

BAB 3 METODOLOGI PENELITIAN

3.1 Metodologi Penelitian

Dalam penelitian Analisis Sentimen Program Naturalisasi Pemain Timnas Sepak Bola Indonesia Menggunakan Algoritma *Naive Bayes* memiliki beberapa tahapan penelitian dalam perancangannya yaitu:

1. Studi Literatur

Studi literatur dilakukan untuk memahami teori dan penelitian terkait yang relevan dengan penelitian ini. Topik yang dikaji meliputi Twitter, pemain keturunan timnas Indonesia, analisis sentimen, dan metode *Naive Bayes*.

2. Pengumpulan Data

Data yang dikumpulkan berasal dari media sosial Twitter. Data yang dikumpulkan menggunakan menggunakan teknik *crawling* menggunakan *library* yang bernama *tweet-harvest*. *Library* ini dapat mengambil *tweet* dengan cara mencari kata kunci yang diinginkan, yaitu kata kunci terkait seputar program naturalisasi pemain timnas Indonesia.

3. Perancangan

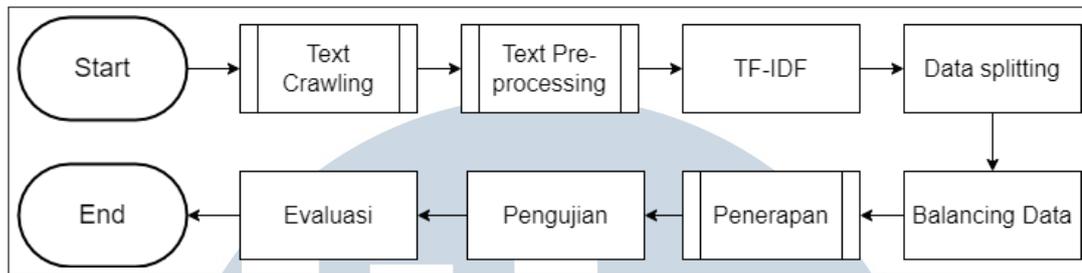
Perancangan sistem dapat dilakukan setelah data yang diinginkan dan sesuai telah terkumpul. Data yang sudah terkumpul akan dilakukan tahap *pre-processing* sebagai pembersihan pada data.

4. Pengujian dan Evaluasi

Tahap pengujian dilakukan setelah model menyelesaikan pelatihan. Pada tahap ini, model digunakan untuk mengklasifikasikan sentimen pada data uji dan hasilnya dievaluasi dengan sistem *Confusion Matrix* untuk mengetahui performa, akurasi, dan hasil pengujian.

3.2 Rancangan Sistem

Penelitian ini menggunakan beberapa tahapan perancangan, yang dijelaskan pada bagian ini. Untuk memvisualisasikan alur rancangan sistem, digunakan *flowchart* seperti yang ditampilkan pada Gambar 3.1



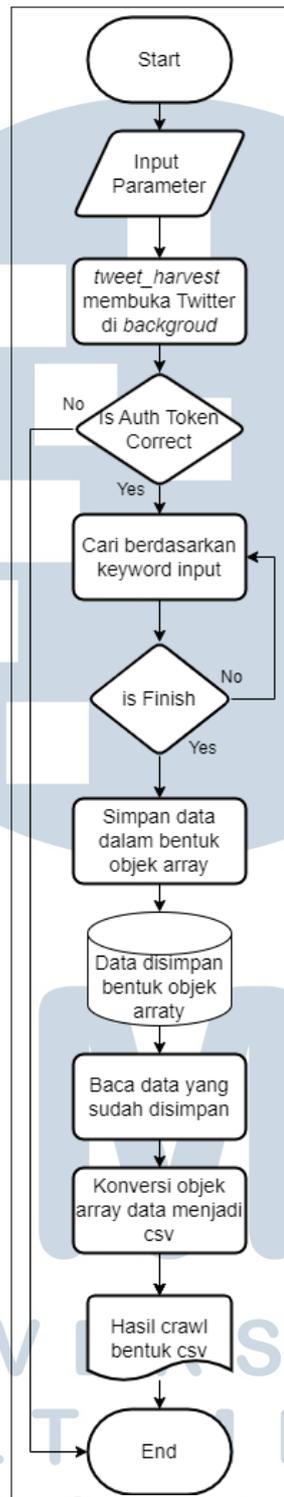
Gambar 3.1. *Flowchart* Rancangan Sistem

Gambar 3.1 merupakan *flowchart* rancangan sistem yang dimulai pada tahap *Text Crawling*. Setelah data didapat, dilakukan *pre-processing* yaitu pembersihan dan pengolahan data. Tahapan selanjutnya yakni penerapan TF-IDF, *data splitting* dan *balancing data* sebelum dilakukan penerapan Algoritma. Setelah diterapkan, dilakukan pengujian dan evaluasi terhadap model yang telah diterapkan.

