BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Pelecehan seksual merupakan tindakan seksual yang dilakukan secara sengaja dan adanya indikasi pemaksaan terhadap korban yang menolak [1]. Gambaran umum objek penelitian berfokus pada analisis sentimen komentar pada unggahan yang berkaitan dengan pelecehan seksual di UMN yang diperoleh melalui media sosial dan situs seperti Instagram, Twitter, Line Today, dan Medium.

Untuk mendapatkan data tersebut digunakan teknik *scraping* yang mengambil data untuk *testing* UMN dari 25 Juni 2021 hingga 7 Desember 2021 dan *data training* yang dimulai yang disimpan berupa *file Excel*. Dari hasil prediksi sentimen akan dilakukan analisa yang mendalam terhadap data dan hasil prediksi untuk menjawab dari rumusan masalah dan tujuan pada penelitian ini.

3.2 Metode Penelitian

Dalam penelitian ini menggunakan metode teknik *data mining* untuk melakukan analisis sentimen pada komentar terhadap kasus pelecehan seksual di UMN. Untuk menentukan metode yang cocok digunakan dalam penelitian ini dapat di bandingkan dengan tiga *process model* terbaik dalam *data mining yang* mengacu pada penelitian Antonia Dåderman dengan berjudul *Evaluating Frameworks for Implementing Machine Learning in Signal Processing: A Comparative Study of CRISP-DM, SEMMA and KDD* [9].

Tabel 3. 1 Perbandingan Frameworks Data Mining

Frameworks	CRISP-DM	KDD	SEMMA
Phase	*Business Understanding*	*Pre-KDD*	-
	Data Understanding	*Selection*	*Sample*
		Pre-Processing	*Explore*
	Data Preparation	*Transformation*	*Modify*
	Modelling	*Data Mining*	*Model*

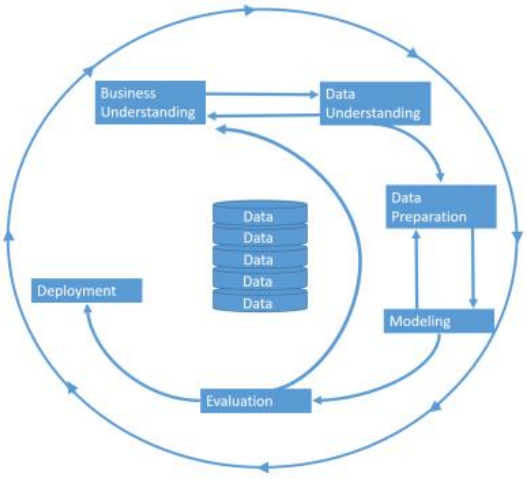
	Evaluation	*Interpretation / Evaluation*	*Assesment*
	Deployment	*Post-KDD*	-

Sumber: [9]

Dapat dilihat pada tabel 3.1 bahwa SEMMA tidak memiliki fase untuk memahami permasalahan atau menentukan tujuan dari proyek dan fase implementasi. Ketidak ada nya kedua fase tersebut tidak dapat di implementasi pada penelitian ini karena dalam penelitian ini membutuhkan kedua fase tersebut maka *framework* SEMMA tidak akan digunakan. Sekarang terdapat 2 pilihan yaitu CRISP-DM dan KDD. CRISP-DM memiliki standar *de-facto* dalam melakukan projek *data mining* [13]. Selain itu, CRISP-DM dapat bekerja baik di berbagai industri ditimbang KDD seperti industri teknologi informasi, *medicion*, dan *software* karena penelitian ini merupakan industri IT maka CRISP-DM pilihan yang tepat dan cocok untuk digunakan dalam penelitian ini serta CRISP-DM memiliki langkah dan terstruktur dengan baik [9].

3.3 Alur Penelitian

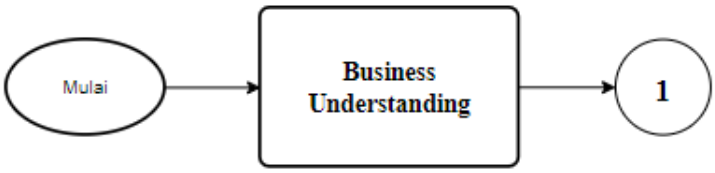
Pada alur penelitian ini, *framework* yang digunakan adalah CRISP-DM atau *Cross Industry Standard Process for Data mining* yang telah di jelaskan pada sub-bab 3.2. Pada gambar 3.1 merupakan alur yang telah diimplementasikan dan dimodifikasi sesuai dengan alur CRISP-DM yang asli dengan menggambarkan 6 tahapan utama yang terdiri dari berbagai *sub-process* dalam setiap tahapan [13].



