BAB III
METODOLOGI PENELITIAN

3.1. Alat dan Bahan Penelitian

3.1.1. Alat Penelitian

Kebutuhan yang digunakan dalam pelaksanaan penelitian ini *hardware* dan *software*. Masing-masing spesifikasi yang dibutuhkan dijelaskan sebagai berikut.

a. Spesifikasi Hardware terdiri dari PC / *Notebook* dengan spesifikasi *Processor* AMD Ryzen 5 5500U with Radeon Graphics 2.10 GHz, RAM 8 Gb, System type 64-bit operating system, x64-based processor, Harddisk 500 GB.
b. Spesifikasi Software, terdiri dari Microsoft Windows 11 Home Single Language; dan software RapidMiner Studio version 9.10 untuk melakukan *crawling* data dan juga melakukan analisis sentimen.

3.1.2. Bahan Penelitian

Bahan penelitian ini salah satunya adalah *dataset*. *Dataset* analisis sentimen berasal dari media sosial twitter. Penulis memilih Twitter karena kebijakan Twitter yang memberikan akses ke data mereka. Pihak Twitter menyediakan *Aplication Programming Interface* (API) untuk mengakses data yang dibutuhkan dengan terlebih dahulu mengajukan permintaan kepada pihak Twitter untuk memperoleh API Key nya. Twitter API adalah sekumpulan protokol, alat, dan

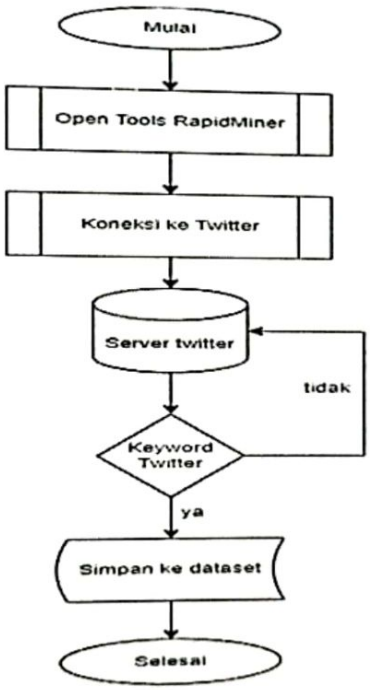
library yang dapat digunakan oleh pengembang aplikasi untuk mengakses dan memanfaatkan data yang tersedia di Twitter, seperti tweet, profil pengguna, dan lain-lain. Dengan menggunakan Twitter API, pengembang dapat membuat aplikasi baru atau menambahkan fitur baru pada aplikasi yang sudah ada, seperti aplikasi monitoring tweet, aplikasi analisis, atau aplikasi pembuatan infografik, dan lain-lain.

Selain itu, pemilihan media sosial Twitter untuk *data mining* karena tiga point utama [2] yaitu: a) Twitter API memiliki desain yang bagus dan mudah diakses, b) Data Twitter tersedia dalam format yang nyaman untuk dianalisis, dan c) Kebijakan Twitter untuk data relatif liberal dibandingkan dengan API yang lain. Pertimbangan lain penulis mengambil data dari Twitter karena beberapa trend dibidang data analitik saat ini banyak menggunakan Twitter untuk melakukan sentimen analitis menilai komentar pengguna itu bernada positif atau negatif.

Untuk memperoleh data, penulis menggunakan *tools* RapidMiner yang merupakan perangkat lunak terbuka untuk menganalisis *data mining, text mining*, dan analisis prediktif. RapidMiner menggunakan berbagai teknik deskriptif dan prediktif untuk memberikan wawasan kepada pengguna dan membantu pengguna membuat keputusan terbaik. RapidMiner memiliki berbagai operator *data mining*, termasuk operator *input, output, preprocessing* data, dan visualisasi yang dapat digunakan untuk melakukan operasi *data mining* dengan mudah dan cepat. Semua proses dalam penelitian ini dari proses awal *crawling* data hingga visualisasi dilakukan dengan RapidMiner.

Dengan bantuan *tools* RapidMiner, data dikumpulkan menggunakan metode *crawling* data dengan Twitter API yang disimpan dalam format csv. Dalam melakukan *crawling data* dibutuhkan *credential keys* berupa token dan key untuk mengakses Twitter API yang diperoleh dengan langkah-langkah yang telah ditetapkan oleh Twitter. Secara manual, data yang terkumpul dilakukan *labelling* berdasarkan sentimen yang mendukung dan tidak mendukung terkait diterbitkannya PERPPU tentang Cipta Kerja.

Adapun alur pengumpulan *dataset* dari Twitter dengan RapidMiner sebagaimana gambar 3.1. berikut.

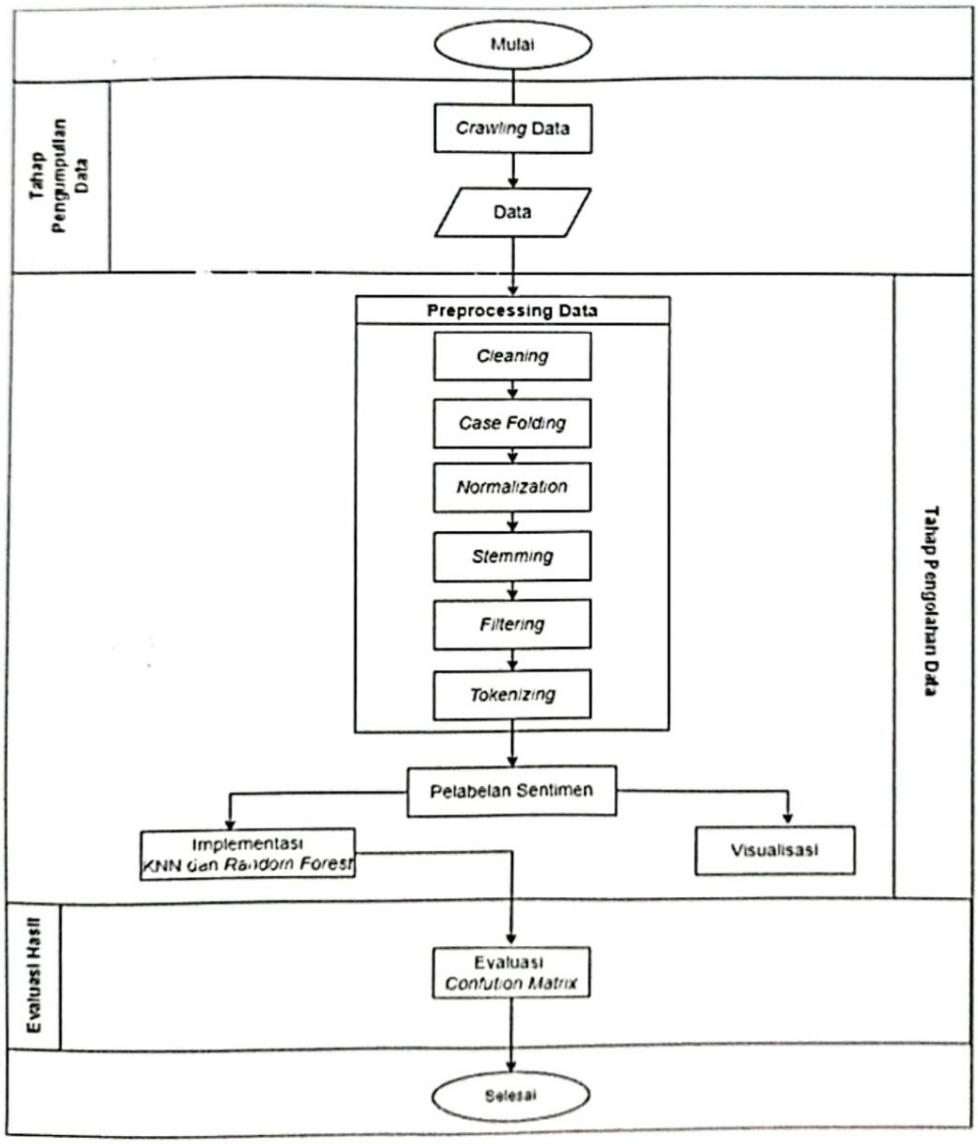


Gambar 3. 1. *Flowchart* Pengumpulan *Dataset* dengan RapidMiner

3.2. Tahapan Penelitian

Penelitian ini melalui beberapa tahapan yang harus dilakukan, yaitu dari awal mengumpulkan data atau penarikan data (*crawling*) dari Twitter hingga