3.2.1 Text Crawling

Data yang didapat dalam penelitian analisis sentimen ini diambil dan dikumpulkan menggunakan menggunakan teknik *crawling* menggunakan *library* yang bernama *tweet-harvest*. *Library* ini dapat mengambil *tweet* dengan cara mencari kata kunci yang diinginkan. Pada penelitian ini, data berfokus terhadap pemain naturalisasi Timnas Indonesia. Oleh karena itu, *keyword* acuan *Text crawling* penelitian ini adalah "Naturalisasi Pemain Timnas Indonesia" karena berbahasa Indonesia dan memperbanyak data yang berlokasi di Indonesia. *Keyword* lainnya yang digunakan seperti "Naturalisasi Pemain Timnas" karena berhubungan dengan topik penelitian analisis sentimen

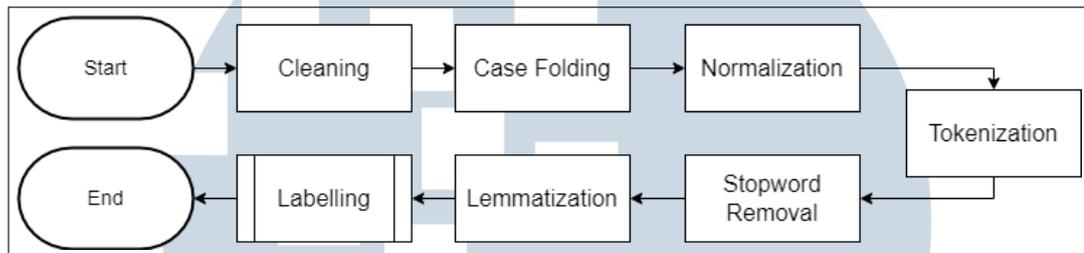
Pada Gambar 3.2, parameter input diisi dengan *keyword*, *auth_token* *Twitter*, jangkauan tanggal, nama *file* untuk disimpan dan limit dari banyaknya *tweet* yang diambil. Setelah input sudah di isi program akan membuka *browser*. Program akan mengecek apakah *auth_token* yang diisi benar dan jika benar program mulai mengambil data berdasarkan *keyword* sampai selesai. Jika sudah selesai data yang didapat akan disimpan dalam bentuk *csv*.



Gambar 3.2. *Flowchart Text Crawling*

3.2.2 Text Pre-processing

Pada tahap *text pre-processing*, data yang sudah didapat akan diolah dan dan dibersihkan. Tahap ini mencakup alur langkah-langkah dalam bentuk *flowchart* seperti yang ditampilkan pada Gambar 3.3.



Gambar 3.3. *Flowchart Text Pre-processing*

1. *Cleaning*

Tahap *cleaning* merupakan proses membersihkan karakter dari simbol selain huruf alfabet.

2. *Case Folding*

Tahap *case folding* merupakan salah satu tahap penting dalam pengolahan data teks, dimana seluruh karakter dalam kalimat diubah menjadi huruf kapital ataupun huruf kecil.

3. *Normalization*

Tahap *normalization* merupakan proses perubahan kata singkat yang tidak baku menjadi kata dasar yang baku.

4. *Tokenization*

Tahap *tokenization* merupakan proses pemenggalan kata per kalimat.

5. *Stopword Removal*

Tahap *stopword removal* merupakan penghapusan kata yang tidak penting dalam kalimat.

