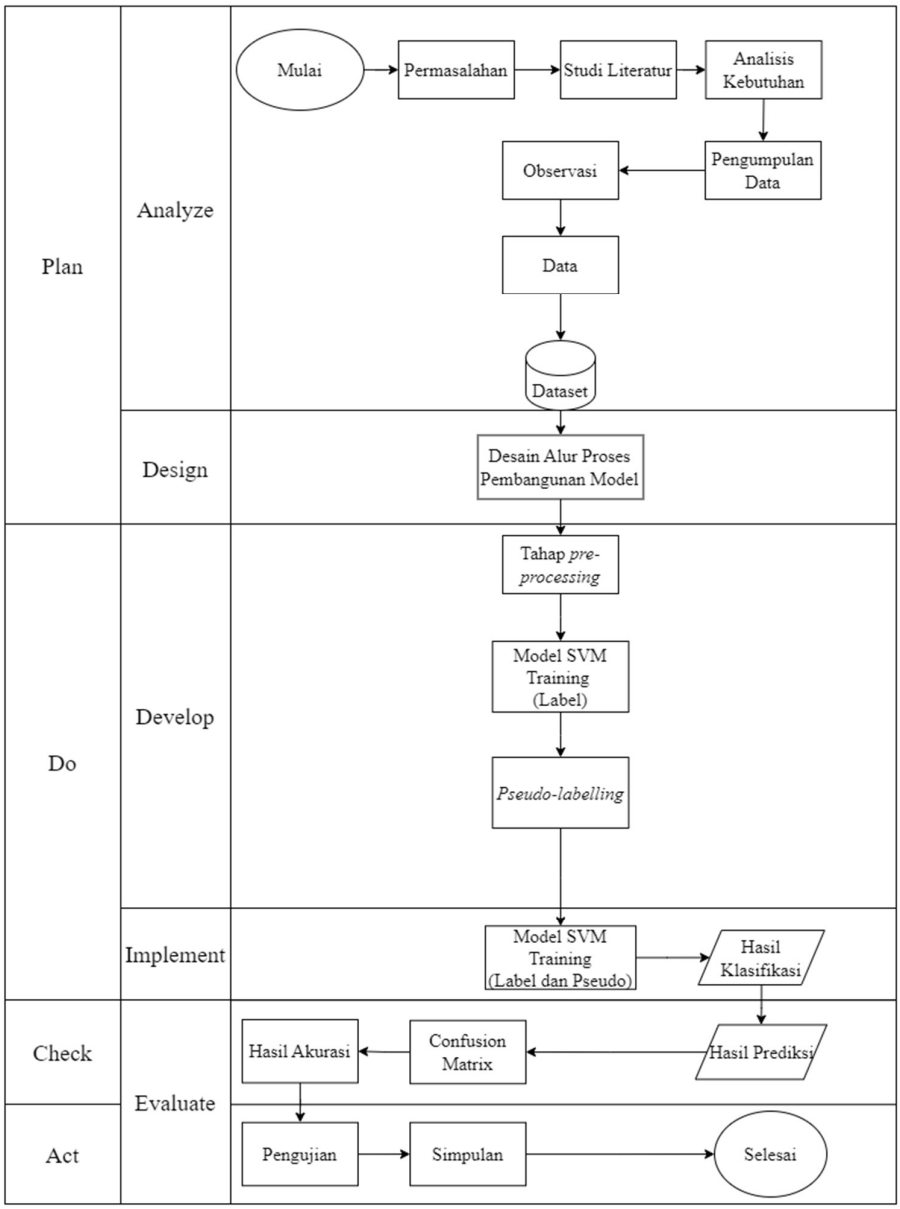
BAB III
METODOLOGI PENELITIAN

Metodologi penelitian ini menggunakan Model ADDIE. Menurut Zahid, (dalam Dwitiyanti, N., dkk., 2020) model ADDIE merupakan model yang umum digunakan dalam pengembangan desain instruksional, namun secara substansial dapat digunakan dalam pengembangan media pembelajaran. Model ADDIE tersebut dibuat oleh *Centre for Educational Technology* di *Florida State University*.