Gambar 3. 1 Alur Penelitian Penerapan CRISP-DM [13].

3.3.1 Business Understanding

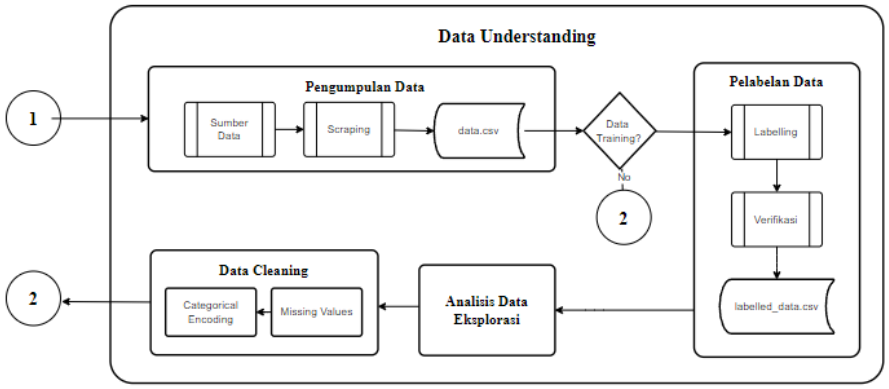
Tahap business understanding merupakan tahapan awal dari CRISP-DM. Tahapan ini untuk memahami dan menentukan target dari masalah pada kasus komentar pelecehan seksual yang terjadi terkait di UMN [13]. Target yang direncanakan dapat membuat model untuk memprediksi dan menganalisa dari sentimen komentar pelecehan seksual terkait UMN serta dapat menjawab masalah pokok pada penelitian ini.



Gambar 3.2 *Flowchart Business Understanding*

3.3.2 Data Understanding

Data understanding akan dibagi menjadi 4 *sub-process* dimana terdiri dari pengumpulan data, pelabelan data, analisis data eksplorasi, dan *data cleaning* yang terdapat pada gambar 3.4 [13].



Gambar 3.3 *Flowchart Data Understanding*

3.3.2.1 Pengumpulan Data

Pada tahap pengumpulan data terdapat dua kali proses pengumpulan data yang dimana data pertama dikumpulkan yang terkait dengan pelecehan seksual secara umum & tidak terkait UMN sedangkan data kedua mengumpulkan data komentar terkait pelecehan seksual yang terjadi di lingkungan UMN.

1) **Pengumpulan Data Komentar Terkait Pelecehan Seksual**

   Tahap awal yang dilakukan yaitu mencari sumber data yang berkaitan dengan pelecehan seksual secara umum dan tidak terkait dengan UMN. Sumber data akan didapatkan melalui situs Twitter dengan kata kunci "Pelecehan Seksual" dan "Pelecehan seksual kampus" yang diambil pada tanggal 1 Februari 2022 hingga 5 Februari 2022. Setelah sumber data berupa URL didapatkan kemudian dilakukan *web scraping* untuk melakukan ekstrasi data dan informasi dari URL yang diberikan. Hasil *scrape* akan di simpan dengan *format* CSV yang terdapat isi kolom berupa URL sumber data dan komentar.

2) **Pengumpulan Data Komentar Terkait Pelecehan Seksual di Lingkungan UMN**

   Pengambilan data yang terkait dengan UMN dimulai pada tanggal 25 Juni 2022 hingga 7 Desember 2022 kemudian mencari data yang terkait dengan pelecehan seksual di lingkungan UMN. Proses tahapan mirip seperti sebelumnya tetapi sumber data yang akan di ambil dari dari akun sumber utama yaitu ultimagz dan artikel yang membuat tentang kasus pelecehan seksual terkait dengan UMN yang menggunakan tagar #saatnyabicara yang terdapat di *platform* Instagram, artikel Medium, Line Today, dan Twitter. Data akan di ambil sebanyak mungkin berdasarkan ketersedian dari sumber data tersebut

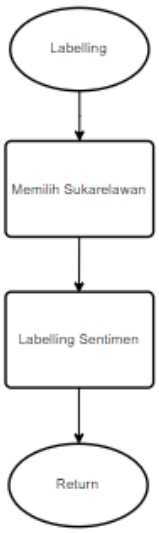
kemudian data yang terkumpul akan disimpan dalam bentuk CSV.