visualisasi. Secara umum tahapan terdiri dari tahap pengumpulan data, tahap pengolahan data, dan tahap evaluasi. Tahap pengumpulan data dilakukan *crawling dataset* twitter menggunakan RapidMiner. Tahap pengolahan data meliputi tahap *preprocesing* data, pelabelan sentimen dan pembobotan, klasifikasi menggunakan algoritma, dan tes pengujian. Sedangkan tahap evaluasi hasil dilakukan evaluasi menggunakan *confution matrix*. Adapun tahapan penelitian ini sebagaimana dalam gambar 3.2. berikut ini.

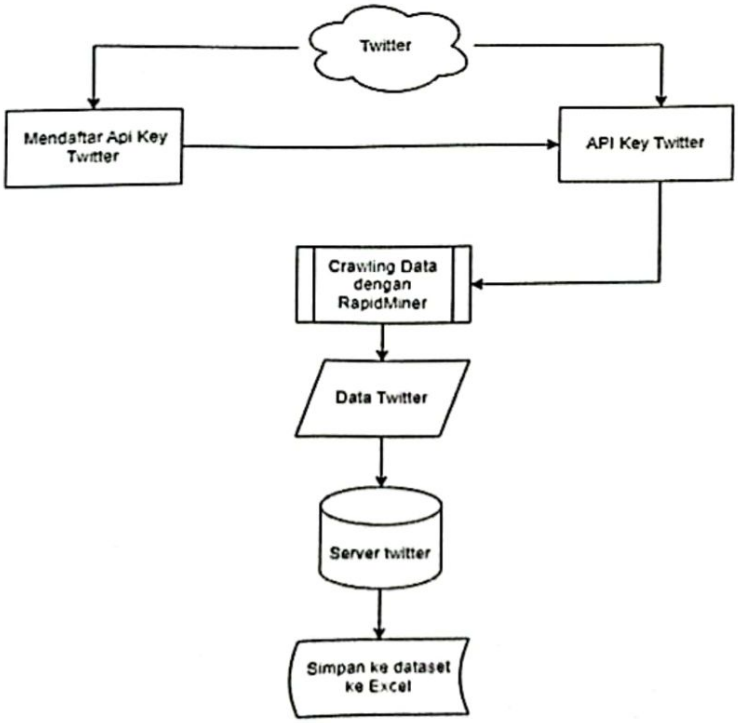


Gambar 3. 2. Block Tahap Penelitian

Teknik analisis data yang dilakukan dalam penelitian ini menggunakan model yang disusun sebagai proses pelabelan data sentimen tweet, dan KNN serta *Random Forest* sebagai proses pemodelan data uji tweet pada analisis sentimen pengguna media sosial twitter terhadap PERPPU tentang Cipta Kerja menggunakan RapdiMiner. Berdasarkan gambar alur tahapan penelitian tersebut, dijabarkan lebih lanjut sebagai berikut.

3.2.1. Tahap pengumpulan data

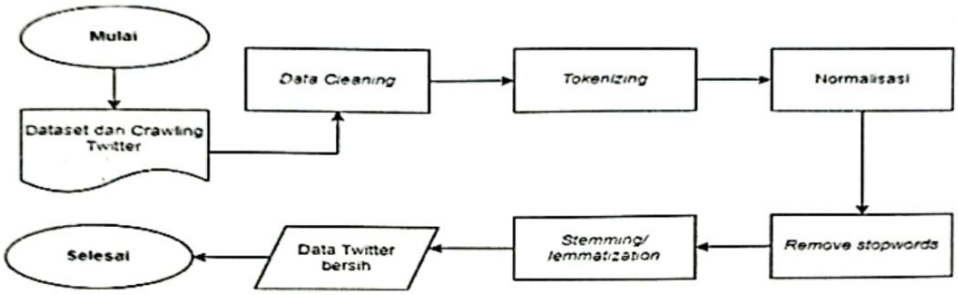
Pada penelitian ini, data yang digunakan adalah komentar pengguna media sosial Twitter dengan kata kunci "PERPPU Cipta Kerja" dengan filter bahasa Indonesia. Data yang diperoleh yaitu tweet dari pengguna Twitter. Data dikumpulkan (*crawling*) secara langsung terhubung dari media sosial Twitter menggunakan RapidMiner, yang kemudian disimpan dalam bentuk *excel*. Tahapannya sebagaimana gambar 3.3. berikut.



Gambar 3. 3. Tahapan *Crawling*

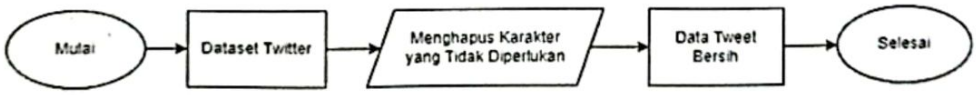
3.2.2. Tahap Pengolahan Data

Pada klasifikasi berita yang menggunakan tipe data berupa teks terdapat beberapa macam proses yang dilakukan, diantaranya *case folding*, menghilangkan tanda baca, *tokenization, lemmatization,* dan *stop word removal* [42]. Langkah-langkah dalam tahap proses *preprocessing* meliputi: 1) *Data cleaning,* 2) *tokenizing*, 3) *normalisasi,* 4) *remove stopwords,* dan 5) *stemming/lemmatization.* Tahapannya pada gambar 3.4. sebagai berikut.



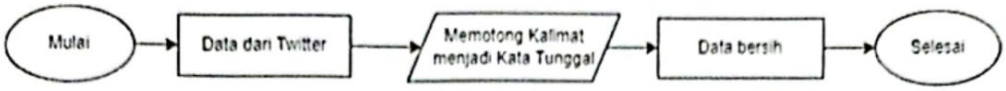
Gambar 3. 4. Tahapan *Preprocessing*

Pada tahap *data cleaning*, dilakukan pembersihan data dan menghilangkan karakter yang tidak diperlukan pada data tweet [9] seperti tanda baca, numeric, url, username, mention, hashtag dan retweet seperti („”~&?!><#%{}([0-9]+:;“). Adapun langkah-langkahnya sebagaimana gambar 3.5. berikut.



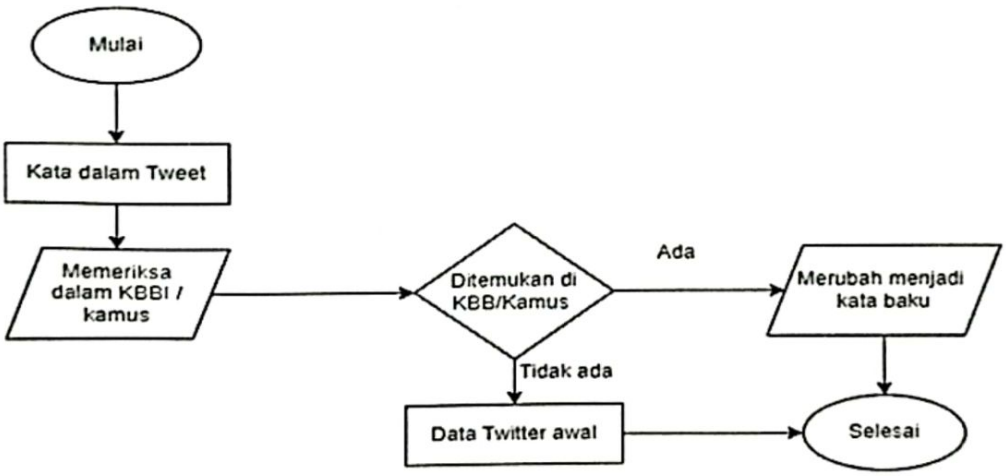
Gambar 3. 5. Tahapan Proses *Data Cleaning*

Tahap *tokenizing* dilakukan untuk memisahkan *string* atau memecahkan kalimat jadi kata per kata agar mendapatkan kata yang memiliki nilai. Langkah-langkahnya sebagaimana gambar 3.6. berikut.



