BAB III

METODOLOGI PENELITIAN

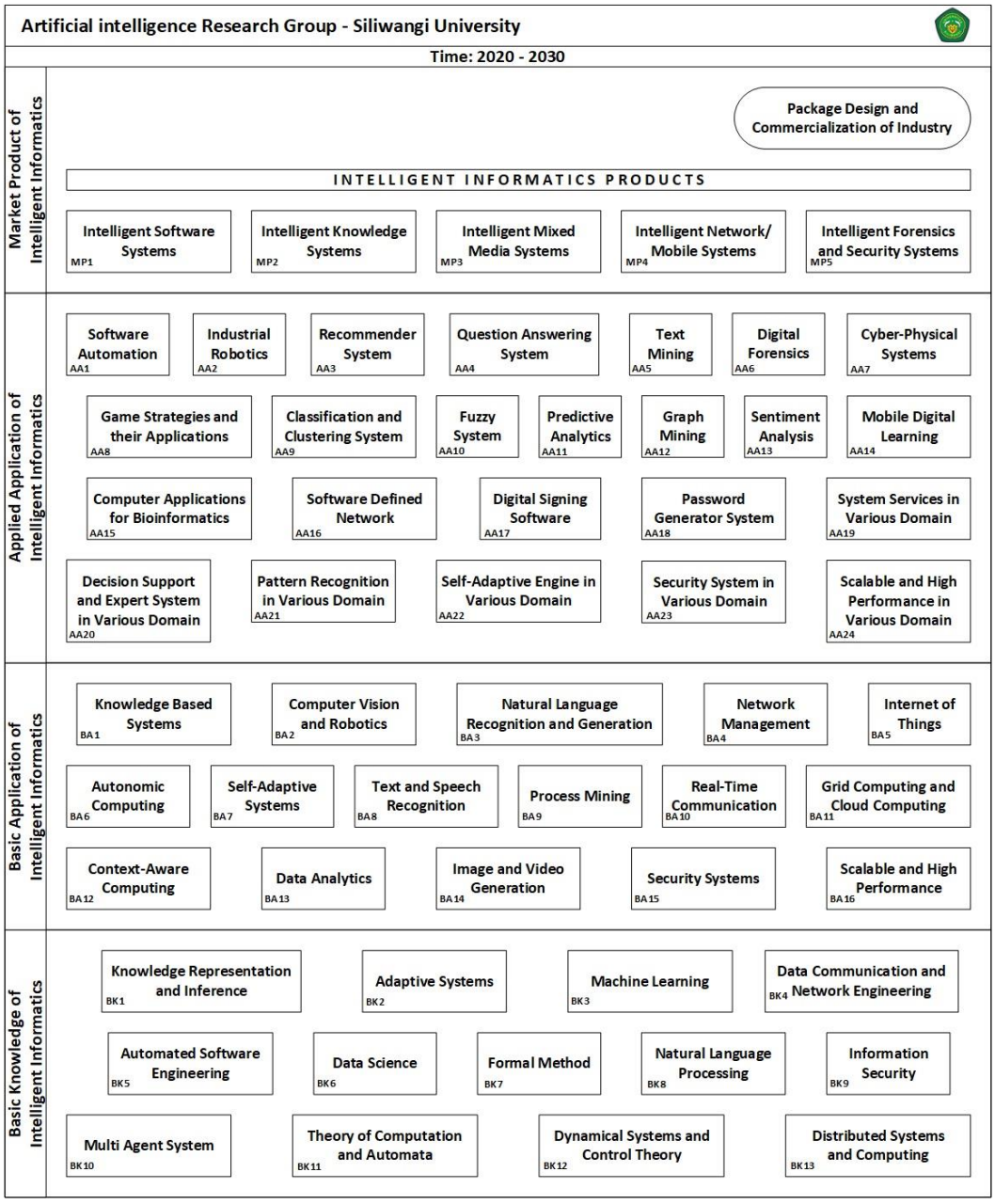
3.1 Metode Penelitian

Metode penelitian merupakan cara ilmiah untuk mendapatkan data dengan tujuan kegunaan tertentu (Sugiyono, 2015). Pada penelitian ini, metode penelitian yang digunakan yaitu metode penelitian kuantitatif. Metode penelitian kuantitatif merupakan jenis penelitian yang spesifikasinya sistematis, terencana, dan terstruktur dengan jelas sejak awal hingga pembuatan desain penelitiannya (Siyoto dan Sodik, 2015). Pada metode penelitian kuantitatif, umumnya teknik pengambilan sampel dilakukan secara acak, pengumpulan data menggunakan instrumen penelitian, analisis data bersifat kuantitatif atau statistik yang bertujuan untuk menguji hipotesis yang telah ditetapkan (Sugiyono, 2015).