3.3.2.2 Pelabelan Data

Pada tahap pelabelan data terdapat dua *sub-process* yaitu *labelling*, verifikasi *label*:

1) Labelling

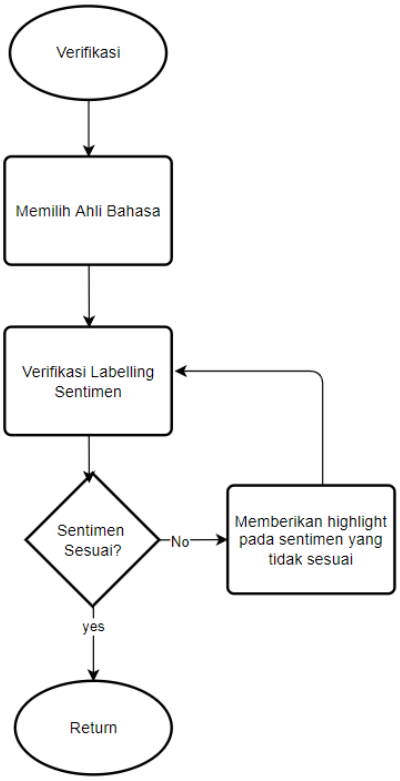
Pada tahap proses *labelling* akan memilih 2 mahasiswa untuk membantu melabelkan sentimen apa yang sesuai dengan komentar yang telah dikumpulkan. Kedua sukarelawan tersebut merupakan mahasiswa yang mendapatkan nilai A pada mata pelajaran bahasa Indonesia yang akan membantu untuk melabelkan sentimen yang tepat pada tiap komentar dari *total* data yang tersedia. Kedua mahasiswa melakukan pelabelan data yang dibagi dua atau 50/50 dari *total* data. Opsi yang diberikan untuk melabelkan sentimen terdapat tiga yaitu “positif”, “netral”, dan “negatif”.



Gambar 3.4 *Flowchart Labelling*

2) Verifikasi Label

Setelah melakukan tahap pelabelan selesai selanjutnya akan memilih ahli bahasa untuk melakukan verifikasi label yang telah di label oleh sukarelawan. Ahli bahasa yang akan melakukan verifikasi yaitu Niknik Mediyawati, S.Pd., M.Hum yang akan verifikasi dan membenarkan sentimen yang salah dengan memberikan 3 warna *highlight* seperti *Highlight* bertanda hijau seharusnya positif, yang merah seharusnya negatif, dan yang kuning seharusnya netral. Setelah itu, sentimen yang salah akan diperbaiki sesuai dengan warna *highlight* yang diberikan. Ahli bahasa akan verifikasi data dengan minimal sebanyak 100 data dan setelah itu akan dilakukan perbaikan label secara manual berdasarkan sentimen yang sesuai.



Gambar 3.5 *Flowchart* Verifikasi Label

3.3.2.3 Analisis Data Eksplorasi

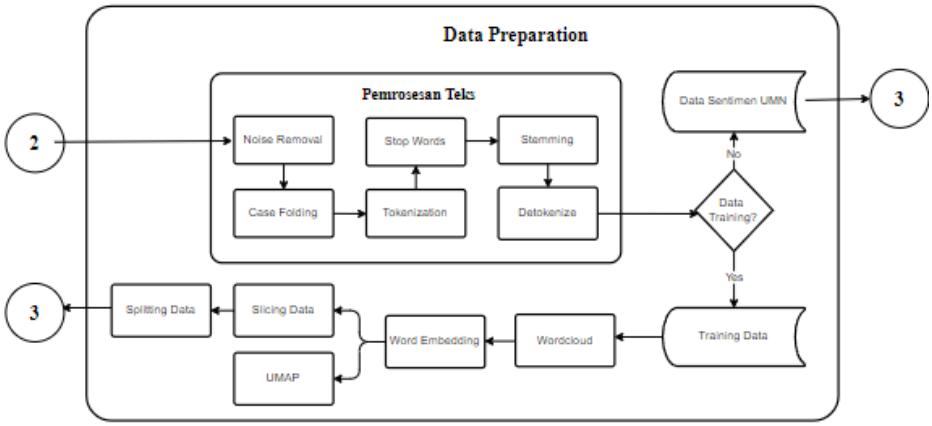
Pada tahap analisis data eksplorasi dimana akan melakukan beberapa visualisasi untuk dapat melihat informasi yang terdapat di dalam data seperti melihat jumlah data, ukuran data, statistik data, penyebaran distribusi data, dan cek *missing values*. Untuk melihat informasi tersebut bisa dilakukan visualisasi grafik pada informasi tertentu agar mempermudah melihat informasi seperti menggunakan *barplot, wordcloud, pie chart*, dll [35].