Gambar 3.1 Alur Penelitian Menggunakan Model ADDIE

3.1. Metode dan Desain Penelitian

Seperti yang dapat dilihat pada gambar 3.1 alur penelitian Model ADDIE terdiri dari 5 fase dimana fase-fase tersebut adalah analisis (*analyze),* desain (*design),* mengembangkan (*develop*), implementasi (*implement*), dan evaluasi (*evaluate*).

3.1.1. Analyze

Fase ini merupakan fondasi dari fase-fase lainnya, fase ini merupakan tahapan penelitian dimana dilakukan pendefinisian masalah, dilakukan kajian riset, dan menentukan kebutuhan-kebutuhan yang akan digunakan pada penelitian. Pada fase ini dilakukan studi literatur berdasarkan permasalahan yang telah diangkat. Berdasarkan permasalahan yang telah diangkat dilakukan pemilihan algoritma klasifikasi, dimana algoritma klasifikasi yang dipilih adalah algoritma *Support Vector Machine* dan dengan pendekatan pembelajaran mesin *semi-supervised*.

Pada tahap ini pula dilakukan pemilihan objek penelitian, dimana objek penelitian yang dipilih adalah terhadap akun pengguna *Twitter* asli dan akun bot sosial *Twitter* pada post *tweet* yang berhubungan dengan topik platform e-commerce. Terakhir dilakukan pengumpulan dan observasi data, dimana data-data tersebut dijadikan dataset yang dapat digunakan dalam pembuatan model klasifikasi dan prediksi.

Pengumpulan data dilakukan dengan proses *data crawling/data scraping Twitter*, dengan menggunakan alat yang bernama *Tweet-Harvest,* dimana dengan alat tersebut data-data post *tweet* pengguna yang berhubung dengan kata kunci yang dimasukkan dapat dikumpulkan. Data-data yang dikumpulkan tersebut bersifat tidak berlabel, dimana label tersebut menunjukkan apakah sebuah postingan pengguna berasal dari pengguna bot sosial atau pengguna yang asli. Data-data yang dikumpulkan tersebut adalah data *tweet/post trending topics* yang berhubungan dengan platform *e-commerce* yang populer. Platform *e-commerce* yang populer tersebut berdasarkan situs *e-commerce* yang populer dikunjungi, dimana situs *e-commerce* tersebut adalah Amazon, eBay, dan AliExpress berdasarkan laporan statistik *online marketplace* yang paling banyak dikunjungi di seluruh dunia pada April 2023, berdasarkan lalu lintas bulanan (Statista Search Department, 2021) .

Pada penelitian ini, metode *semi-supervised learning* digunakan, dengan itu, selain dikumpulkan data yang tidak berlabel, dikumpulkan juga data yang berlabel.

Dataset yang berlabel ini berasal dari dataset *cresci-2017* (Cresci, S., dkk., 2017) dari *botometer* sebuah *repository* untuk berbagi kumpulan dataset bot sosial Twitter yang diberi anotasi, dan juga menyediakan daftar alat yang tersedia untuk mendeteksi bot. Dataset *cresci-2017* merupakan kumpulan dataset yang berisi data akun dan post *twitter* yang berasal dari pengguna asli dan berbagai bot sosial. Data-data dalam dataset *cresci-2017* tersebut berformat *csv* dan terdapat berbagai macam fitur-fitur yang dapat digunakan untuk melatih model klasifikasi dalam penelitian yang dilakukan.

3.1.2. Design

Fase ini merupakan fase dimana hasil atau *output* dari fase analisis digunakan untuk melakukan proses perancangan dan desain model. Pada tahap ini dijelaskan alur proses dari pembangunan model klasifikasi menggunakan sebuah diagram model. Dimulai dari pengumpulan data, lalu tahap *pre-processing, feature extraction & engineering,* pelatihan model pertama, klasifikasi, *pseudo-labelling,* pelatihan model lebih lanjut, dan klasifikasi terakhir.