Pada penelitian ini, data dikumpulkan dengan observasi pada media sosial *Twitter* tentang program vaksinasi covid-19 di Indonesia. Proses ini menggunakan alat bantu untuk mendapatkan data *tweet* pada media sosial *Twitter*. Pengambilan sampel ini dilakukan secara acak dengan periode waktu tertentu. Data yang telah dikumpulkan kemudian dianalisis secara statistik dengan menggunakan algoritma *machine learning* untuk mendapatkan sentimen yang ditunjukkan pada kalimat *tweet* tersebut apakah positif, netral, atau negatif.

3.2 Peta Jalan (*Road Map*) Penelitian

Peta jalan atau yang biasa disebut dengan *roadmap* merupakan suatu konsep arah penelitian yang dimaksudkan untuk menjelaskan ke arah mana penelitian akan dituju. *Roadmap* pada penelitian ini mengacu kepada *roadmap Artificial Intelligence Research Group* – Universitas Siliwangi tahun 2020 – 2030. *Roadmap Artificial Intelligence Research Group* – Universitas Siliwangi tahun 2020 – 2030 ini merupakan kolaborasi antara Kelompok Keahlian (KK) Informatika dan Sistem Inteligen (ISI) bersama dengan Kelompok Keahlian (KK) Jaringan, Keamanan, dan Forensika Digital (JKF) jurusan Informatika, Fakultas Teknik, Universitas Siliwangi. *Roadmap Artificial Intelligence Research Group* – Universitas Siliwangi tahun 2020 – 2030 ditunjukkan pada Gambar 3.1.



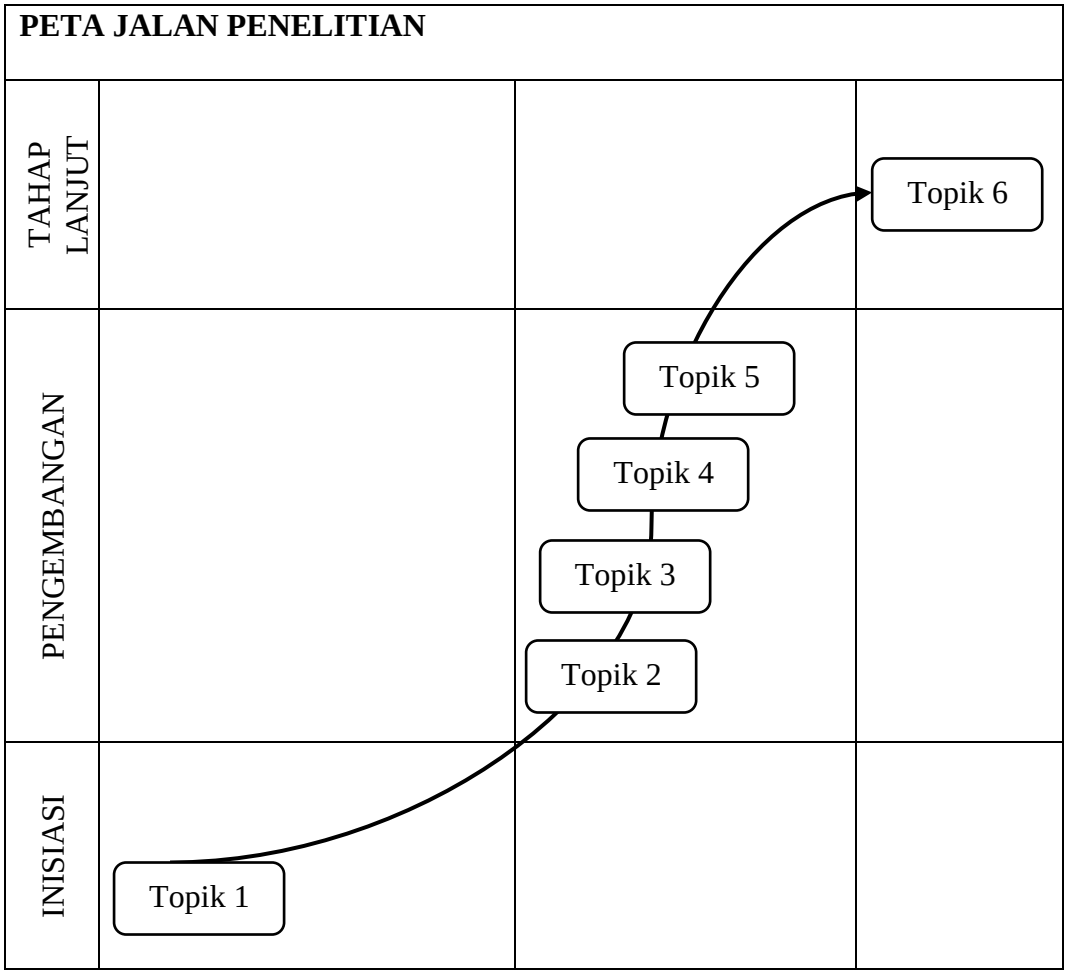
Gambar 3.1 *Roadmap AI Research Group Universitas Siliwangi 2020 – 2030*

