BAB III
METODOLOGI PENELITIAN

3.1. Jenis Penelitian

Jenis penelitian dalam tesis ini adalah penelitian eksperimen. Penelitian eksperimen adalah Penelitian eksperimen adalah penelitian yang dilakukan dengan pendekatan saintifik dengan menggunakan dua set variabel. Set pertama bertindak sebagai konstanta, yang dapat gunakan untuk mengukur perbedaan dari set kedua. Metode penelitian kuantitatif, misalnya, bersifat eksperimental.

3.2. Sumber Data Penelitian

Dalam penelitian ini terdapat dua sumber data, yaitu sumber data primer dan sumber data sekunder.

3.2.1. Sumber Data Primer

Sumber data primer dalam penelitian ini ialah *dataset* kalimat normal maupun ujaran kebencian berasal dari komentar di media sosial Twitter yang dapat diakses oleh semua pihak, berupa teks yang diambil melalui penelusuran secara *online* pada kolom komentar Twitter antara bulan Januari - Juni 2023.

Tabel 3. 1Contoh Ujaran kebencian (Sumber : Data Diolah 2022)

No	Alamat	Kalimat ujarannya	Tanggal
1	@worksfess (netral)	petugas imigrasi tu emang jutek ya? gua cuman nanya doang padahal, nanyanya juga baik2. bikin emosi aja malem2	21 Des 2022 09.40 PM
2	@ruhutsitompul (Negatif)	Ha ha ha yg gini mau jadi Presiden RI ?, yg ta'unya	22 Des 2022 5.18 PM

No	Alamat	Kalimat ujarannya	Tanggal
		hanya ganti2 nama Rumah Sakit jadi Rumah Sehat ganti nama2 jalan eh Rumah Makan jadi Rumah Kenyang dasar kadrun	
3	@ainunrozi (merendahkan)	kadang pengeeeeen banget nolol-nololin orang depan mukanya	22 Des 2022, 12.09 PM

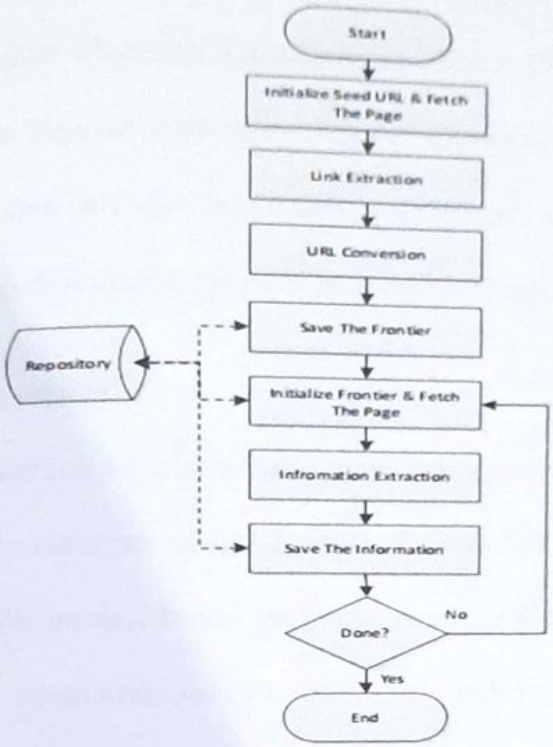
3.2.2. Pelabelan Twitter API

API Twitter memungkinkan akses terprogram ke Twitter dengan cara yang unik dan canggih. Manfaatkan elemen inti Twitter seperti: Tweet, Pesan Langsung, Ruang, Daftar, pengguna, dan lainnya.