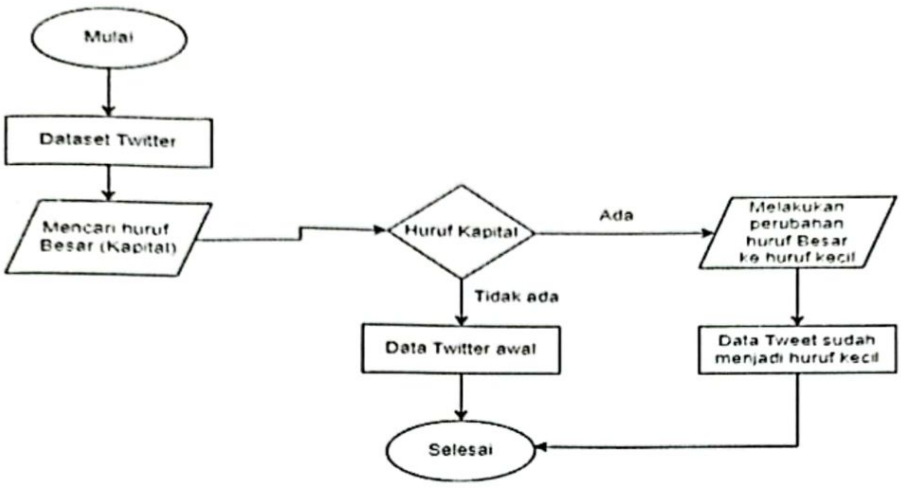
Gambar 3. 6. Tahapan Proses *Tokenizing*

Tahap Normalisasi / *case folding* bertujuan untuk mengubah bentuk teks yang tidak standar menjadi bentuk yang lebih mudah diproses dan dipahami oleh mesin. Normalisasi termasuk proses memperbaiki singkatan, memperbaiki tata bahasa, memperbaiki ejaan, dan memperbaiki format teks. Dalam hal normalisasi singkatan, teknik ini memperbaiki singkatan atau akronim yang sering digunakan dalam bahasa Indonesia seperti "sdg" menjadi "sedang", "utk" menjadi "untuk", "tdk" menjadi "tidak", dan lain-lain. Tahapannya sebagaimana gambar 3.7. berikut.



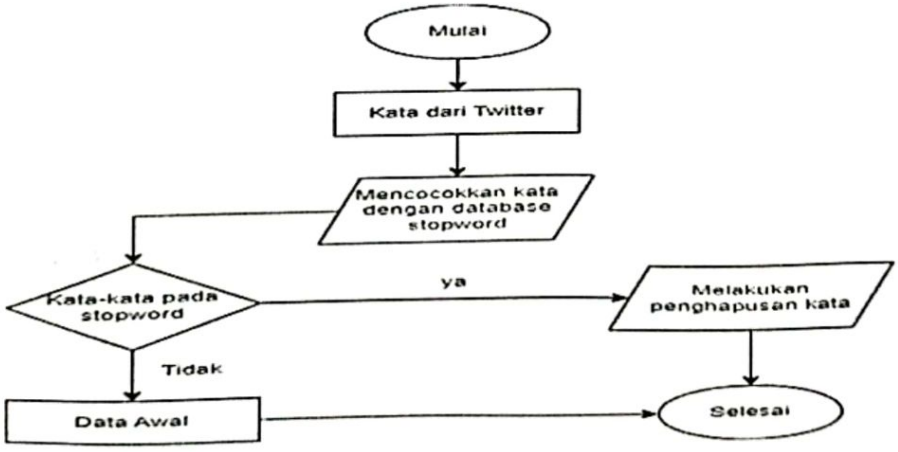
Gambar 3. 7. Tahapan Proses Normalisasi Data

Pada *case folding*, dilakukan konversi atau perubahan huruf kapital ke dalam huruf kecil (*lowercase*) pada semua data yang terdapat di dalam dokumen. Langkah-langkahnya sebagaimana gambar 3.8. berikut.



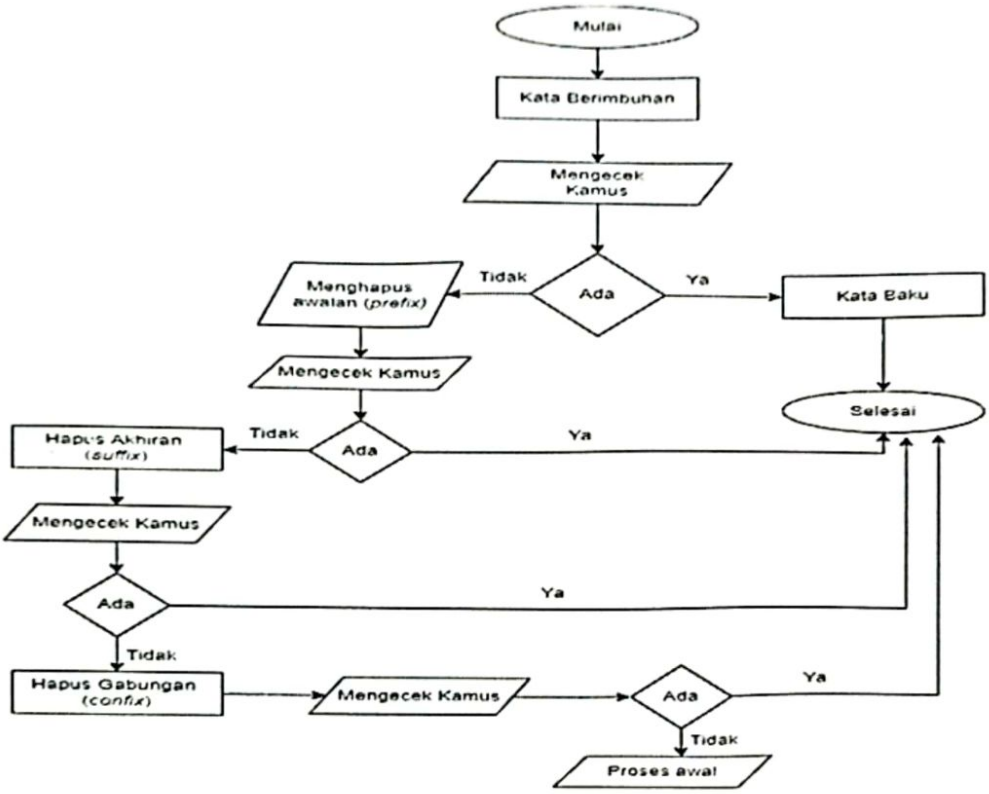
Gambar 3. 8. Tahap Proses *Case Folding*

Tahap selanjutnya yaitu *Filtering.* Dalam tahap ini, kata-kata yang tidak bermakna atau tidak penting dibuang dari teks. *Stopword* adalah salah satu teknik yang digunakan dalam tahap *filtering* ini. *Stopword* adalah daftar kata-kata yang biasa digunakan tetapi tidak jelas dan dapat dibuang dari teks, seperti "dari", "yang", "untuk", "dan", "di", dan lain-lain. Dalam penelitian ini, *stopword* diambil dari sumber KBBI atau sumber lain yang memiliki daftar kata-kata yang sering digunakan. Adapun tahapannya sebagaimana gambar 3.9. berikut.