3.3.2.4 Data Cleaning

Setelah memahami bentuk dan kualitas data dari hasil analisis data eksplorasi selanjutnya akan memperbaiki kualitas data tersebut dengan cara *dropout data* jika terdapat *missing values* dan melakukan *label encoding* pada *dependant variable* atau *predictor* yang hasilnya akan berupa *polarity* seperti sentimen negatif akan menjadi *label* -1, sentimen netral akan menjadi *label* 0, dan sentimen positif akan menjadi *label* 1.

3.3.3 Data Preparation

Data Preparation akan dibagi menjadi 5 *sub-process* dimana terdiri dari pemrosesan teks, *wordcloud, word embedding*, UMAP, dan *splitting data* yang terdapat pada gambar 3.7 [13].



Gambar 3.6 *Flowchart Data Preparation*

3.3.3.1 Pemrosesan Teks

Pada tahap pemrosesan teks dimana untuk membersihkan agar dapat mudah dipahami oleh komputer. Proses pemrosesan data mencakup *Natural Language Processing* yang dimulai dari *noise removal, case folding, tokenization, stop words, stemming,* dan *detokenize.* Data yang telah di bersihkan selanjutnya disimpan kedalam bentuk CSV agar dapat langsung digunakan pada saat tahap selanjutnya atau sebagai *checkpoint.* Pada proses menyimpanan data terdapat kondisi jika data tersebut merupakan *data training* maka akan disimpan dengan bernama *training* data dan akan melanjutkan proses selanjutnya sedangkan jika bukan *data training* maka akan disimpan dengan nama data sentimen umn dan akan melewati proses yang terdapat di *data preparation* dan langsung ke proses ke *modelling.*

3.3.3.2 Wordcloud

Pada tahap pemrosesan data dimana untuk membersihkan agar dapat mudah dipahami oleh komputer. Proses pemrosesan data mencakup *Natural Language Processing* yang dimulai dari *noise removal, case folding, segmentation, tokenization, stop words,* dan *lemmatization.* Data yang telah di bersihkan selanjutnya disimpan

kedalam bentuk CSV agar dapat langsung digunakan pada saat tahap selanjutnya atau sebagai *checkpoint.*

3.3.3.3 Word Embedding

Pada tahap *word embedding* akan menggunakan metode *FastText* karena pada penelitian yang terkait, *FastText* memiliki akurasi *vectorize* yang cukup baik dan juga dapat melakukan vektorisasi dari kata yang belum pernah di temui sebelumnya dibandingkan metode *countvectorizer dan tfidf.* Pada penelitian "Perbandingan Kinerja *Word Embedding Word2vec, Glove,* Dan *FastText* Pada Klasifikasi Teks" menunjukkan bahwa setiap algoritma *word embedding* memiliki kinerja yang sama baiknya dan bergantung pada permasalahan serta bahasa yang digunakan. *FastText* memiliki keunggulan yaitu dapat vektorisasi suatu kata yang tidak ada *vocab* atau kamus sedangkan *word2vec* dan *glove* harus mempelajari tidak dapat vektorisasi kata yang tidak terdapat dalam kamus. Selain itu, hasil terbaik dari perbandingan ketiga jenis *word embedding* yaitu *FastText* [10]. *FastText* juga memiliki hasil yang cukup baik bahkan sama dengan *Word2Vec* di dataset yang berbahasa Indonesia [11]. Oleh sebab itu, jenis *word embedding* yang digunakan pada penelitian ini adalah *FastText*.