3.1.3. Develop

Pada fase ini dataset yang telah dikumpulkan selanjutnya melalui tahap *pre-processing.* Pada tahap *pre-processing* ini terbagi menjadi empat tahap yaitu *data cleaning* dimana data mentahan dilakukan pembersihan dengan beberapa proses seperti pengisian/penambahan data kosong, menyelesaikan inkonsistensi data, menghaluskan *noise data,* dan lain-lainnya*,* selanjutnya dilakukan *data integration* dimana data yang memiliki format berbeda diubah menjadi format yang sama sehingga dapat digunakan*,* lalu dilakukan *data transformation* dimana dilakukan normalisasi dan generalisasi data*,* dan terakhir *data reduction* dimana dilakukan pengurangan data.

Setelah tahap *preprocessing,* data yang telah melalui proses *preprocessing* dilakukan proses *feature extraction* dimana fitur-fitur data yang telah di-*preprocessed* diekstraksi agar bisa digunakan untuk melatih model untuk melakukan klasifikasi.

Setelah proses *feature extraction* dilanjutkan dengan proses *data mining*. Dataset berlabel dilakukan pemisahan menjadi data latih dan data uji. Selanjutnya, data latih yang sebelumnya telah dipisah dari dataset berlabel, akan digunakan untuk melatih model SVM pertama. Model SVM tersebut selanjutnya akan

melakukan prediksi terhadap data uji yang sebelumnya telah dipisahkan dari dataset berlabel. Hasil dari prediksi tersebut selanjutnya digunakan untuk menguji nilai akurasi dari model.

Tahap berikutnya, setelah model telah dilatih, model tersebut akan melakukan prediksi terhadap dataset tak berlabel, dengan tujuan untuk melakukan proses *pseudo-labelling*. Setelah *pseudo-label* telah diprediksi, data *pseudo-label* tersebut digabungkan dengan data berlabel sebelumnya menjadi sebuah dataset baru yang berisi data berlabel dan data *pseudo-labeled*. Data baru yang didapatkan tersebut digunakan untuk melatih model terakhir kalinya.

3.1.4. Implement

Selanjutnya, pada tahap ini, model klasifikasi yaitu model klasifikasi SVM yang dilatih dengan dataset berlabel dan model klasifikasi SVM yang dilatih dengan gabungan data latih berlabel dan data *pseudo-labeled* akan dilatih. Lalu tiap model akan membuat prediksi yang hasilnya akan diukur dengan menggunakan *confusion matrix*.

3.1.5. Evaluate

Pada tahap ini, akan dilakukan pengujian kinerja terhadap model klasifikasi SVM yang dilatih dengan data berlabel dan juga terhadap model klasifikasi SVM yang dilatih dengan dataset *pseudo* yang didapat dengan menggabungkan data berlabel dan data *pseudo-labeled*. Model klasifikasi tersebut akan dilakukan pengujian terhadap variasi rasio pemisahan data dan dengan pengaturan *hyperparameter C* dari algoritma SVM.

Menurut Muraina, I. (2022) pemisahan dataset adalah praktik yang dianggap sangat diperlukan untuk menghilangkan atau mengurangi bias pada data pelatihan dalam Model Pembelajaran Mesin. Proses pemisahan ini dilakukan untuk mencegah algoritme pembelajaran mesin menghasilkan tipe *overfitting* yang dapat berkinerja buruk pada data pengujian sebenarnya.

Untuk pengujian variasi rasio pemisahan data, dilakukan pengujian rasio pemisahan pelatihan/pengujian: 50-50, 60-40, 70-30, 80-20, dan 90-10. Rasio pemisahan tersebut berdasarkan penelitian Muraina, I., (2022) yang mempelajari dampak rasio pemisahan pelatihan/pengujian yang berbeda terhadap performa model. Dimana rasio pemisahan pelatihan/pengujian tersebut merupakan rasio yang digunakan dalam penelitian mereka.