Gambar 3. 9. Tahapan Proses *Filtering*

Tahapan *stemming* memiliki fungsi untuk menghapus seluruh kata imbuhan yang terdapat pada data tweet seperti *prefix*, *suffix* dan *konfix*. Dengan kata lain, tahap ini merupakan proses menghilangkan atribut tambahan dari kata seperti menghapus "mem" dan "-kan" dan sebagainya. Adapun tahapannya sebagaimana gambar 3.10. berikut.



Gambar 3. 10. Tahapan Proses *Stemming*

Setelah dilakukan tahap *prepocessing data*, tahap selanjutnya dilakukan analisis sentimen untuk mengetahui nilai sentimen dari data yang berupa negatif, positif atau netral. Langkah-langkahnya sebagai berikut [2].

a. Tahap 1: mencari sentimen negatif, positif, dan netral di setiap baris.
b. Tahap 2: menganalisis sentimen seluruh dokumen sebagai negatif, positif, serta netral.

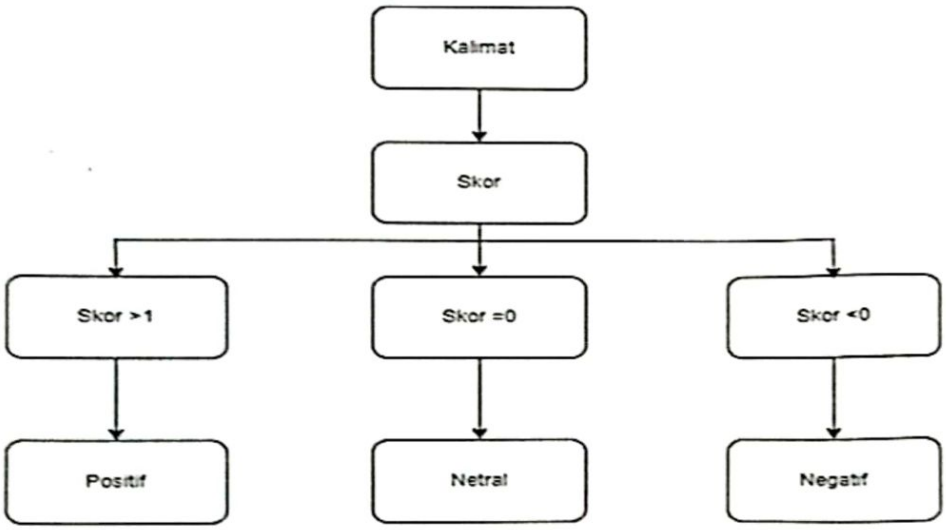
c. Tahap 3: menerapkan pengelompokan untuk mengumpulkan semua atribut yang ada dengan skor sentimen yang sama.

d. Tahap 4: menggunakan visualisasi data dari analisis sentimen untuk interaksi antar pengguna.

Rumus perhitungan skor sentimennya sebagaimana persamaan 10 berikut.

$$Skor = (\sum kata\ positif - \sum kata\ negatif) \ldots\ldots\ldots\ldots\ldots\ldots\ldots (10)$$

Hasil perhitungan skor: Jika skor kalimat >0, maka diklasifikasikan ke dalam kelas positif, jika skor kalimat <0, maka diklasifikasikan ke dalam kelas negatif, dan jika skor kalimat =0, diklasifikasikan dalam kelas netral. Lebih jelasnya digambarkan sebagaimana gambar 3.11 berikut.



Gambar 3. 11. Alur Pelabelan Sentimen

Memberikan label pada data dalam penelitian ini dilakukan dengan mengkategorikan sentimen atau pendapat dalam teks menjadi beberapa kategori atau label yang sesuai yaitu positif, negatif, atau netral. Pada pelabelan ini peneliti untuk harus memahami domain atau topik dari data yang akan dilabeli, yaitu tentang Peraturan Pemerintah Pengganti Undang-Undang (Perppu) Tentang

Cipta Kerja. Kemudian menentukan kategori atau label sentimen yang akan digunakan yang dapat berupa "Positif," "Negatif," dan "Netral."

Setiap teks tweet dilakukan anotasi secara manual, yaitu teks dibaca dan ditentukan kategori sentimen yang paling sesuai dengan teks tersebut. Jika sebuah tweet berisi dukungan terhadap peraturan tersebut, peneliti memberi label "Positif." Namun jika tweet tersebut berisi kritik atau protes, peneliti memberinya label "Negatif", dan jika tidak ada sentimen yang jelas atau teksnya tidak berhubungan dengan topik, peneliti memberinya label "Netral. Hal ini harus dilakukan secara konsisten tidak peduli siapa yang memberi label tersebut.

Setelah data diberi label, peneliti memisahkan data menjadi dua bagian yaitu menjadi data pelatihan (*training* data) dan data pengujian (*testing* data). Data pelatihan digunakan untuk melatih model sentimen, sedangkan data pengujian digunakan untuk menguji performa model. Data pengujian yang telah dilabeli untuk menguji performa model sentimen analisis yang dikembangkan. Peneliti menggunakan metrik evaluasi seperti akurasi, *precision, recall,* dan *F1-score* untuk menilai sejauh mana model dapat memprediksi sentimen dengan akurat. Jika langkah-langkah tersebut belum berhasil baik, maka dilakukan iterasi dengan mengulang langkah-langkah ini untuk memperbaiki model yang dibuat dan membuat label-label lebih akurat.

Data yang sudah dilakukan pelabelan dan pembobotan selanjutnya menggunakan *wordcloud* pada RapidMiner dilakukan visualisasi hasil dari analisis sentimen sehingga terlihat gambaran frekuensi kata-kata yang sering muncul. Setelah pelabelan data, kemudian dilakukan optimasi menggunakan

PSO. PSO dapat digunakan untuk mencari parameter optimal dalam algoritma klasifikasi yang dapat meningkatkan performa model. Setelah analisis sentimen, langkah selanjutnya adalah memproses dan mengklasifikasikan lebih lanjut data yang diberi label. Metode yang digunakan untuk melakukan klasifikasi yaitu menggunakan metode KNN dan *Random Forest*, dengan sebelumnya dioptimasi menggunakan PSO dengan bantuan *tools* RapidMiner.

3.2.3. Tahap Evaluasi Hasil

Metrik yang dapat digunakan untuk mengevaluasi kinerja model pada tes pengujian, yaitu: Akurasi (*Accuracy*), Presisi (*Precision*), *Recall* (*Recall*), dan F1-Score. Akurasi mengukur persentase data yang diklasifikasikan dengan benar oleh model, presisi mengukur persentase data positif yang diklasifikasikan dengan benar oleh model dari semua data yang diklasifikasikan sebagai positif. *Recall* mengukur persentase data positif yang diklasifikasikan dengan benar oleh model dari semua data positif yang sebenarnya, dan F1-score memberikan gambaran kinerja model yang lebih baik.

3.3. Metode Yang Digunakan

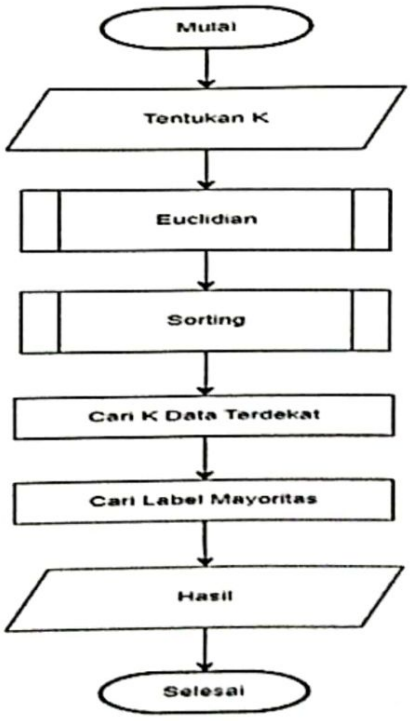
Metode atau model yang diusulkan dalam penelitian ini yaitu algoritma KNN dan *Random Forest* dengan masing-masing menggunakan seleksi fitur PSO. PSO digunakan untuk menghasilkan akurasi yang lebih tinggi. KNN merupakan sebuah metode pembelajaran mesin yang digunakan untuk melakukan klasifikasi dan regresi. KNN berdasarkan pada ide bahwa suatu *instance* akan

diklasifikasikan ke kelas yang paling banyak dipresentasikan oleh K tetangga terdekatnya. Untuk melakukan klasifikasi menggunakan KNN, sistem akan memperhitungkan jarak dari setiap *instance* baru ke setiap *instance* pada dataset yang ada. Kemudian, sistem akan mengidentifikasi K tetangga terdekat yang memiliki kelas yang berbeda-beda. Setelah itu, sistem akan menentukan kelas dari *instance* baru berdasarkan mayoritas kelas dari K tetangga tersebut.