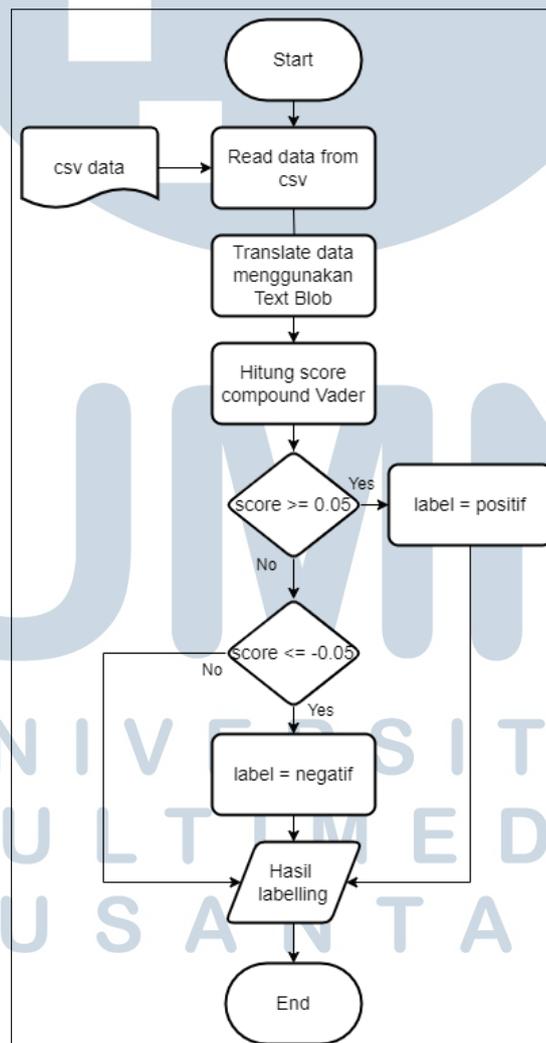
6. *Lemmatization*

Tahap *lemmatization* adalah proses untuk mengubah kata berimbuhan menjadi bentuk kosakatanya.

7. Labeling

Tahap *labelling* merupakan proses memberikan label positif, negatif ataupun netral pada suatu kalimat.

Pada tahap *labelling*, *library* yang digunakan dalam *labelling* yaitu *Valence Aware Dictionary and Sentimen Reasoner (VADER)*. VADER adalah alat bantu analisis sentimen yang menggunakan aturan dan kamus sentimen untuk memahami sentimen dalam teks. Kamus sentimen ini berisi daftar kata-kata yang diklasifikasikan sebagai positif, negatif, atau netral. VADER mengklasifikasikan teks berdasarkan skor compound: teks dengan skor ≥ 0.05 dikategorikan sebagai positif, teks dengan skor ≤ -0.05 dikategorikan sebagai negatif. *Flowchart labelling* dapat dilihat pada Gambar 3.4.



Gambar 3.4. *Flowchart Labelling*

Gambar 3.4 menampilkan diagram *flowchart* proses *labelling*. Proses dimulai dengan membaca data hasil *crawling*, kemudian menerjemahkan data karena VADER terbatas pada bahasa Inggris. Penerjemahan dilakukan menggunakan *library Text Blob*. Setelah data diterjemahkan, skor *compound* dihitung menggunakan VADER. Teks dengan skor ≥ 0.05 dikategorikan sebagai positif, teks dengan skor ≤ -0.05 dikategorikan sebagai negatif.

Selain memanfaatkan VADER, sebuah alat analisis sentimen yang populer, proses pelabelan sentimen juga dapat dilakukan dengan memanfaatkan *pre-trained model*. Model-model ini, yang telah dilatih sebelumnya pada kumpulan data teks yang sangat besar, mampu menangkap nuansa bahasa yang lebih kompleks dibandingkan dengan kamus sentimen seperti VADER. Dengan demikian, *pre-trained model* dapat memberikan hasil yang lebih akurat, terutama dalam menangani bahasa sehari-hari yang informal atau bahasa gaul. Selain itu, model-model ini juga lebih fleksibel dan dapat disesuaikan untuk berbagai domain dan tugas yang spesifik. *Pre-trained model* yang digunakan yaitu *mdhugol/indonesia-bert-sentiment-classification*, model yang dikhususkan dalam klasifikasi sentimen ini dapat melakukan klasifikasi sentimen .

3.2.3 TF-IDF

Dalam tahap TF-IDF (*Term Frequency-Inverse Document Frequency*), bobot akan diberikan kepada setiap kata dalam data. Bobot ini mencerminkan seberapa penting suatu kata dalam dokumen tersebut dan seberapa jarang kata tersebut muncul dalam keseluruhan kumpulan dokumen. Bobot TF-IDF dihitung dengan mengalikan dua nilai: TF (*Term Frequency*) dan IDF (*Inverse Document Frequency*).

3.2.4 Data Splitting

Sebagai bagian dari evaluasi, dataset dibagi menjadi dua set yang tidak tumpang tindih. Set pertama, yaitu data pelatihan, digunakan untuk melatih model Multinomial Naive Bayes dan Complement Naive Bayes. Set kedua, yang terdiri dari 20% dari total data, berfungsi sebagai data pengujian untuk mengukur kemampuan generalisasi model.