(Sumber : AIS, 2020)

Berdasarkan *Roadmap* AI Research Group Universitas Siliwangi 2020 – 2030, *Basic Knowledge* yang digunakan pada penelitian ini yaitu *Natural Language Processing* (NLP) dan *Machine Learning* (ML), *Basic Application* pada penelitian

ini yaitu *Data Analytics*, serta *Applied Application* pada penelitian ini yaitu *Text Mining* dan *Sentiment Analysis*.

Terdapat tiga tahapan dalam *roadmap* penelitian ini, yaitu tahap inisiasi, tahap pengembangan, dan tahap lanjut atau tahap hilir yang ditunjukkan pada Gambar 3.2.



Gambar 3.2 *Peta Jalan Penelitian*

Keterangan :

Topik 1 : Studi Literatur

Topik 2 : Pengumpulan, Analisis Data, dan Persiapan Data

Topik 3 : Mengimplementasikan model *Machine Learning* dengan mengombinasikan beberapa algoritma *machine learning* pada teknik *Stacking Ensemble Classifier*

Topik 4 : Pengujian model yang dibuat dan pengukuran akurasi terhadap data sampel

Topik 5 : Evaluasi hasil tingkat akurasi model yang dibuat.

Topik 6 : Pengembangan model yang telah dibuat.

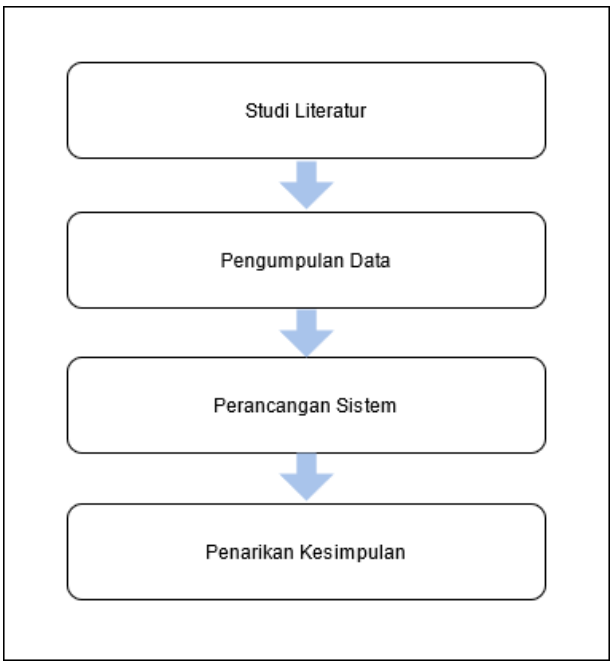
Tahap inisiasi merupakan tahap awal pada penelitian ini yaitu studi literatur. Tahap studi literatur ini merupakan tahapan untuk mencari dan memahami pengetahuan yang terkait dengan penelitian.

Tahap pengembangan merupakan tahapan inti pada penelitian ini. Tahapan ini terdiri dari 1) pengumpulan, analisis, serta persiapan data, 2) membuat model *machine learning* dengan teknik *stacking ensemble classifier*, 3) pengujian model yang telah dibuat, dan 4) evaluasi hasil tingkat akurasi yang didapatkan oleh model yang telah dibuat.

Tahap lanjut merupakan tahapan pengembangan yang lebih mendalam terhadap model yang telah dibuat. Tahapan ini akan memperbaiki kekurangan dari model yang telah dibuat. Dalam memperbaiki kekurangan tersebut, dapat dilakukan perbaikan berupa mengombinasikan dengan teknik lain untuk mendapatkan hasil yang lebih baik.