3.3. Alat Pengumpulan Data

3.3.1. *Web Crawler*

Web Crawler adalah program yang menelusuri World Wide Web dengan cara yang metodis, otomatis dan teratur. Istilah lain untuk web crawler adalah ant, automatic indexer, bots, web spiders atau web robots. Web Crawler adalah salah satu jenis bot atau agen perangkat lunak. Secara umum, proses crawling dimulai dengan list URL yang akan dikunjungi, disebut seeds. Kemudian web crawler akan mengunjungi URL tersebut satu per satu. Setiap page URL yang dikunjungi akan diidentifikasi apakah ada hyperlink di dalamnya. Jika ada maka akan ditambahkan ke dalam list URL yang akan dikunjungi. Ini disebut crawl frontier. [21]



Gambar 3. 1Langkah-Langkah dalam Metode Web Crawling [9]

URL yang didapat dari crawl frontier ajab dikunjungi secara rekursif dengan beberapa kebijakan tertentu [2]. Langkah-langkah dalam melakukan metode web crawling

1. *Initialize Seed URL and Fetch The Page*

Langkah pertama yaitu menginisialisasi *Seed* URL. Seed URL adalah URL awal yang sudah ditentukan oleh pengguna untuk melakukan *crawling*. Seed URL dalam penelitian ini adalah *webpage* utama TripAdvisor yang berisi daftar nama restoran yang ada di Bandar Lampung. *Seed* URL tersebut diakses halaman webnya.

2. *Link Extraction*

Selanjutnya, proses ekstraksi URL yang ada di halaman tersebut sebelumnya. URL yang belum dikunjungi disebut sebagai *Frontier*. Ektraksi dilakukan dengan menggunakan *Regular Expression*. *Regular Expression* akan mengidentifikasi pola URL dalam data teks. Selain URL, para tahapan ini diekstraksi juga nama restoran, yang dilakukan dengan cara mem-*parsing* berdasarkan tag HTML "*title*".

3. *URL Conversion*

URL hasil ekstraksi bentuknya masih belum baku (*relative* URL), sehingga tidak bisa langsung disimpan sebagai *Frontier*. *Relative* URL tersebut harus diubah terlebih dahulu menjadi bentuk yang baku (*absolute* URL). Konversi dilakukan dengan cara menambahkan URL utama dari *website* TripAdvisor ke depan *relative* URL. Contoh *relative* URL ;

Restaurant_Review-g297722-d7309448- Reviews-El_s_Coffee_House Bandar_Lampung_Lampung_Sumatra.html Setelah Dikonversi menjadi *absolute* URL : https://www.tripadvisor.co.id/ Restaurant_Review-g297722-d7309448- Reviews-El_s_Coffee_HouseBandar_Lampung_Lampung_Sumatra.html

4. *Save The Frontier*

Absolute URL kemudian disimpan ke dalam *repository* untuk digunakan pada tahap selanjutnya.

5. *Initialize Frontier and Fetch The page*

URL dari *Frontier* kemudian diambil kembali dari *repository* untuk diakses halaman webnya. *Frontier* pada penelitian ini berisi detail informasi terkait sebuah restoran.

6. *Information Extraction*

Langkah selanjutnya adalah mengekstraksi informasi dengan cara mem-*parsing* berdasarkan struktur halaman HTML, yang dilihat dari *tag* HTML. Pada penelitian ini beberapa informasi yang diekstraksi yaitu alamat, kota, kode pos, *website*, nomor telepon, *rating*, dan jumlah *reviewer*. Tabel 1 menunjukkan jenis informasi yang diekstraksi beserta *tag*-nya di dalam HTML. Beberapa informasi ada yang tersimpan dalam HTML yang sama, seperti informasi kota dan kode pos yang tersimpan dalam tag "*locality*", serta *rating* dan jumlah reviewer yang tersimpan dalam tag "*rating*". Oleh karena itu, perlu dilakukan proses ektraksi selanjutnya. Ektraksi dilakukan dengan menggunakan *Regular Expression*.

7. Save The Information

Hasil informasi yang sudah didapatkan kemudian disimpan di dalam *repository*.