3.3.3.4 UMAP

Pada tahap UMAP dimana data yang telah di bersihkan akan di buat mirip dengan konsep *word of bags* terlebih dahulu yang berisi dari kumpulan kata-kata unik didalam sebuah *array* [17]. *Word of bags* yang berisi kata unik tersebut selanjutnya akan di *vectorize* menggunakan *menggunakan teknik Word Embedding* seperti penjelasan sub bab *word embedding.* Data yang telah di *vectorize* akan memiliki ukuran 300x300. Data yang berukuran besar akan dilakukan *dimensionality reduction* dengan menggunakan UMAP. Hasil pengurangan dimensi akan menjadi n baris dan 2 kolom yang terbagi

menjadi *axis* X dan Y sehingga data *word of bags* dapat di visualisasi berdasarkan *axis* X dan Y.

3.3.3.5 Slicing Data

Sebelum melakukan *splitting data*, akan dilakukan pengecekan *missing values* dari proses *word embedding* jika hasil *embedding* terdapat NaN atau Inf maka akan dilakukan *cleaning data* dengan cara *drop data* tersebut. Setelah itu, data akan di bagi menjadi X dan y yang dimana X adalah data independen dan y adalah *data predictor*.

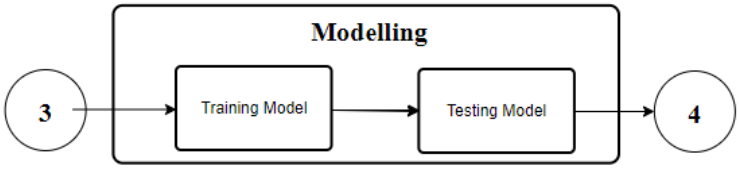
3.3.3.6 Splitting Data

Pada tahap *splitting data* akan terbagi jadi *training data* dan *testing*. Berdasarkan penelitian lainnya membuktikan bahwa dengan menggunakan berbagai algoritma *machine learning* terbukti rasio terbaik dan cocok untuk digunakan untuk *splitting dataset* yaitu rasio 70 banding 30 atau 70% *training data* dan 30% *testing data* [36]. Jadi dalam penelitian ini rasio *splitting data* yang akan digunakan sebesar 30% untuk *testing data* sedangkan *training data* sebesar 70%.

3.3.4 Modelling

Pada tahap *modelling* akan melakukan pelatihan model *machine learning* serta uji coba dari hasil model dengan *dataset* yang telah dipisah berdasarkan penjelasan pada subab 2.2 [13]. Teknik klasifikasi yang digunakan adalah teknik *supervised learning* karena pada penelitian terdahulu telah membandingkan hasil akurasi dari *supervised learning* yang menggunakan algoritma regresi linear, *decision tree*, dan SVM mendapatkan akurasi dengan rata-rata sebesar 82,33% sedangkan *unsupervised learning* dengan algoritma *K-Means*, *single linkage*, dan *apriori* mendapatkan hasil akurasi dengan rata-rata sebesar 78% [34].

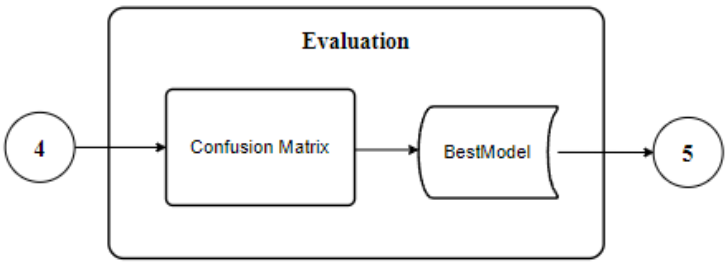
Setelah memilih teknik *machine learning* yang digunakan selanjutnya, akan memilih algoritma *supervised learning* yang akan digunakan. Dalam penelitian ini akan menggunakan algoritma SVM sebagai solusi dari permasalahan penelitian karena berdasarkan pada penelitian terdahulu dengan judul "*Comparation Of Classification Algorithm On Sentiment Analysis Of Online Learning Reviews And Distance Education"* menunjukkan bahwa hasil dari rata-rata akurasi SVM mendapatkan 87,67% yang lebih baik ditimbang *naïve bayes* dengan nilai rata-rata 86,33% yang menggunakan *dataset* berbahasa Indonesia [8]. Ditambah dengan jurnal pendukung lainnya yang menggunakan *dataset* berbahasa Indonesia menunjukkan bahwa rata-rata dari hasil penelitian SVM juga lebih baik ditimbang *Naïve Bayes* dengan selisih 4,42% [33].