KNN memiliki konsep menggolongkan dokumen uji X sebagai berikut [2].

a. Menetapkan parameter K yang merupakan jumlah dari tetangga terdekat.
b. Mengkalkulasi kuadrat jarak *euclidean* dari *data training*.
c. Mengurutkan data secara *ascending*.
d. Menggabungkan kategori Y yang berlandaskan tetapan parameter K.

Tahapan perhitungan algoritma KNN sebagaimana gambar 3.12. berikut.



Gambar 3. 12. Alur Proses Algoritma KNN

Random Forest merupakan salah satu algoritma klasifikasi yang menggabungkan beberapa *decision tree* menjadi satu model dengan *training* pada data yang dimiliki. *Random Forest* bekerja dengan cara membuat beberapa decision tree yang masing-masing mempelajari sebagian dari dataset. Setiap *tree* dalam model *Random Forest* membuat prediksi independen dan hasil akhir ditentukan melalui voting dari semua *tree*. Pada *Random Forest* dilakukan proses pengklasifikasian dengan memecah data secara acak ke dalam *decision tree* [2]. Algoritma *Random Forest* digunakan untuk memperoleh hasil akhir dari sistem identifikasi. Dalam hal ini, sebuah prediksi akan diperoleh dengan mengamati voting mayoritas dari setiap kelas yang telah dilakukan pelabelan data sebelumnya.

KNN dan *Random Forest* merupakan metode yang mampu bekerja dengan baik pada set data dengan dimensi tinggi. Untuk mengantisipasi masalah yang mungkin timbul misalnya pada penentuan parameter, penulis melakukan optimasi terhadap kedua metode yang digunakan dengan PSO. Dalam tahapan ini peneliti membutuhkan *dataset* berupa cuitan yang sudah dikumpulkan untuk membuat logika *model machine learning* terkait klasifikasi sentimen terhadap PERPPU Tentang Cipta Kerja. Untuk menghindari data menjadi tidak relevan, data yang sudah dikumpulkan akan dilakukan *preprocessing* sehingga hasil prediksi dapat dipertanggungjawabkan.

Pada tahap ini teks atau sentimen diklasifikasikan dengan tahapan *preprocessing* agar teks yang memiliki isi yang tidak sempurna seperti data yang hilang, tidak valid, salah ketik, atau juga atribut-atribut data yang tidak relevan.

Keberadaan data tersebut dapat mengurangi mutu atau akurasi sehingga perlu dibuang. Pada teks yang belum diolah, biasanya memiliki karakteristik dimensi yang tinggi, terdapat *noise* pada data dan terdapat struktur teks yang tidak baik.

3.4. Evaluasi Metode

Skenario yang diterapkan untuk melakukan evaluasi dan validasi agar memastikan validitas metode dan hasil penelitian, diusulkan model dengan menerapkan algoritma KNN berbasis PSO dan *Random Forest* berbasis PSO. Algoritma KNN dan *Random Forest* diyakini dapat menghasilkan akurasi yang baik, namun dalam penelitian ini penulis juga melakukan peningkatan akurasi dan hasil perhitungan menggunakan PSO.

Standar validasi yang digunakan dalam penelitian ini yaitu 10 *fold cross-validation* yaitu teknik validasi model dengan membagi dataset menjadi 10 bagian yang berukuran sama, dengan 9 bagian digunakan sebagai data training dan 1 bagian digunakan sebagai data testing. Proses ini diulang sebanyak 10 kali, dengan setiap bagian data testing digunakan sekali sebagai data testing dan selebihnya sebagai data training.

Dengan menggunakan *10-fold Cross Validation*, setiap data dalam dataset akan digunakan sebagai data testing sekali, sehingga keseluruhan dataset dapat digunakan untuk menguji model. Cara ini akan memberikan estimasi yang lebih baik mengenai performa model dalam mengatasi masalah generalisasi, di mana model harus dapat melakukan prediksi yang akurat pada data baru yang belum pernah dilihat sebelumnya. *10-fold Cross Validation* ini sangat berguna karena

membantu menghindari *overfitting*, mempermudah interpretasi hasil, dan membantu memilih model yang paling baik untuk digunakan. *Performance* diukur menggunakan *Accuracy* dan AUC serta akan ditampilkan dalam bentuk kurva ROC.

Contoh penerapan metode yang digunakan dalam penelitian ini sebagai berikut. Disajikan contoh dataset sebagaimana tabel 3.1. berikut.

Tabel 3. 1. Data latih

No	Komentator	Komentar	Kelas
1	@305_NiBOSS	Dengan PERPPU Cipta Kerja, Indonesia hanya akan terus mengundang investasi yang mengeksploitasi sumber daya alam dan tenaga asing	Negatif
2	@yangmi3prnew	Membaca aturan dalam PERPPU Cipta Kerja di sini sangat melindungi hak-hak Karyawan, kenapa banyak penolakan	Positif
3	@ileng_s	PERPPU Cipta Kerja Konfirmasi pembangkangan terhadap konstitusi	Negatif
4	@RadioElshinta	PERPPU cipta kerja disesuaikan dgn putusan MK waktu judicial review dan segera di ajukan ke DPR RI agar segera disetujui dlm persidangan berikutnya agar tidak jadi polemik lagi	Positif
5	@ AdhiZulkarnaen	DPR slama 2 periode ini kan bkn Mewakili Rakyat, Bung RH, tp Mewakili Oligarki (Dewan Pemeras Rakyat).	Negatif

Tahapan *Preprosesing*

Data cleaning. Dilakukan pembersihan data dan menghilangkan karakter yang tidak diperlukan pada data tweet seperti tanda baca, numeric, dan sebagainya. Hasil data *cleaning* disajikan dalam tabel 3.2. berikut.

Tabel 3. 2. Hasil Data *Cleaning*

No	Komentator	Komentar	Hasil
1	05_niboss	Dengan PERPPU Cipta Kerja, Indonesia hanya akan terus	Dengan PERPPU Cipta Kerja Indonesia hanya akan terus

No	Komentator	Komentar	Hasil
		mengundang investasi yang mengeksploitasi sumber daya alam dan tenaga asing	mengundang investasi yang mengeksploitasi sumber daya alam dan tenaga asing
2	yangmi3prnew	Membaca aturan dalam PERPPU Cipta Kerja di sini sangat melindungi hak-hak Karyawan, kenapa banyak penolakan	Membaca aturan dalam PERPPU Cipta Kerja di sini sangat melindungi hak-hak Karyawan kenapa banyak penolakan
3	ileng_s	PERPPU Cipta Kerja Konfirmasi pembangkangan terhadap konstitusi	PERPPU Cipta Kerja Konfirmasi pembangkangan terhadap konstitusi
4	radioelshinta	PERPPU cipta kerja disesuaikan dgn putusan MK waktu judicial review dan segera di ajukan ke DPR RI agar segera disetujui dlm persidangan berikutnya agar tidak jadi polemik lagi	PERPPU cipta kerja disesuaikan dgn putusan MK waktu judicial review dan segera di ajukan ke DPR RI agar segera disetujui dlm persidangan berikutnya agar tidak jadi polemik lagi
5	adhizulkarnaen	DPR slama 2 periode ini kan bkn Mewakili Rakyat, Bung RH, tp Mewakili Oligarki (Dewan Pemeras Rakyat).	DPR slama periode ini kan bkn Mewakili Rakyat Bung RH tp Mewakili Oligarki Dewan Pemeras Rakyat

Case Folding. Bertujuan untuk mengubah semua huruf dalam suatu string menjadi huruf kecil atau huruf besar. Perbedaan antara huruf besar dan huruf kecil, dihilangkan sehingga memudahkan proses selanjutnya. Pada proses ini juga dilakukan penghapusan karakter-karakter pada dokumen yang tidak dibutuhkan, seperti misalnya emot ikon, tanda baca dan lain sebagainya. Hasil *case folding* sebagaimana tabel 3.3. berikut ini.

