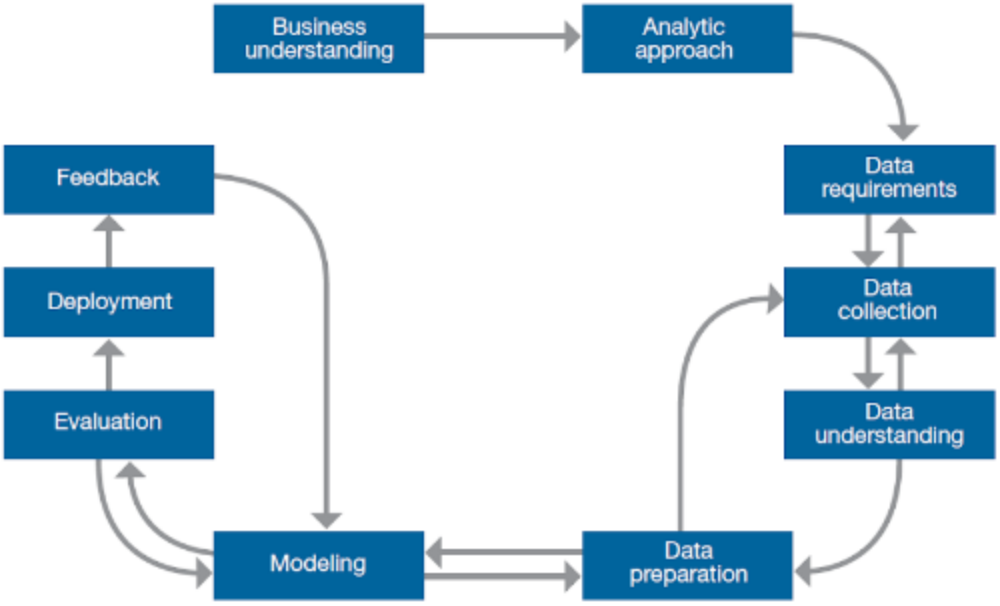
BAB III

METODOLOGI PENELITIAN

3.1. *Foundational Methodology for Data Science*

Penelitian ini akan menggunakan *Foundational Methodology for Data Science/FMDS.* Gambar 3.1 menunjukkan alur kerja dari FMDS. Metodologi ini merupakan pengembangan dari kerangka CRISP-DM (IBM, 2021). Metodologi ini memiliki 10 langkah yang dibagi menjadi 5 modul, yaitu: 1) dari masalah hingga pendeketan, 2) dari kebutuhan hingga pengumpulan, 3) dari pemahaman hingga persiapan, 4) dari pemodelan hingga evaluasi, dan 5) dari *deployment* hingga *feedback.*



Gambar 3.1 *Foundational Methodology for Data Science*
(Sumber : (IBM, 2021))

Pada *framework* CRISP-DM memiliki langkah-langkah: Pemahaman Bisnis, Pemahaman Data, Persiapan Data, Pemodelan, Evaluasi, dan *Deployment*.

Sedangkan *framework* FMDS memiliki lebih banyak langkah dalam penelitiannya, antara lain: pendekatan analitis setelah *business understanding*, kebutuhan data, pengumpulan data, sebelum *data understanding*, dan *feedback* setelah *deployment*.

3.2. Tahapan Penelitian

3.2.1. Business Understanding

Setiap proyek maupun penelitian dimulai dengan *business understanding* atau pemahaman bisnis yang menjadi dasar solusi efektif bagi sebuah permasalahan bisnis. Fase ini betujuan untuk menjawab pertanyaan, “Apa permasalahan yang coba untuk dicarikan solusinya?” Pada fase ini penelitian dimulai dengan mendefinisikan masalah, menentukan tujuan dan cakupan penelitian. Pada penelitian ini, tahap pemahaman bisnis tertuang pada Bab I Pendahuluan.

3.2.2. Analytical Approach

Setelah masalah bisnis dipahami dengan jelas, maka penelitian ini dilanjutkan dengan menentukan pendekatan untuk memecahkan masalah. Fase ini bertujuan untuk menjawab pertanyaan, “Bagaimana peneliti bisa menggunakan data yang ada untuk memecahkan permasalahan?” Penelitian ini akan menggunakan pendekatan dengan mengimplementasikan penggunaan *Ensemble Machine Learning Classifier* dan SMOTE untuk menganalisis sentimen terhadap SDGs di Indonesia. Fase penelitian ini dijelaskan pada Bab 2 mengenai algoritma dan pendekatan yang digunakan pada penelitian ini.

3.2.3. Data Requirement

Pendekatan analitik yang dipilih menentukan kebutuhan data. Tahapan ini bertujuan untuk menjawab pertanyaan, “Apa data yang dibutuhkan untuk memecahkan permasalahan yang dimiliki?” Secara khusus, metode analitik yang akan digunakan memerlukan konten, format, dan representasi data tertentu, yang dipandu oleh pengetahuan domain. Data yang dibutuhkan untuk penelitian ini adalah berupa data twit mengenai SDGs yang bersumber dari pengguna internet di Indonesia.