Gambar 3.7 *Flowchart Modelling*

3.3.5 Evaluation

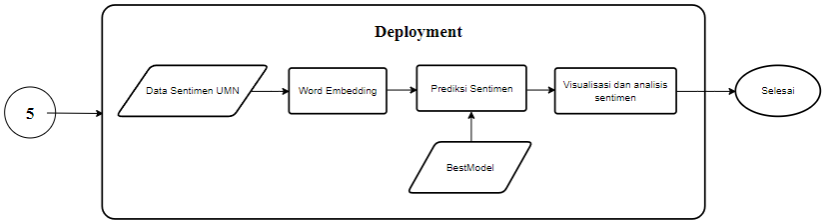
Pada tahap *evaluation* akan melihat performa dari hasil prediksi *training dataset* dan *testing dataset* dengan menggunakan *confusion matrix* dan visualisasi metrik serta melakukan perhitungan dari *confusion matrix* yang terdiri dari *precision, recall, dan f1-score* [13].



Gambar 3.8 *Flowchart Evaluation*

3.3.6 Deployment

Pada tahap ini dimana model yang telah di *train* dan evaluasi akan diimplementasi untuk memprediksi data sentimen terkait pelecehan seksual di UMN. Hasil dari prediksi tersebut akan dianalisa menggunakan teknik visualisasi untuk dapat memahami jawaban atau solusi dari penelitian ini.



Gambar 3.9 *Flowchart Deployment*

3.4 Variabel Penelitian

3.4.1 Variabel Independen

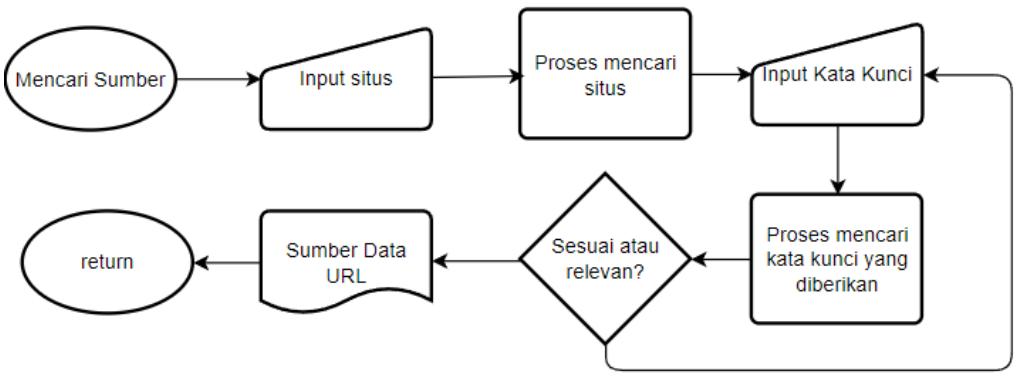
Variabel independen merupakan variabel yang memiliki pengaruh terhadap variabel lainnya. Pada penelitian ini variabel independen yang digunakan adalah komentar dari hasil postingan pengguna yang berkaitan dengan pelecehan seksual dan sumber atau URL.

3.4.2 Variabel Dependen

Variabel dependen adalah variabel yang terpengaruh dengan variabel independen. Dalam penelitian ini, variabel dependen yang digunakan adalah variabel sentimen yang berklasifikasi ‘positif’, ‘negatif’ dan ‘netral’.