Tabel 3. 3. Hasil *Case Folding*

No	Komentator	Komentar	Hasil
1	05_niboss	Dengan PERPPU Cipta Kerja Indonesia hanya akan terus mengundang investasi yang mengeksploitasi sumber daya alam dan tenaga asing	dengan perppu cipta kerja indonesia hanya akan terus mengundang investasi yang mengeksploitasi sumber daya alam dan tenaga asing
2	yangmi3prnew	Membaca aturan dalam PERPPU Cipta Kerja di sini	membaca aturan dalam perppu cipta kerja di sini sangat

No	Komentator	Komentar	Hasil
		sangat melindungi hak-hak Karyawan kenapa banyak penolakan	melindungi hak-hak karyawan kenapa banyak penolakan
3	ileng_s	PERPPU Cipta Kerja Konfirmasi pembangkangan terhadap konstitusi	perppu cipta kerja konfirmasi pembangkangan terhadap konstitusi
4	radioelshinta	PERPPU cipta kerja disesuaikan dgn putusan MK waktu judicial review dan segera di ajukan ke DPR RI agar segera disetujui dlm persidangan berikutnya agar tidak jadi polemik lagi	perppu cipta kerja disesuaikan dengan putusan mk waktu judicial review dan segera di ajukan ke dpr ri agar segera disetujui dlm persidangan berikutnya agar tidak jadi polemik lagi
5	adhizulkarnaen	DPR slama periode ini kan bkn Mewakili Rakyat Bung RH tp Mewakili Oligarki Dewan Pemeras Rakyat	dpr slama periode ini kan bkn mewakili rakyat bung rh, tp mewakili oligarki dewan pemeras rakyat

Tahap normalisasi dilakukan untuk mengubah dan memperbaiki kata yang disingkat ke dalam kata yang memiliki arti sama berdasarkan KBBI agar menjadi informasi yang dapat diproses dengan mudah misalnya "sdg" menjadi "sedang", utk" menjadi "untuk", "tdk" menjadi "tidak" dan sebagainya. Hasil dari hasil normalisasi sebagaimana tabel 3.4. berikut ini.

Tabel 3. 4. Data Hasil Normalisasi

No	Komentator	Komentar	Hasil
1	05_niboss	dengan perppu cipta kerja indonesia hanya akan terus mengundang investasi yang mengeksploitasi sumber daya alam dan tenaga asing	dengan perppu cipta kerja indonesia hanya akan terus mengundang investasi yang mengeksploitasi sumber daya alam dan tenaga asing
2	yangmi3prnew	membaca aturan dalam perppu cipta kerja di sini sangat melindungi hak-hak karyawan, kenapa banyak penolakan	membaca aturan dalam perppu cipta kerja di sini sangat melindungi hak-hak karyawan kenapa banyak penolakan
3	ileng_s	perppu cipta kerja konfirmasi pembangkangan terhadap konstitusi	perppu cipta kerja konfirmasi pembangkangan terhadap konstitusi

No	Komentator	Komentar	Hasil
4	radioelshinta	perppu cipta kerja disesuaikan dgn putusan mk waktu judicial review dan segera di ajukan ke dpr ri agar segera disetujui dlm persidangan berikutnya agar tidak jadi polemik lagi	perppu cipta kerja disesuaikan dengan putusan mahkamah konstitusi waktu judicial review dan segera di ajukan ke dewan perwakilan rakyat republik indonesia agar segera disetujui dalam persidangan berikutnya agar tidak jadi polemik lagi
5	adhizulkarnaen	dpr slama 2 periode ini kan bkn mewakili rakyat bung rh tp mewakili oligarki dewan pemeras rakyat	dpr selama periode ini mewakili rakyat bung tetapi mewakili oligarki dewan pemeras rakyat

Stopword Removal dimana pada tahap ini akan dihilangkan kata-kata yang tidak penting seperti kata "di", "dan", "karena" "oleh" dan lain sebagainya. Tahap ini dilakukan agar dapat memperbesar nilai akurasi. Hasil perhitungan dari *stopword removal* sebagaimana tabel 3.5. berikut.

Tabel 3. 5. Data Hasil Stopword Removal

No	Komentar	Komentar	Hasil
1	05_niboss	dengan perppu cipta kerja indonesia hanya akan terus mengundang investasi yang mengeksploitasi sumber daya alam dan tenaga asing	perppu cipta kerja indonesia terus undang investasi eksploitasi sumber daya alam tenaga asing
2	yangmi3prnew	membaca aturan dalam perppu cipta kerja di sini sangat melindungi hak-hak karyawan kenapa banyak penolakan	membaca aturan perppu cipta kerja sini sangat lindungi hak-hak karyawan kenapa banyak penolakan
3	ileng_s	perppu cipta kerja konfirmasi pembangkangan terhadap konstitusi	perppu cipta kerja konfirmasi pembangkang terhadap konstitusi
4	radioelshinta	perppu cipta kerja disesuaikan dengan putusan mahkamah konstitusi waktu judicial review dan segera di ajukan ke dewan perwakilan rakyat republik indonesia agar segera disetujui dalam persidangan berikutnya agar tidak jadi polemik lagi	perppu cipta kerja sesuai putusan mahkamah konstitusi waktu judicial review segera ajukan dewan perwakilan rakyat republik indonesia agar segera setujui sidang berikutnya agar tidak jadi polemik

No	Komentar	Komentar	Hasil
5	adhizulkarnaen	dpr selama periode ini bukan mewakili rakyat bung rh, tetapi mewakili oligarki dewan pemeras rakyat	dewan perwakilan rakyat selama periode ini mewakili rakyat bung mewakili oligarki dewan pemeras rakyat

Stemming. Pada proses ini dilakukan pengubahan kata yang ada dalam dokumen menjadi bentuk kata dasar. Contohnya pengubahan kata "himbauan" menjadi "imbau", "larangan" menjadi "larang", "penyebaran" menjadi "sebar" dan lain sebagainya. Hasil *stemming* sebagaimana tabel 3.6 berikut.