3.3.2. *Uniform Resource Locator*

URL merupakan sistem pengalamatan yang digunakan pada World Wide Web. Di internet URL menggabungkan informasi mengenai jenis protokol yang digunakan, alamat situs dimana resource ditempatkan, lokasi sub directory dan nama file yang digunakan.

Sintak lengkap suatu URL: Access-methode://server_name1: [port]/directory/file Contoh: http://www.microsoft.com/mspress/net/default.asp

URL diatas terdiri dari komponen-komponen:

a. http: tipe internet protokol yang digunakan untuk menyimpan dan mengirim informasi.

b. ://: standar pemberian tanda baca URL.

c. www.microsoft.com: nama domain situs dimana resources disimpan.

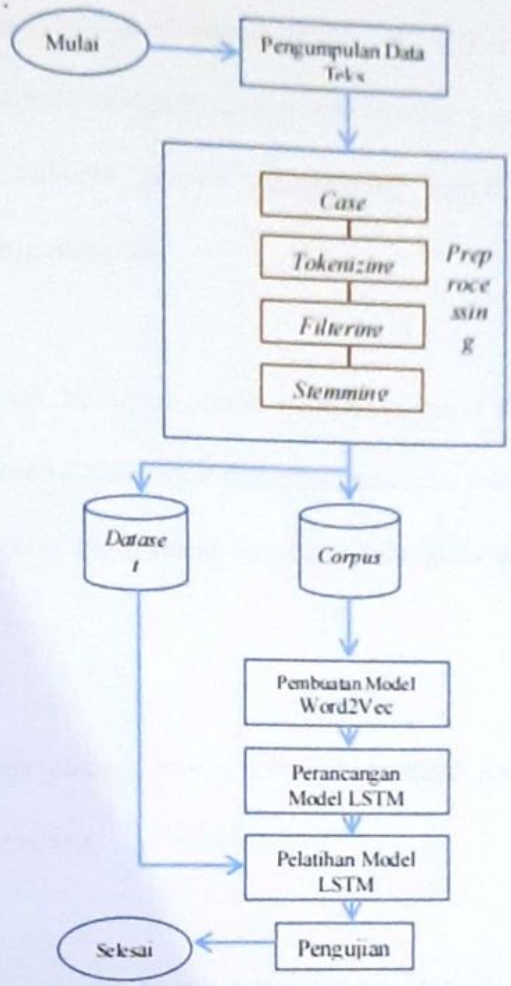
d. /mspress/net: tempat directory ke resources yang disimpan di komputer yang jauh (dalam hal ini sebuah file).

e. Default.asp: nama file yang dibuka.

URL menyediakan sebuah daftar metode yang konsisten dan mudah dimengerti dari berbagai macam situs internet, terutama pada situs world wide web [3].

3.4. Tahapan dalam Penelitian

Secara umum, tahapan penelitian ini diawali dengan studi literatur, dilanjutkan dengan pengumpulan data untuk membuat *corpus* dan *dataset* yang berasal dari data *corpus* dari Twitter. *Data preprocessing* (pra-pemrosesan data) yang dilakukan terhadap data teks (*corpus* dan *dataset*) adalah pembersihan *stopwords*, *stemming*, dan menghilangkan karakter-karakter yang tidak diperlukan.



Gambar 3. 2 Alur Metodologi Penelitian [13]

Hasil dari pra-pemrosesan dikumpulkan ke dalam sebuah *file* teks. Selanjutnya, tahapan pembuatan model *Word2Vec*. Teks hasil prapemrosesan diubah ke dalam bentuk vektor agar dapat dibaca oleh algoritma Long Short-Term Memory (LSTM). Dilanjutkan dengan merancang model *Long-Short Term Memory* yang sesuai dengan *corpus* dan *dataset* agar model yang dirancang dapat mendeteksi penggunaan ujaran kebencian dengan baik, model dilatih dengan menggunakan *dataset* yang berisi kumpulan kalimat mengandung ujaran kebencian, dan diakhiri dengan pengujian model.