3.5 Teknik Pengumpulan Data

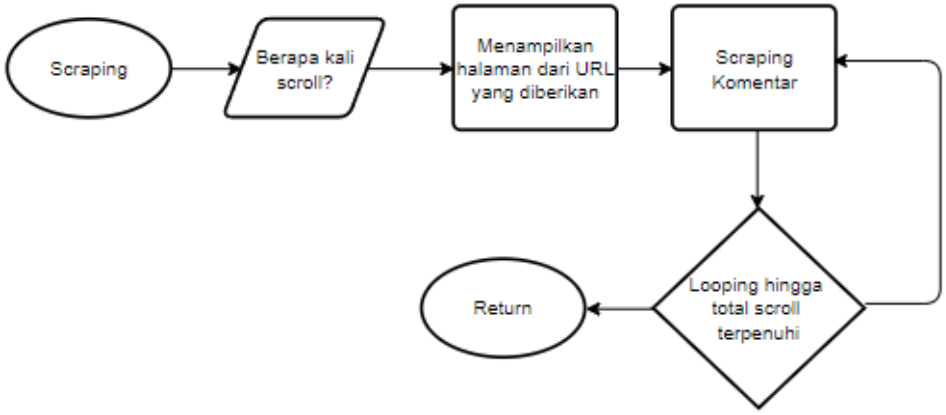
Pada teknik pengumpulan data secara garis besar sudah di jelaskan pada tahapan *data understanding*. Pada sub-bab ini akan menjelaskan secara rinci dalam pengumpulan data. Dimana tahapan mencari sumber data dan mengambil data menggunakan teknik *web scraping* menggunakan *tools python* yang dibuat sendiri menggunakan *library selenium.* Pengumpulan sumber data URL harus dilakukan terlebih dahulu berdasarkan pada gambar 3.11 yang dapat membuka website dan mencari *postingan* dengan kata kunci yang telah diberikan jika data sesuai maka URL yang berisi sumber data akan di tampung ke dalam dokumen yang berisi *array*.



Gambar 3.10 *Flowchart* Pengumpulan Data

Setelah sumber data selesai dilakukan maka proses selanjutnya melakukan teknik *scraping* yang dilakukan menggunakan *tools python* dengan *library selenium* agar mengekstrak informasi yang terdapat didalam website secara otomatis. Pada gambar 3.12 merupakan alur dari *scraping tweets* di Twitter yang akan menjalankan URL yang telah di tetapkan serta seberapa banyak

scroll yang akan digunakan dikarenakan Twitter menggunakan teknik *infinite scroll* pada *website*. *Tools* yang dibuat akan secara otomatis *scrolling* hingga jumlah *scroll* sesuai dengan kondisi. Pada penelitian ini menggunakan 20 kali *scroll* untuk mengumpulkan data komentar berdasarkan *tweets* yang tertampil di URL tersebut.



Gambar 3.11 *Flowchart Scraping*

3.6 Teknik Pengambilan Sampel

Teknik pengambilan sampel pada penelitian ini secara umum mirip dengan teknik pengumpulan data akan tetapi data yang di ambil dilakukan secara acak dan jumlah sampel data yang kurang lebih 20 hingga 50 sampel. Pengumpulan sampel menggunakan teknik *scraping data* yang telah di tentukan URL dari postingan atau kata kunci dengan postingan yang acak menggunakan python. Dari jumlah sampel data yang berkaitan dengan topik penelitian kemudian dibantu oleh 3 sukarelawan untuk membantu menghapus data yang tidak berkaitan dan memilih beberapa data yang berkaitan kemudian dilakukan proses *labelling* yang dilakukan secara manual oleh ketiga sukaraleawan dan dibantu oleh ahli bahasa untuk verifikasi hasil *labelling* tersebut.

3.7 Teknik Analisis Data

Tabel 3.2 *Tools* Perbandingan

Tools	**Kelebihan**	**Kekurangan**
Python	- Memiliki Kecepatan dalam komputasi performa - Membawa teknologi didorong bersama dengan sejumlah besar informasi dari data yang tersedia - Pemrograman python dalam menjalankan kode atau analisis data yang jauh lebih baik dan ramah bagi pengguna	- Komputasi yang cepat akan mengonsumsi memory yang besar.
R	- Sangat bergantung pada kemajuan berorientasi teknologi	- Pemrograman R yang secara tidak langsung mempengaruhi kinerja komputasi - Susah dipelajari oleh pengguna

Sumber: [37]

Pada tabel 3.3 menunjukkan perbandingan dari kedua *Tools* yang akan digunakan pada penelitian ini. Untuk Penelitian ini yang menggunakan pendekatan *text mining* akan menggunakan *tools* python karena memiliki performa yang cukup baik dalam komputasi dan mudah dipelajari [37].