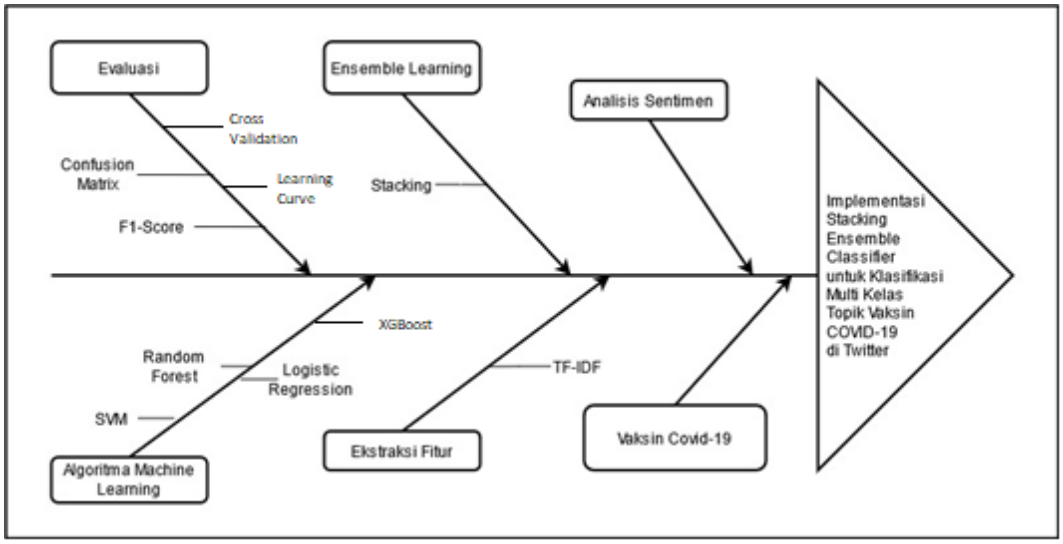
3.3 Tahapan Penelitian

Proses penelitian yang akan dilakukan digambarkan di dalam sebuah diagram alur. Pembuatan diagram alur dibuat agar mempermudah menyampaikan informasi terhadap langkah – langkah yang akan dilakukan. Tahapan penelitian secara umum ditunjukkan pada Gambar 3.3.



Gambar 3.3 *Gambaran Umum Tahapan Penelitian*

Tahapan penelitian yang akan dilakukan yaitu dimulai dari studi literatur, proses pengumpulan data, analisis data, membuat model *stacking ensemble classifier*, pengujian model *stacking ensemble classifier* yang dibuat menggunakan data sampel, dan sampai proses penarikan kesimpulan. Sedangkan diagram *fishbone* pada penelitian ini ditunjukkan pada Gambar 3.4.



Gambar 3.4 *Fishbone Diagram Penelitian*

Vaksin Covid-19 merupakan data yang digunakan pada penelitian ini. Data ini berupa data cuitan pengguna media sosial twitter mengenai vaksin covid-19 di Indonesia.

Analisis Sentimen merupakan *research field* pada penelitian ini. Dari dataset yang didapatkan, akan diketahui bagaimana sentimen masyarakat terhadap vaksin covid-19 di Indonesia.

Ekstraksi Fitur merupakan metode untuk mengubah data teks menjadi matriks angka. Ekstraksi fitur yang digunakan adalah metode TF – IDF untuk mencari frekuensi kata tiap kalimat pada dataset. Hasil keluaran dari metode ini merupakan matriks yang berisi frekuensi kata pada kalimat.

Ensemble Learning merupakan metode yang digunakan untuk pemodelan terhadap dataset. Metode *ensemble learning* yang digunakan pada penelitian ini

adalah *stacking*. Konsep dari metode *stacking* adalah menumpuk beberapa algoritma *machine learning*.

Algoritma *Machine Learning* yang digunakan pada penelitian ini yaitu *Support Vector Machine, Logistic Regression,* dan *Random Forest* sebagai *first level learners* pada metode *stacking*, setra *second level* atau *meta learner* menggunakan algoritma *Logistic Regression* untuk model pertama dan *XGBoost* untuk model kedua.

Evaluasi terhadap model *stacking* pada penelitian ini menggunakan metode *confusion matrix* untuk melihat matriks hasil prediksi yang didapatkan oleh model. *Learning Curve* digunakan untuk mengetahui performa *training* pada model dengan menggunakan *Cross Validation*. Metode *F1-Score* digunakan untuk mengetahui performa yang dihasilkan dari model *stacking* pada penelitian ini.

3.3.1. Studi Literatur

Tahapan ini merupakan tahap pengumpulan berbagai data dan sumber yang berhubungan dengan penelitian seperti teori Analisis Sentimen, *Text Preprocessing*, *Ensemble Method,* algoritma *Logistic Regression, Random Forest, Support Vector Machine, Stacking Ensemble Classifier*, *Confusion Matrix,* dan *F1-Score*. Studi literatur didapatkan dari jurnal dan *e-proceeding* terkait metode dan algoritma yang digunakan, selain itu diperoleh dari buku dan internet serta dokumen lain yang berhubungan dengan objek penelitian.