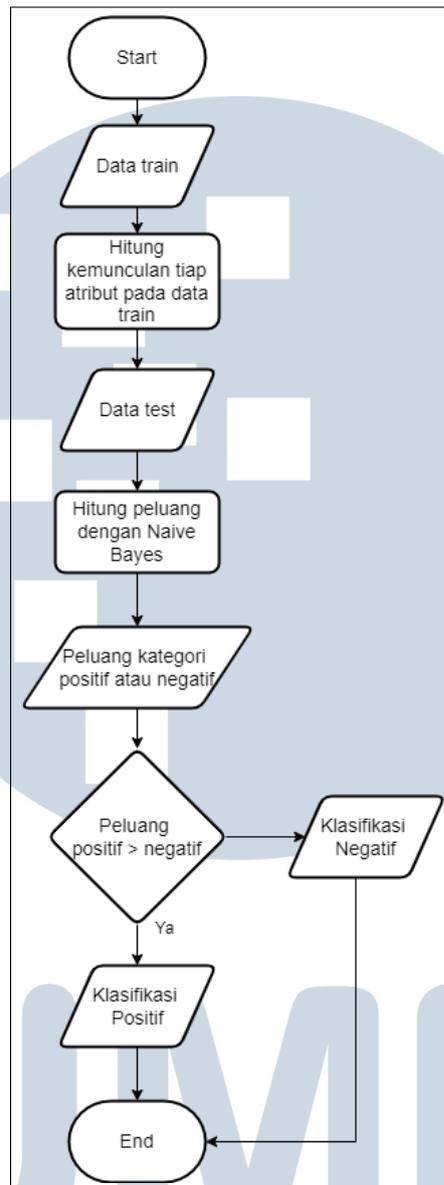
3.2.5 Data Balancing

Setelah data dibagi menjadi data *train* dan *test*, langkah selanjutnya adalah menyeimbangkan data pelatihan. Hal ini penting untuk dilakukan karena data pelatihan seringkali memiliki ketidakseimbangan kelas, dimana satu kelas (kelas minoritas) memiliki jumlah data yang jauh lebih sedikit daripada kelas lainnya (kelas mayoritas). Ketidakseimbangan kelas ini dapat menyebabkan model pembelajaran mesin menjadi bias terhadap kelas mayoritas dan berkinerja buruk pada kelas minoritas. Metode yang digunakan pada penyeimbangan data adalah *Adaptive Synthetic* (ADASYN). ADASYN adalah teknik *oversampling* yang menghasilkan data sintesis untuk kelas minoritas.

3.2.6 Penerapan Model

Pada tahap implementasi, model klasifikasi *Multinomial Naive Bayes* dan *Complement Naive Bayes* dipilih untuk digunakan dalam penelitian ini. Kedua model ini dilatih menggunakan data latih yang telah melalui proses pembobotan TF-IDF. Dengan memberikan bobot yang lebih tinggi pada kata-kata yang jarang muncul, diharapkan model dapat lebih baik dalam membedakan dokumen dari kelas yang berbeda. Setelah proses pelatihan, kinerja kedua model dievaluasi menggunakan data uji dengan menghitung berbagai metrik kinerja. Pemilihan model ini didasarkan pada kesederhanaan implementasi, kecepatan komputasi, serta hasil penelitian sebelumnya yang menunjukkan kinerja yang baik pada masalah klasifikasi teks.

Gambar 3.5 menjelaskan bagaimana model bekerja. Model ini menghitung berapa kali kata-kata tertentu muncul pada data *train*. Hasil perhitungan ini kemudian digunakan untuk menebak kategori dari data baru. Untuk menguji ketepatan tebakan model, kita menggunakan data yang belum pernah dilihat model sebelumnya atau data *test*. Model ini menggunakan rumus khusus (teorema Bayes) untuk menghitung kemungkinan data baru masuk ke kategori positif atau negatif. Jika kemungkinan positif lebih besar, maka data dikategorikan sebagai positif, dan sebaliknya.



Gambar 3.5. Flowchart Model Naive Bayes

Untuk mendapatkan kinerja optimal dalam model klasifikasi, proses *Hyperparameter Tuning* menjadi langkah yang krusial. Proses ini bertujuan untuk mencari kombinasi nilai *hyperparameter* terbaik yang dapat memaksimalkan kinerja model pada data yang belum terlihat. Dengan kata lain, *Hyperparameter Tuning* membantu menemukan pengaturan model yang paling sesuai dengan karakteristik data.

3.2.7 Pengujian Model

Setelah tahap pelatihan, kedua model dievaluasi menggunakan dataset uji yang independen. *Confusion matrix* digunakan sebagai alat evaluasi untuk menghitung metrik kinerja seperti *precision*, *recall*, dan *F1-score* pada masing-masing kelas (positif dan negatif). Selain itu, dilakukan pula pengujian generalisasi model dengan menggunakan data baru yang belum pernah digunakan dalam proses pelatihan maupun pengujian.

3.2.8 Evaluasi

Evaluasi kinerja merupakan tahap akhir dalam penelitian ini. Tujuan utama dari tahap evaluasi adalah untuk mengukur performa model *Naive Bayes* yang telah dikembangkan. Melalui proses evaluasi, diharapkan dapat diperoleh pemahaman yang lebih mendalam mengenai kekuatan dan kelemahan model dalam menyelesaikan masalah klasifikasi yang diteliti.