Tabel 3. 6. Data Hasil *Stemming*

No	Komentar	Komentar	Hasil
1	05_niboss	perppu cipta kerja indonesia hanya akan terus mengundang investasi yang mengeksploitasi sumber daya alam tenaga asing	perppu cipta kerja indonesia hanya terus undang investasi eksploitasi sumber daya alam tenaga asing
2	yangmi3prnew	membaca aturan perppu cipta kerja sini sangat melindungi hak-hak karyawan kenapa banyak penolakan	membaca aturan perppu cipta kerja sini sangat lindungi hak-hak karyawan kenapa banyak penolak
3	ileng_s	perppu cipta kerja konfirmasi pembangkangan terhadap konstitusi	perppu cipta kerja konfirmasi pembangkang terhadap konstitusi
4	radioelshinta	perppu cipta kerja disesuaikan putusan mahkamah konstitusi waktu judicial review segera ajukan dewan perwakilan rakyat republik indonesia agar segera disetujui persidangan berikutnya agar polemik	perppu cipta kerja sesuai putusan mahkamah konstitusi waktu judicial review segera ajukan dewan perwakilan rakyat republik indonesia agar segera setujui sidang berikutnya agar polemik
5	adhizulkarnaen	dewan perwakilan rakyat selama periode ini bukan mewakili rakyat bung mewakili oligarki dewan pemeras rakyat	dewan perwakilan rakyat selama periode ini wakili rakyat bung wakili oligarki dewan peras rakyat

Kemudian dihitung TF-IDF dengan contoh perhitungan TF-IDF untuk kata "lapangan kerja" dalam sebuah dokumen dan korpus teks terkait PERPPU Cipta Kerja:

1. Dokumen data Twitter: "PERPPU Cipta Kerja diharapkan dapat membuka lapangan kerja baru di Indonesia"
2. Korpus teks: "RUU Cipta Kerja bertujuan untuk mempercepat pertumbuhan ekonomi dan menciptakan lapangan kerja baru".

Langkah-langkah yang peneliti lakukan adalah:

1. Menghitung *term frequency* (TF) dari kata "lapangan kerja" dalam dokumen:
   a. Frekuensi kemunculan kata "lapangan kerja" dalam dokumen = 1
   b. Jumlah kata dalam dokumen = 11
   c. TF = 1/11 = 0.09
2. Menghitung *inverse document frequency* (IDF) dari kata "lapangan kerja" dalam seluruh dokumen:
   a. Jumlah dokumen dalam korpus teks = 2
   b. Frekuensi kemunculan kata "lapangan kerja" dalam seluruh dokumen = 2
   c. IDF = log (2/2) = 0
3. Menghitung TF-IDF dari kata "lapangan kerja" dalam dokumen: TF-IDF = TF x IDF = 0.09 x 0 = 0

Dalam contoh ini, nilai IDF dari kata "lapangan kerja" adalah 0 karena kata tersebut muncul dalam semua dokumen dalam korpus teks. Oleh karena itu, nilai TF-IDF dari kata "lapangan kerja" dalam dokumen ini juga menjadi 0, yang

menunjukkan bahwa kata tersebut tidak memberikan kontribusi yang signifikan dalam membedakan dokumen ini dari dokumen lain dalam korpus teks.

Penerapan *Euclidean Distance* dalam KNN untuk analisis sentimen RUU Cipta Kerja dapat dilakukan dengan menghitung jarak antara setiap dokumen dengan dokumen uji (yang akan diklasifikasikan). Sebagai contoh, misalkan terdapat dataset yang terdiri dari 5 dokumen terkait RUU Cipta Kerja sebagaimana tabel 3.7 berikut.

Tabel 3. 7. Contoh dataset 5 dokumen

Dokumen	Jumlah kata	Sentimen
D1	1000	Negatif
D2	1500	Negatif
D3	2000	Netral
D4	2500	Positif
D5	3000	Positif

Sebuah dokumen uji (*instance*) yang terdiri dari 1800 kata akan diklasifikasikan. Untuk itu, perlu menghitung jarak *Euclidean Distance* antara dokumen uji tersebut dengan setiap dokumen pada dataset latih dan mencari 3 *nearest neighbors* (K=3) dari dokumen uji. Tabel berikut ini merupakan perhitungan *Euclidean Distance* antara dokumen uji dengan setiap dokumen pada dataset latih sebagaimana tabel 3.8 berikut.

Tabel 3. 8. *Distance* untuk Uji

Dokumen	Jumlah kata	Sentimen	*Distance* untuk Uji
D1	1000	Negatif	800
D2	1500	Negatif	300
D3	2000	Netral	200
D4	2500	Positif	700
D5	3000	Positif	1200

Dari tabel di atas, dapat dilihat bahwa K=3 dokumen terdekat dari dokumen uji adalah D3, D2, dan D1. Label sentimen mayoritas dari K tetangga tersebut diambil, yaitu sentimen Negatif, sehingga dokumen uji diklasifikasikan ke dalam kelas Negatif.

Perhitungan *Euclidean Distance* dalam KNN ini dilakukan dengan menghitung jarak antara dokumen uji dengan dokumen pada dataset latih menggunakan rumus *Euclidean*. Dalam contoh di atas, jumlah kata pada dokumen digunakan sebagai fitur untuk menghitung jarak *Euclidean Distance* antara dokumen.

Dalam kasus analisis sentimen RUU Cipta Kerja, data uji adalah dokumen yang ingin diklasifikasikan ke dalam kelas sentimen positif atau negatif. Berikut adalah contoh dataset yang terdiri dari 5 dokumen terkait RUU Cipta Kerja sebagaimana tabel 3.9 berikut.

Tabel 3. 9. Sentimen Lima dokumen terkait RUU Cipta Kerja

No	Dokumen	Sentimen
1	RUU Cipta Kerja kontroversial	negatif
2	RUU Cipta Kerja dianggap menyulitkan buruh	negatif
3	RUU Cipta Kerja dituduh tidak pro rakyat	negatif
4	RUU Cipta Kerja diharapkan meningkatkan investasi	positif
5	RUU Cipta Kerja dinilai memudahkan perizinan	positif

Misalkan terdapat dokumen uji sebagai berikut: "RUU Cipta Kerja dianggap memberikan banyak manfaat bagi dunia usaha." Untuk mengklasifikasikan dokumen uji ini menggunakan KNN dengan *Euclidean Distance*, langkah-langkah yang dapat dilakukan adalah sebagai berikut:

1. Mengubah dokumen-dokumen dalam dataset menjadi vektor numerik. Salah satu teknik yang umum digunakan adalah TF-IDF.
2. Menghitung nilai Euclidean Distance antara dokumen uji dengan masing-masing dokumen dalam dataset.
3. Mengambil k dokumen dengan jarak terdekat dengan dokumen uji. Misalkan k = 3.
4. Mengambil mayoritas sentimen dari k dokumen terdekat sebagai sentimen dari dokumen uji. Jika mayoritas adalah sentimen positif, maka dokumen uji akan diklasifikasikan sebagai sentimen positif, begitu pula sebaliknya.

Dalam contoh ini, diasumsikan menggunakan TF-IDF untuk mengubah dokumen menjadi vektor numerik. Setelah itu, hasil perhitungan *Euclidean Distance* dapat dilihat pada tabel 3.10 berikut:

Tabel 3. 10. Hasil perhitungan *Euclidean Distance*

No	Dokumen	Sentimen	Jarak dengan Dokumen Uji
1	RUU Cipta Kerja kontroversial	negatif	2.45
2	RUU Cipta Kerja dianggap menyulitkan buruh	negatif	3.46
3	RUU Cipta Kerja dituduh tidak pro rakyat	negatif	3.32
4	RUU Cipta Kerja diharapkan meningkatkan investasi	positif	4.24
5	RUU Cipta Kerja dinilai memudahkan perizinan	positif	4.24

Cara menghitung jarak *euclidean* antara vektor tweet baru dengan vektor setiap tweet dalam dataset menggunakan rumus sebagaimana persamaan 5.

Misalkan terdapat dua vektor tweet sebagai berikut:

x = [0.5, 0.7, 0.3, 0.8, 0.2]

y = [0.2, 0.4, 0.9, 0.6, 0.1]

Maka, jarak *Euclidean* antara kedua vektor tersebut dapat dihitung sebagai berikut:

$$d(x, y) = \sqrt{[(0.5 - 0.2)^2 + (0.7 - 0.4)^2 + (0.3 - 0.9)^2 + (0.8 - 0.6)^2 + (0.2 - 0.1)^2]}$$

$$d(x, y) = \sqrt{[0.09 + 0.09 + 0.36 + 0.04 + 0.01]}$$

$$d(x, y) = \sqrt{0.59}$$

$$d(x, y) = 0.77$$

Jadi, jarak *Euclidean* antara vektor x dan y adalah 0.77.

