yang dapat dilihat pada gambar 2.7 (Vyas & Uma, 2018).



Gambar 2.7 Alur *crawling data* (Suharno & Listiyoko, 2018)

2.14 *Confusion Matrix*

Confusion matrix merupakan suatu metode yang digunakan untuk melakukan perhitungan akurasi pada konsep *data mining*. *Confusion matrix* berisikan informasi mengenai hasil klasifikasi akutal dan telah di prediksi oleh sistem klasifikasi. Performa dari sistem tersebut biasanya dievaluasi menggunakan data dalam sebuah *matrix*. Dibawah ini menampilkan sebuah *confusion matrix* untuk pengklasifikasian ke dalam dua kelas (Rosandy, 2016).

Tabel 2.2 Tabel Visualisasi *Confusion Matrix*

Kelas	Klasifikasi Positif	Klasifikasi Negatif
Positif	TP (True Positif)	FN (False Negatif)
Negatif	FP (False Negatif)	TN (True Negatif)

Berikut adalah penjelasan dari tabel visualisasi *confusion matrix*.

a. TP (*true positive*), yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem.
b. TN (*true negative*), yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem.
c. FP (*false positive*), yaitu jumlah data positif namun terklasifikasi salah oleh sistem
d. FN (*false negative*)**,** yaitu jumlah data negatif namun terklasifikasi salah oleh sistem.

Dengan kata lain, nilai akurasi merupakan perbandingan antara data yang terklasifikasi benar dengan keseluruhan data. Nilai akurasi dapat diperoleh dengan persamaan berikut ini.

$$Akurasi = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \; X \; 100\%$$

Nilai presisi menggambarkan jumlah data kategori positif yang diklasifikasi secara benar dibagi dengan total data yang diklasifikasi positif, *presisi* dapat diperoleh dengan persamaan berikut ini.

$$Presisi = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \; X \; 100\%$$

Sementara itu, *recall* menunjukkan beberapa persen data kategori positif yang terklasifikasi dengan benar oleh sistem.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \; X \; 100\%$$