3.2.4. Data Collection

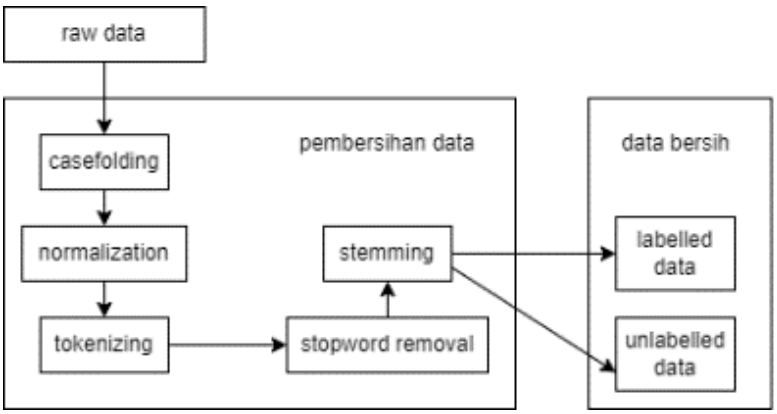
Pada tahap ini, penelitian mulai melakukan pengumpulan data. Tahapan ini bertujuan untuk menjawab pertanyaan, “Darimana data berasal dan bagaimana cara mendapatkannya?” Data yang digunakan dalam penelitian ini adalah sekumpulan twit berbahasa Indonesia mengenai 17 SDGs pada bulan Juli 2022. Data diperoleh berasal dari Twitter dengan menggunakan Twitter API (*Application Programming Interface*) melalui RapidMiner.

3.2.5. Data Understanding

Setelah pengumpulan data selesai, selanjutnya data ditinjau dan menentukan apakah data yang diperoleh sesuai dengan kebutuhan yang bisnis dan kebutuhan data. Tahapan ini bertujuan untuk menjawab pertanyaan, “Apa data yang diperoleh sudah memenuhi kebutuhan untuk memecahkan masalah?” Pada tahap ini penulis memastikan data yang diperoleh sudah sesuai kebutuhan, serta memahami isi dan informasi awal data yang diperoleh.

3.2.6. Data Preparation

Tahap ini meliputi semua kegiatan untuk mengkonstruksi kumpulan data yang akan digunakan pada tahap pemodelan. Fase penelitian bertujuan untuk menjawab pertanyaan, "Apa saja yang harus dilakukan peneliti untuk mempersiapkan data sehingga bisa melanjutkan pemodelan dengan menggunakan data yang ada?" Secara garis besar fase ini memiliki 2 tahapan yang ditunjukkan oleh Gambar 3.2, pertama yaitu melakukan pembersihan data, kedua adalah melakukan pelabelan terhadap data yang akan digunakan pada tahap pemodelan.



Gambar 3.2 skema data preparation

3.2.7. Modeling

Tahap pemodelan berfokus pada pengembangan model prediktif atau deskriptif menurut pendekatan analitik yang ditentukan sebelumnya. Pada tahap ini peneliti melakukan beberapa kali percobaan dengan beberapa algoritma dan kombinasi algoritma dengan parameternya masing-masing untuk menemukan model terbaik. Fase ini bertujuan untuk menjawab pertanyaan, "Bagaimana data dapat divisualisasikan untuk mendapatkan solusi yang diperlukan?" Penelitian

ini akan membandingkan performa dari algoritma *Naïve Bayes, Support Vector Machine, Random Forest* dan *Ensemble Machine Learning Classifier (Stacking* dan *Voting). Data split*-nya akan menggunakan perbandingan antara 70:30 dan 80:20 karena secara empiris merupakan *range* untuk *data splitting* terbaik (Gholamy, Kreinovich and Kosheleva, 2018). Selain performa algoritma, penelitian juga melihat pengaruh penggunaan SMOTE terhadap dataset yang ada. Tabel 3.1 merupakan beberapa model yang akan dilakukan pada Penelitian ini yaitu:

Tabel 3.1 Daftar Model Penelitian

No.	Model	Keterangan Model
1	**Model A**	Algoritma Ensemble Stacking tanpa SMOTE, data split 70:30
2	**Model B**	Algoritma Ensemble Stacking tanpa SMOTE, data split 80:20
3	**Model C**	Algoritma Ensemble Stacking dengan SMOTE, data split 70:30
4	**Model D**	Algoritma Ensemble Stacking dengan SMOTE, data split 80:20
5	**Model E**	Algoritma Ensemble Voting tanpa SMOTE, data split 70:30
6	**Model F**	Algoritma Ensemble Voting tanpa SMOTE, data split 80:20
7	**Model G**	Algoritma Ensemble Voting dengan SMOTE, data split 70:30
8	**Model H**	Algoritma Ensemble Voting dengan SMOTE, data split 80:20

3.2.8. Evaluation

Pada Tahap ini peneliti akan mengevaluasi kualitas model berdasarkan akurasi yang didapatkan pada tahapan sebelumnya. Fase ini bertujuan untuk menjawab pertanyaan, “apakah model yang didapatkan benar-benar menjawab pertanyaan awal atau perlu disesuaikan kembali?” Model dengan akurasi tertinggi akan disimpan dan digunakan untuk tahapan selanjutnya. Pada tahapan ini peneliti akan mengevaluasi model terbaik dengan menggunakan *confusion*

matrix dengan beberapa penilaian diantaranya akurasi, presisi, *recall* dan *f1-score* yang didapatkan dari *confusion matrix* model terbaik.

3.2.9. Deployment

Pada tahapan ini peneliti mengimplementasikan model terbaik yang didapatkan di tahap sebelumnya terhadap data yang belum memiliki label (data yang telah melewati fase *data preparation*). Pada tahap ini memiliki *output* berupa hasil prediksi dan analisis terhadap semua data yang telah diproses di tahapan sebelumnya.

3.2.10. Feedback

Pada tahapan ini peneliti menarik kesimpulan yang didapatkan dari tahapan sebelumnya. Pada tahap ini peneliti memberikan *feedback* berdasarkan tahapan sebelumnya dan juga *wordcloud* dari setiap SDGs. *Wordcloud* di sini akan diambil berdasarkan hasil SDGs di tahapan *data preparation*. Berdasarkan *wordcloud* tersebut, peneliti dapat memberikan *feedback* terhadap topik yang ramai menjadi perbincangan pengguna internet di Indonesia.