Contoh perhitungan *Random Forest* pada data PERPPU Cipta Kerja sebagai berikut.

Disajikan dataset yang terdiri dari 1000 baris dan 10 kolom, dengan kolom pertama hingga kesembilan adalah fitur atau atribut dan kolom ke-10 adalah variabel target yang ingin diprediksi, yaitu "status dukungan" (mendukung atau menolak) terhadap PERPPU Cipta Kerja. Adapun Langkah-langkahnya sebagai berikut.

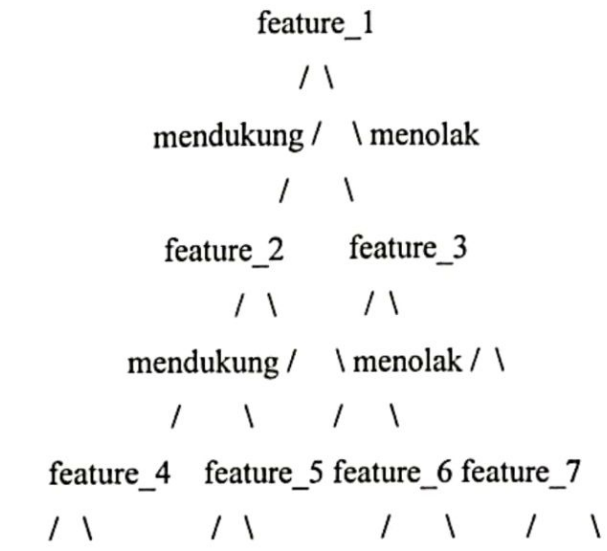
1. Melakukan *preprocessing* data.
2. Melakukan ekstraksi fitur dengan menggunakan TF-IDF.
3. Memilih 600 data tweet secara acak sebagai data train dan 400 data tweet yang tersisa sebagai data test.
4. Membuat 10 *decision tree* dan melakukan training pada data train dengan subset data yang dipilih secara acak dengan pengulangan. Setiap *decision tree* memiliki kedalaman maksimum 5.
5. Memprediksi sentimen dari data test menggunakan setiap *decision tree* yang telah dibuat.

6. Mengambil mayoritas dari semua hasil prediksi setiap *decision tree* sebagai hasil prediksi *Random Forest*.
7. Menghitung akurasi model menggunakan *confusion matrix*.
8. Menampilkan hasil akurasi model.

Contoh hasil perhitungan manual untuk satu decision tree sebagai berikut:

Decision Tree 1:

1. Menggunakan data train dengan 200 data tweet secara acak.
2. Memilih fitur-fitur dari data train dengan menggunakan TF-IDF.
3. Membuat *decision tree* dengan ke dalaman maksimum 5.
4. *Decision tree* yang dihasilkan memiliki bentuk sebagai berikut:



5. *Decision tree* di atas menggunakan 7 fitur: feature_1, feature_2, feature_3, feature_4, feature_5, feature_6, dan feature_7.

Setelah membuat 10 *decision tree*, dilakukan prediksi pada data test dan diambil mayoritas hasil prediksi dari setiap *decision tree* sebagai hasil prediksi

Random Forest. Akurasi model kemudian dihitung dengan menggunakan *confusion matrix*. Contoh hasil akurasi model sebagai berikut:

	Actual	
	Mendukung	Menolak
Predicted Mendukung	180	30
Menolak	35	155

Dari hasil *confusion matrix* di atas, dapat dihitung akurasi model sebagai berikut:

Akurasi = (180 + 155) / (180 + 30 + 35 + 155) = 0.8375

Dengan demikian, akurasi model *Random Forest* pada data PERPPU Cipta Kerja adalah 0.8375 atau 83.75%.

Confusion matrix adalah tabel yang digunakan untuk mengevaluasi performa dari sebuah model klasifikasi dengan membandingkan prediksi model dengan data aktual. Misalkan terdapat data sentimen terhadap produk yang terdiri dari 100 data, dengan 50 data positif dan 50 data negatif. Kemudian dibuat model klasifikasi untuk memprediksi sentimen data tersebut. Setelah memprediksi, hasil prediksi dibandingkan dengan data aktual dan didapatkan *confusion matrix* sebagaimana tabel 3.11 berikut:

Tabel 3. 11. Contoh hasil *confusion matrix*

Prediksi	Aktual Positif	Aktual Negatif
Prediksi Positif	40 (TP)	10 (FP)
Prediksi Negatif	10 (FN)	40 (TN)

Berdasarkan tabel 3.11 di atas, terdapat 40 *True Positive* (TP) yaitu data positif yang diprediksi benar sebagai positif. Selain itu, juga terdapat 40 *True*

Negative (TN) yaitu data negatif yang diprediksi benar sebagai negatif. Sedangkan *False Positive* (FP) terdapat 10 data negatif yang diprediksi salah sebagai positif, dan *False Negative* (FN) terdapat 10 data positif yang diprediksi salah sebagai negatif.

Dari *confusion matrix* tersebut, dapat diakukan evaluasi performa model dengan berbagai metode, yaitu:

Accuracy: (TP + TN) / (TP + FP + TN + FN)
= (40 + 40) / 100
= 0.8 atau 80%

Precision: TP / (TP + FP)
= 40 / (40 + 10)
= 0.8 atau 80%

Recall: TP / (TP + FN)
= 40 / (40 + 10)
= 0.8 atau 80%

F1-score: 2 * (*Precision* * *Recall*) / (*Precision* + *Recall*)
= 2 * (0.8 * 0.8) / (0.8 + 0.8) = 0.8

Dalam kasus ini, model klasifikasi memiliki akurasi 80%, *precision* 80%, *recall* 80%, dan F1-score 0.8, yang menunjukkan performa yang baik.

Contoh penerapan perhitungan evaluasi metode dengan 10-*fold cross-validation* pada analisis sentimen Perppu Cipta Kerja sebagai berikut.

1. Dataset yang berisi teks mengenai Perppu Cipta Kerja dan label sentimennya (positif, negatif, atau netral).

   Dataset terdiri dari 1000 teks mengenai Perppu Cipta Kerja, dengan label sentimen positif, negatif, atau netral.

2. Bagi dataset menjadi 10 subset yang sama besar.

Dataset dibagi menjadi 10 subset yang masing-masing terdiri dari 100 teks.

3. Untuk setiap *subset*, gunakan *subset* tersebut sebagai data uji (*test set*) dan gabungkan sisa *subset* menjadi data latih (*training set*).

   Pada iterasi pertama, digunakan *subset* 1 sebagai *test set* dan *subset* 2-10 digabungkan sebagai *training set*. Pada iterasi kedua, digunakan *subset* 2 sebagai *test set* dan *subset* 1, 3-10 digabungkan sebagai *training set*, dan seterusnya.

4. Melatih model analisis sentimen menggunakan *training set* dan evaluasi performa model pada *test set*.

   Pada setiap iterasi, dilatih model menggunakan *training set* dan mengukur performanya pada test set untuk mengukur akurasi, presisi, *recall*, F1-score, dan AUC-ROC.

5. Mengulangi langkah 3-4 untuk setiap *subset* sehingga setiap *subset* telah menjadi data uji satu kali.

   Langkah 3-4 diulang untuk setiap *subset*, sehingga setiap *subset* telah menjadi test set satu kali.

6. Hitung rata-rata performa model pada ke-10 iterasi *cross-validation*.

   Setelah melakukan *10-fold cross-validation*, hitung rata-rata performa model pada ke-10 iterasi. Misalnya, menghitung rata-rata akurasi, presisi, dan *recall*,

Dengan melakukan *10-fold cross-validation*, penulis dapat memastikan bahwa model analisis sentimen yang digunakan memiliki performa yang stabil dan konsisten pada dataset yang dimiliki.