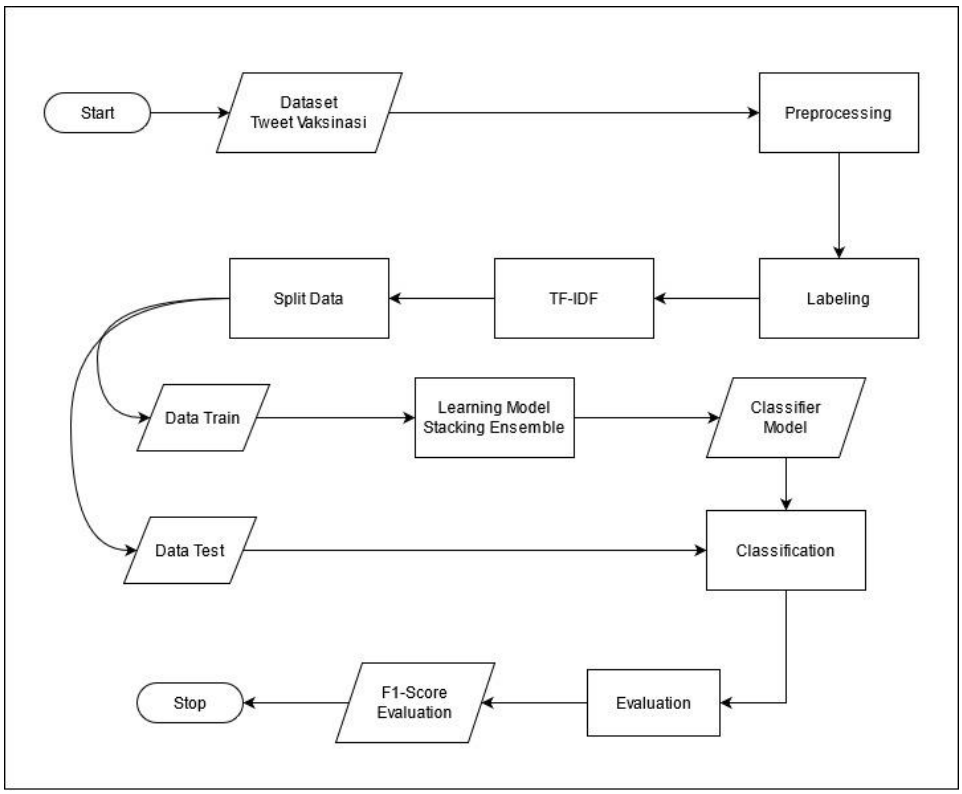
3.3.2. Pengumpulan Data

Data yang digunakan pada penelitian ini merupakan *tweet* berbahasa Indonesia yang ditulis oleh pengguna *Twitter*. Data *tweet* tersebut dikumpulkan

dengan menggunakan alat bantu yaitu *twint* yang dikembangkan oleh TWINT Project (Zacharias dan Poldi, 2020) untuk mempermudah dan mempersingkat waktu dalam tahap pengumpulan data. Adapun data *tweet* yang dikumpulkan adalah data yang tersedia sejak tanggal 1 Mei 2021 sampai 10 Juli 2021 dengan menggunakan kata kunci “vaksin covid-19”.

3.3.3. Perancangan Sistem

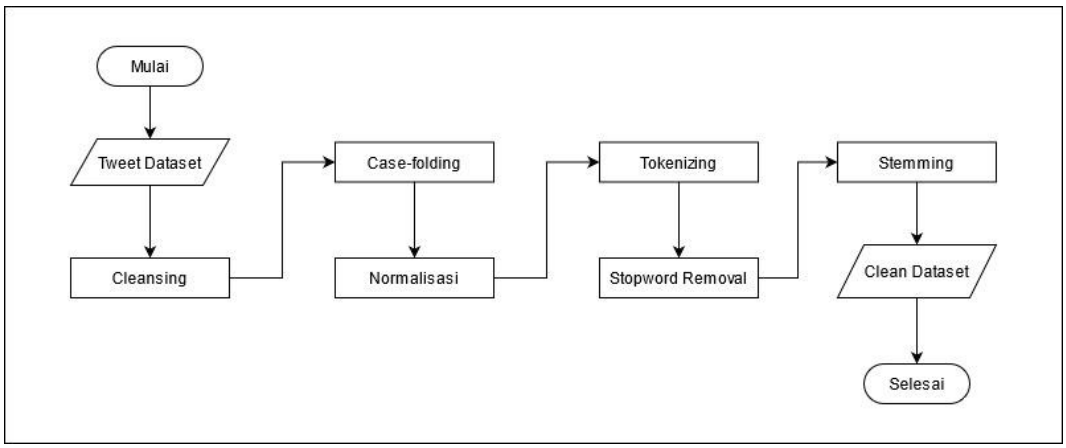
Pada tahapan ini akan dijelaskan tentang rancangan sistem yang dibangun. Gambaran umum perancangan sistem pada penelitian ini ditunjukkan pada Gambar 3.5.