Sebelum dilakukan pembuatan model *Word2Vec*, dilakukan pra-pemrosesan terlebih dahulu menggunakan *library Python* yaitu *gensim* [9] pada *corpus* dan *dataset*. *Library gensim* mengandung fungsi untuk melakukan prapemrosesan teks yeng meliputi:

1. Case Folding

Proses *Case Folding* ini bertujuan untuk mengubah huruf dalam teks menjadi huruf standar (huruf kecil). Data yang diterima pada *case folding* hanya huruf 'a' sampai 'z', karakter selain huruf-huruf tersebut dihilangkan dan hanya dianggap delimiter.

2. Tokenizing

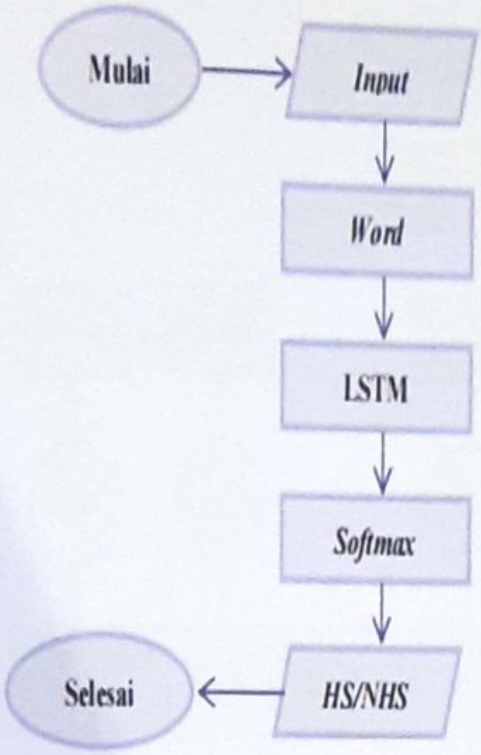
Setelah melakukan *case folding*, proses selanjutnya adalah data teks dipecah per kata pada tahapan *tokenizing*.

3. Filtering

Pada tahap *filtering* dilakukan proses penghapusan kata yang tidak diperlukan atau diluar objek penelitian (*stopwords*). Beberapa kata *stopwords* pada penelitian ini diantaranya adalah 'maka', 'akan', 'yang', 'untuk', 'dan', 'juga', 'dari', 'di', 'ya', serta 'kan'.