Untuk pengujian dengan pengaturan *hyperparameter C* dalam algoritma SVM digunakan nilai *C*= 0.001, *C*= 0.01, *C*= 0.1, *C*= 1, dan *C*= 10, *C*= 100. Nilai-nilai tersebut berdasarkan penelitian oleh Tharwat, A. (2019), yang melakukan penelitian investigasi terhadap parameter penglasifikasi *Support Vector Machine* dengan fungsi kernel. Penelitian yang dilakukan oleh Tharwat, A. (2019) menemukan bahwa nilai *C* yang kecil meningkatkan margin SVM dan karena hal tersebut meningkatkan jumlah vektor pendukung yang dapat menyebabkan underfitting. Sedangkan untuk nilai *C* yang besar meminimalkan lebar margin SVM dan menambah bobot sampel yang tidak dapat dipisahkan. Oleh karena itu, satu *outlier*, *noise*, atau sampel kritis dapat menentukan batas keputusan, yang membuat pengklasifikasi sensitif terhadap *noise* dalam data.

Setelah pengujian tersebut hasil akan diukur dengan *Confusion Matrix* untuk dapat diambilkan kesimpulan. *Confusion Matrix* adalah pengukur yang sangat populer digunakan saat memecahkan masalah klasifikasi. *Confusion Matrix* dapat diterapkan untuk klasifikasi biner serta untuk masalah klasifikasi multi kelas (Kulkarni, A., dkk., 2020). Sebuah *Confusion Matrix* atau *Error Matrix* adalah suatu layout tabel yang memungkinkan untuk memvisualisasikan atau menghitung performa sebuah algoritma, khususnya dalam klasifikasi statistika. *Confusion Matrix* sering digunakan dalam berbagai macam permasalahan klasifikasi.

Tabel 3.1

Confusion Matrix

		Actual Values	
		Positive (1)	**Negative (0)**
Predicted Values	**Positive (1)**	**TP**	**FP**
	Negative (0)	**FN**	**TN**

Seperti yang dapat dilihat pada tabel 3.1, tabel *Confusion Matrix* adalah tabel dengan 4 kombinasi berbeda dari nilai prediksi dan aktual. Berikut merupakan penjelasan dari tabel 3.1:

a. TP - *True Positive*

   Diprediksi positif dan benar

b. TN - *True Negative*

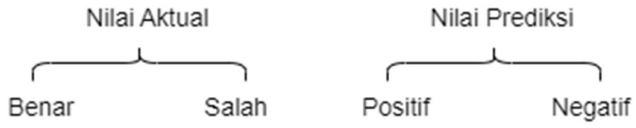
   Diprediksi negatif dan benar

c. FP - *False Positive*

   Diprediksi positif dan salah

d. FN - *False Negative*

   Diprediksi negatif dan salah



Gambar 3.2 Nilai Prediksi dan Nilai Aktual

Seperti yang dapat dilihat pada gambar 3.2, perlu diketahui bahwa untuk nilai prediksi dijelaskan sebagai Positif dan Negatif sedangkan nilai aktual dijelaskan sebagai Benar dan Salah.

Confusion Matrix sangat berguna untuk mengukur Accuracy, Precision, Recall, F1 Score. Berikut penjelasan dari hal-hal tersebut:

a. Accuracy

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} \quad (1)$$

Rumus 3.1 Accuracy

Accuracy mewakili jumlah instance data yang diklasifikasikan dengan benar di atas jumlah total instance data.

b. Precision

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Rumus 3.2 Precision

Precision adalah rasio pengamatan positif yang diprediksi dengan benar terhadap total pengamatan positif yang diprediksi.

c. Recall

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

Rumus 3.3 Recall

Recall adalah rasio pengamatan positif yang diprediksi dengan benar terhadap semua pengamatan di kelas yang sebenarnya.

d. F1 Score

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (1)$$

Rumus 3.4 F1 Score

F1 Score adalah rata-rata tertimbang dari Precision dan Recall. Oleh karena itu, skor ini memperhitungkan positif palsu dan negatif palsu. Hasil-hasil yang didapatkan selanjutnya akan dilakukan analisis. Terakhir dibuatkan sebuah kesimpulan dari hasil penelitian.