Gambar 3.5 *Gambaran Umum Perancangan Sistem*

A. *Preprocessing*

Dalam *text mining*, tahap *preprocessing* dilakukan untuk mengolah data yang bermacam – macam menjadi sebuah data yang teratur yang dapat diterapkan pada algoritma *machine learning* (Jaka, 2015). Adapun tahap *preprocessing* pada penelitian ini terdiri dari *cleansing, case-folding, perbaikan kata singkat, tokenizing*, penghilangan *stopword*, dan *stemming*. Proses tersebut ditunjukkan pada Gambar 3.6.



Gambar 3.6 *Tahapan Preprocessing*

1. *Cleansing*

Proses *cleansing* dilakukan untuk mengkilangkan elemen yang tidak perlu pada teks *tweet* yang dianggap tidak penting. Elemen – elemen tersebut terdiri dari *hashtag, mention, link, RT (retweet),* tanda baca, nomor, gambar, serta *whitespace*.

2. *Case-folding*

Proses *case-folding* dilakukan untuk mengubah semua huruf kapital menjadi huruf kecil. Pada proses ini hanya huruf a-z yang diubah.

3. Normalisasi

Proses ini dilakukan untuk memperbaiki kata singkat pada *tweet*. Hal ini dikarenakan jumlah kata yang dapat dimasukkan pada *tweet* terbatas sehingga

banyak pengguna *Twitter* menyingkat kata untuk memperpendek kalimat yang akan diunggah. Proses ini memanfaatkan kamus yang telah dibuat oleh (Aliyah Salsabila *dkk.*, 2018) yang merupakan kumpulan "kata alay" yang telah diperbaiki menjadi kata formal atau baku. Algoritma dari proses normalisasi ini adalah sebagai berikut:

a. Masukkan kalimat;
b. Bandingkan setiap kata pada kalimat dengan kamus kata alay;
c. Jika kata terdapat pada kamus kata alay, ganti kata tersebut dengan kata baku pada kamus alay;
d. Keluaran berupa kalimat dengan kata baku.

4. *Tokenizing*

Proses *tokenizing* dilakukan untuk memisahkan atau memotong teks input berdasarkan tiap – tiap kata yang tersusun.

5. Penghilangan *Stopword*

Proses ini dilakukan untuk menghilangkan kata depan, kata ganti, kata sambung, dan kata yang tidak ada hubungannya dengan analisis sentimen.

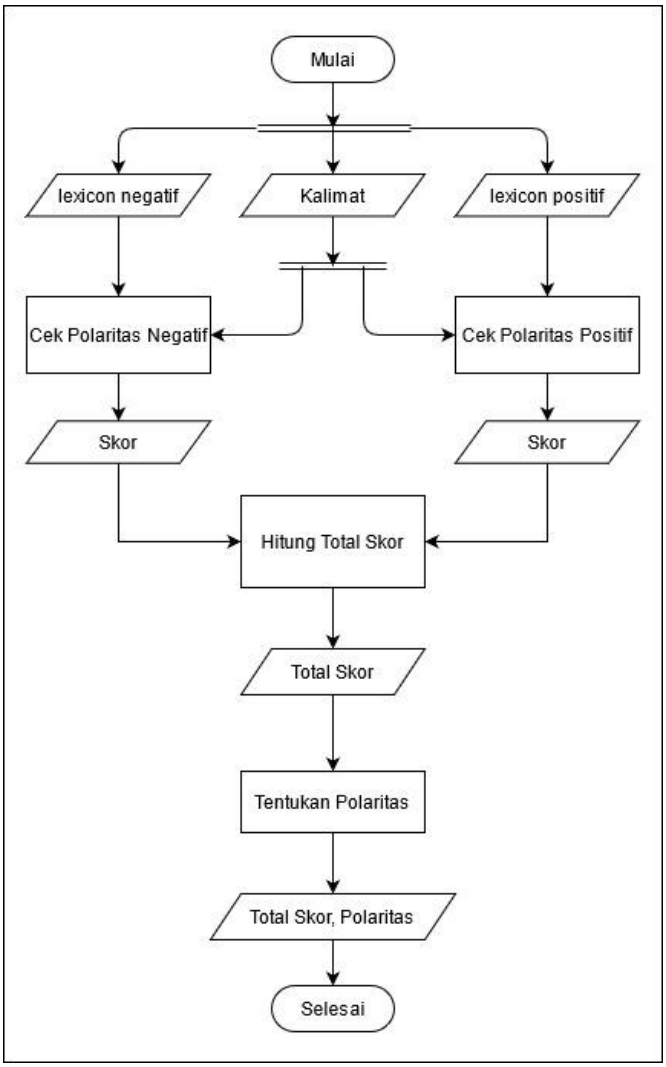
6. *Stemming*

Proses ini dilakukan untuk mengubah sebuah kata menjadi kata dasarnya.