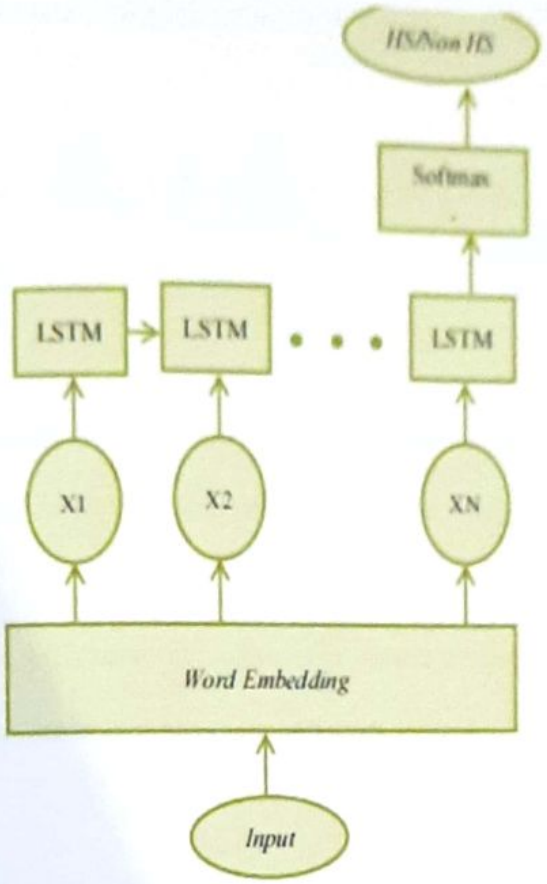
4. Stemming

Setelah proses *filtering* dilakukan, proses yang terakhir adalah *stemming* dimana kata yang ada pada data ditransformasi ke kata dasarnya. Dalam perancangan Word2Vec, digunakan *library gensim*. *Library gensim* diperlukan untuk melakukan *import* fungsi Word2Vec, dimana fungsi ini digunakan untuk mengubah teks pada "*corpus.txt*" dari kata menjadi vektor untuk dapat dijalankan pada algoritma LSTM dan hasil dari Word2Vec disimpan ke model_word2vec_300.model.



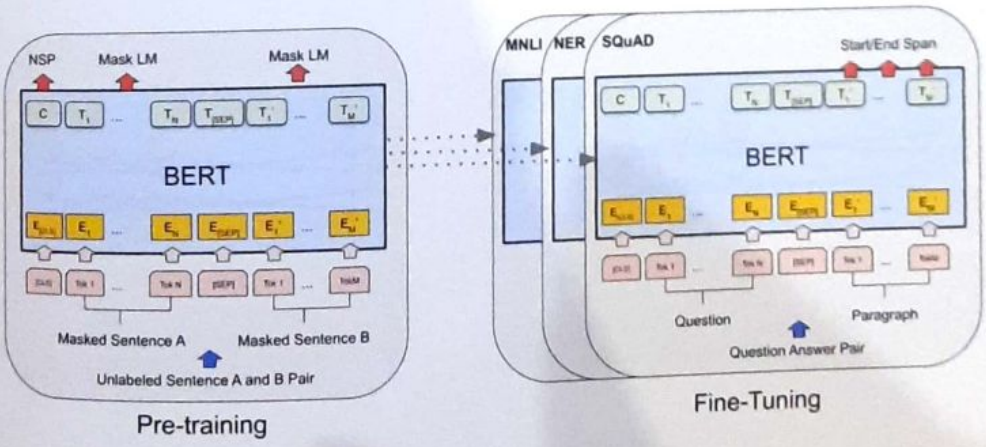
Gambar 3. 3 Gambaran Umum LSTM [12]

Pada penelitian ini, LSTM digunakan sebagai algoritma pada program pendekteksi ujaran kebencian, cara kerja dari LSTM secara umum dapat dilihat pada Gambar 3.2. Algoritma LSTM dipilih karena memiliki memori yang besar, cocok digunakan untuk data berbentuk *sequence*. Data diinput terlebih dahulu sebelum masuk ke algoritma LSTM melalui tahap *word embedding*, algoritma LSTM kemudian memproses hasil dari *word embedding*. Sesudah input pertama selesai diproses, hasil tersebut dimasukkan ke memori, dimana hasil input sebelumnya akan digunakan untuk proses berikutnya dalam memeriksa hubungan antara *input* pertama dengan kedua, ketiga, keempat dan selanjutnya. Hasil proses berikutnya dipengaruhi oleh hasil *input* sebelumnya. *Input* berupa kalimat yang terdiri dari beberapa kata. Cara kerja LSTM secara lebih detail dapat dilihat pada Gambar 3.



Gambar 3. 4 Cara Kerja Algoritma LSTM [12]

3.5. Cara Kerja BERT



Gambar 3. 5 Langkah dalam BERT

Secara umum ada 2 langkah dalam BERT, yaitu : [22]

1. Pra-pelatihan

Selama prapelatihan, model dilatih pada data yang tidak berlabel melalui berbagai tugas prapelatihan.

2. Penyetelan halus

Model BERT pertama kali diinisialisasi dengan parameter yang telah dilatih sebelumnya, dan semua parameter disetel dengan baik menggunakan data berlabel dari tugas hilir.

Setiap tugas hilir memiliki model fine-tuned terpisah, meskipun mereka diinisialisasi dengan parameter pra-pelatihan yang sama.