3.2. Populasi dan Sampel

3.2.1. Populasi

Populasi adalah wilayah generalisasi yang terdiri atas objek/subjek yang memiliki kuantitas dan karakteristik tertentu yang ditetapkan oleh peneliti untuk dipelajari dan kemudian ditarik kesimpulannya. Populasi dalam penelitian yang dilakukan adalah akun pengguna *Twitter* yang berlokasi secara global, berbahasa Inggris dan berjumlah 1.421 post.

Alasan penelitian memilih akun dan post *Twitter (X)* yang berlokasi secara global dan berbahasa Inggris adalah karena studi kasus dari penelitian ini memfokuskan kepada topik platform *e-commerce* yang populer secara global dan platform *e-commerce* tersebut mayoritas menggunakan bahasa Inggris.

3.2.2. Sampel

Sampel adalah suatu bagian dari keseluruhan serta karakteristik yang dimiliki oleh sebuah Populasi. Dalam penentuan sampel dari populasi yang dilakukan penelitian, Pada penelitian ini, sampel akan diambil dengan menggunakan metode *simple random sampling.* Menurut Singh dalam Noor (2022), *Simple Random Sampling* adalah "metode pemilihan sampel yang paling sederhana dan paling umum, di mana sampel dipilih unit demi unit, dengan probabilitas pemilihan yang sama untuk setiap unit pada setiap pengundian".

Dari sampel yang diambil tersebut akan dilakukan sebuah pemisahan, yaitu pemisahan menjadi *train* dan *test split.* Pemisahan tersebut bertujuan untuk membantu dalam pengukuran performa model klasifikasi yang akan dilatih.

3.3. Instrumen Penelitian

Instrumen penelitian merupakan alat yang digunakan untuk mengumpulkan data dalam penelitian. Instrumen penelitian dibuat sesuai tujuan pengukuran dan teori yang digunakan (Purwanto, 2018). Instrumen penelitian merupakan hal yang penting bagi sebuah penelitian karena instrumen penelitian merupakan alat bantu untuk pengumpulan data yang akan digunakan untuk penelitian.

Dalam penelitian yang dilakukan instrumen penelitian yang digunakan merupakan API *Twitter* dan alat *Tweet-Harvest.* API menyediakan kondisi kemungkinan untuk berbagi konten dan data secara online. Sebagai objek protokolologis, API memungkinkan pihak yang berkepentingan untuk mengakses data dan fungsionalitas layanan online populer, dengan cara yang sangat terkontrol (Trupthi, M. dkk., 2017). Dengan menggunakan API *Twitter,* proses pengumpulan data *Twitter* seperti *data streaming* dan *data crawling* dapat dilakukan.

Pada penelitian ini alat yang digunakan untuk melakukan *Twitter Data Crawling* adalah *Tweet-Harvest.* Dengan *Tweet-Harvest, Twitter Data Crawling* dapat dilakukan dengan menggunakan *auth_token* dari akun twitter pengguna yang dapat digunakan untuk mengumpulkan *post/tweet* dari kata kunci yang dimasukkan ke *Tweet-Harvest.* Setelah dilakukan proses *data crawling Twitter*, data yang terkumpul akan dikumpulkan menjadi sebuah dataset yang akan digunakan dalam penelitian. Dataset yang telah dikumpulkan dengan *Tweet-Harvest* sebelumnya bersifat tidak berlabel. Untuk dataset yang bersifat berlabel digunakan dataset dari *cresci-2017* (Cresci, S., dkk., 2017*)* dari *botometer* yang berisi data-data bot sosial dan pengguna asli *twitter* dalam bentuk format *csv*.