B. *Dataset Labeling*

Dataset labeling atau pelabelan dataset merupakan tahap pemberian label pada data yang telah diperoleh. Hal ini dilakukan karena data yang diperoleh hanya berupa teks *tweet* saja, belum ada label yang menunjukkan bahwa *tweet* tersebut memiliki sentimen negatif, positif, atau netral. Pada proses pelabelan ini, data diberi label "*positive"* untuk sentimen positif, "*neutral*" untuk sentimen netral, dan

"*negative*" untuk sentimen negatif. Proses pelabelan dataset ini ditunjukkan pada Gambar 3.7.



Gambar 3.7 *Proses Pelabelan Data*

Proses pelabelan ini dilakukan dengan cara mencari polaritas kalimat dengan menggunakan kamus InSet (*Indonesia Sentiment Lexicon*) (Koto dan Rahmaningtyas, 2017). Pada proses ini, tiap kata pada kalimat dalam dataset dibandingkan dengan kata pada kamus data InSet untuk mendapatkan polaritasnya kemudian menjumlahkan skornya. Skor polaritas tiap kalimat dalam dataset

menentukan label untuk kalimat tersebut. Jika jumlah skor lebih dari 0 maka diberi label "*positive*". Jika jumlah skor sama dengan 0 maka diberi label "*neutral*". Dan jika jumlah skor kurang dari 0 maka diberi label "*negative*".

C. Ekstraksi Fitur dengan TF-IDF

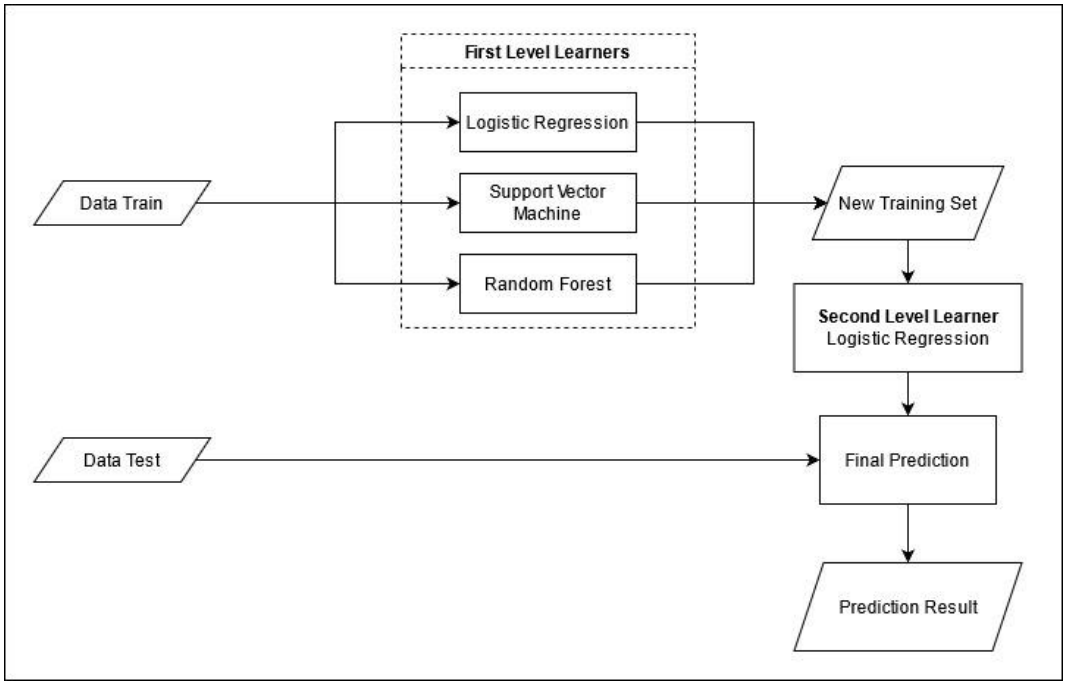
Proses ekstraksi fitur dengan *TF-IDF* dilakukan untuk mengubah *term* menjadi bilangan numerik yang akan diproses sebagai data uji serta data latih. Ekstraksi fitur dengan *TF-IDF* merupakan proses ekstraksi teks dengan memberikan nilai pada masing masing kata dalam suatu kalimat . Pada penelitian ini, proses ekstraksi fitur memanfaatkan *library sklearn* yaitu modul *CountVectorizer* dan *TfidfTransformer*.

D. *Split Data*

Tahap ini merupakan pemisahan data menjadi 2 bagian, yaitu data latih dan data tes. Data latih digunakan untuk melatih algoritma *machine learning* agar algoritma dapat belajar dari data yang tersedia. Data tes digunakan untuk menguji algoritma *machine learning* tersebut yang menghasilkan tingkat akurasi atau skor dari algoritma yang digunakan.

E. *Learning Model Stacking Ensemble*

Tahapan ini merupakan tahap pembuatan model untuk klasifikasi dataset dengan metode *stacking ensemble classifier*. Rancangan model *stacking ensemble classifier* pada penelitian ini ditunjukkan pada Gambar 3.8.



Gambar 3.8 *Rancangan Model Stacking Ensemble Classifier*

Model *stacking* yang dibuat pada penelitian ini terdiri dari 2 *level learner*. Level 1 yaitu *first level learner* menggunakan 3 algoritma yang terdiri dari *Logistic Regression* (LR), *Support Vector Machine* (SVM), dan *Random Forest* (RF). Pada *first level learner* ini, data latih digunakan untuk melatih ke 3 algoritma tersebut yang akan menghasilkan set baru data latih untuk digunakan pada *second level learner* atau *meta learner*. Sedangkan *second level learner* atau *meta learner* menggunakan algoritma *Logistic Regression*. Hasil dari *meta learner* ini digunakan untuk memprediksi data tes yang nantinya akan dievaluasi performa dari model.

F. Evaluasi

Evaluasi performa model yang dibuat pada penelitian ini menggunakan metode *accuracy*, *precision*, *recall*, dan *F1-Score*. Untuk menghitung metode evaluasi yang sudah disebutkan, dibutuhkan *True Prositive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) dari *confusion matrix* setiap

label (Yutika, Adiwijaya dan Faraby, 2021). *Confusion matrix* merupakan metode yang meringkas kinerja klasifikasi *clasifier* sehubungan dengan beberapa data pengujian (Ting, 2010). Tabel *confusion matrix* yang digunakan pada penelitian ini ditunjukkan pada Tabel 3.

Tabel 3.1 *Confusion Matrix Tiga Kelas Sentimen*

		Predicted		
		Positive	**Neutral**	**Negative**
Actual	**Positive**	P_{PP}	P_{NtP}	P_{NgP}
	Neutral	P_{PNt}	P_{NtNt}	P_{NgNt}
	Negative	P_{PNg}	P_{NtNg}	P_{NgNg}

Sumber : (Saputro, Notodiputro dan Indahwati, 2018)

Berdasarkan Tabel 3.1, penghitungan nilai *accuracy*, *precision*, *recall*, pada penelitian ini dapat dirumuskan sebagai berikut (Saputro, Notodiputro dan Indahwati, 2018) :

$$Accuracy = \frac{P_{PP}+NgNg+NtNt}{P_{PP}+NgNg+NtNt+P_{NtP}+P_{NgP}+P_{PNt}+P_{NgNt}+P_{PNg}+P_{NtNg}} \; x\; 100\% \quad (3.1)$$

$$Precission.positive = = \frac{P_{PP}}{P_{PP}+P_{PNt}+P_{PNg}} \; x\; 100\% \quad (3.2)$$

$$Precission.negative = = \frac{P_{NgNg}}{P_{NgNg}+P_{NgNt}+P_{NgP}} \; x\; 100\% \quad (3.3)$$

$$Precission.neutral = = \frac{P_{NtNt}}{P_{NtNt}+P_{NtP}+P_{NtNg}} \; x\; 100\% \quad (3.4)$$

$$Recall.positive = = \frac{P_{PP}}{P_{PP}+P_{NtP}+P_{NgP}} \; x\; 100\% \quad (3.5)$$

$$Recall.negative = = \frac{P_{NgNg}}{P_{NgNg}+P_{NtNg}+P_{PNg}} \; x\; 100\% \quad (3.6)$$

$$Recall.neutral = = \frac{P_{NtNt}}{P_{NtNt}+P_{PNt}+P_{NgNt}}\ x\ 100\% \quad (3.7)$$

Untuk menghitung *F1 – Score*, berdasarkan hasil penghitungan nilai *precision* dan *recall* dengan rumus di atas dapat dirumuskan sebagai berikut (Han, Kamber dan Pei, 2012) :

$$F1 - Score.positive = 2\ x\ \frac{precision.Pst\ x\ recall.Pst}{precision.Pst + recall.Pst} \quad (3.8)$$

$$F1 - Score.negative = 2\ x\ \frac{precision.Ng\ x\ recall.Ng}{precision.Ng + recall.Ng} \quad (3.9)$$

$$F1 - Score.neutral = 2\ x\ \frac{precision.Nt\ x\ recall.Nt}{precision.Nt + recall.Nt} \quad (3.10)$$

3.3.4. Penarikan Kesimpulan

Tahapan ini merupakan tahapan akhir penelitian, di mana hasil dari penelitian tersebut berupa performa model *stacking ensemble classifier* yang dibuat untuk analisis sentimen opini masyarakat mengenai program vaksinasi di Indonesia yang diambil dari media sosial *Twitter*. Penarikan kesimpulan berupa hasil akhir performa model yang dibuat dengan kelebihan dan kekurangannya.