BAB 3
METODOLOGI PENELITIAN

3.1 Metodologi Penelitian

3.1.1 Studi Literatur

Pada tahap pertama, dilakukan studi literatur dengan membaca penelitian-penelitian sebelumnya seperti analisis sentimen, *text-preprocessing*, metode LSTM, *scraping data*. Selain itu, terdapat juga beberapa informasi yang dikumpulkan seperti ”vaksin keempat Covid-19” atau ”booster kedua Covid-19”. Segala informasi yang dikumpulkan didapatkan melalui penelitian sebelumnya, jurnal *online* maupun *website*.

3.1.2 Pengumpulan Data

Pada tahapan ini, akan dilakukan pengumpulan data berupa *tweet* dari pengguna Twitter tentang vaksinasi Covid-19 dosis ke-4. Data *tweet* dikumpulkan dengan menggunakan Snscrape yang merupakan *scraping tools*.

3.1.3 Pemrograman Sistem

Pada tahapan ini, dilakukan 1proses pengolahan data hingga proses *modelling*. Data yang sudah dikumpulkan akan dijadikan data yang informatif untuk model *machine learning*. Setelah itu, akan dibuat model *machine learning* yang nantinya dapat mengklasifikasikan sentimen berdasarkan data.

3.1.4 Pengujian dan Evaluasi

Setelah model berhasil dilatih, akan dilakukan tahap pengujian dengan memprediksi sentimen dari data test yang sudah disiapkan, serta selanjutnya yaitu dilakukan evaluasi untuk dapat mengetahui performa dari model LSTM yang sudah dibuat dengan menggunakan *confusion matrix*.

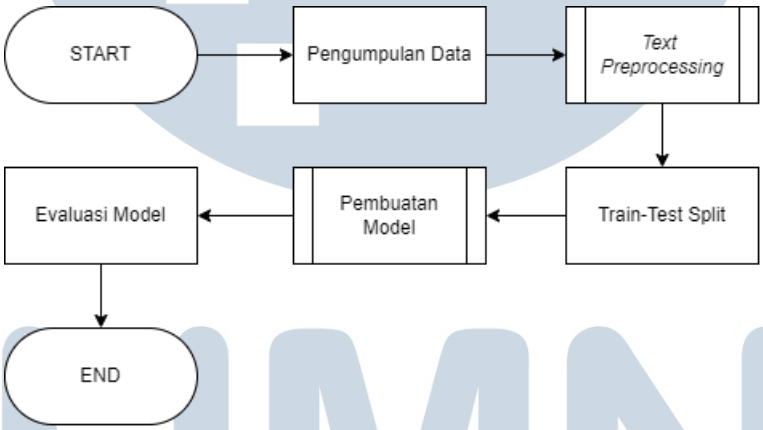
3.1.5 Penulisan Laporan

Penulisan laporan dibuat berdasarkan dari penelitian yang telah dilakukan sebagai hasil dokumentasi mengenai penelitian yang telah dilakukan. Penulisan laporan juga menjadi bukti bahwa penelitian telah dilakukan dan diselesaikan. Proses dari penulisan laporan akan dibuat mulai dari bagian pendahuluan hingga kesimpulan dan saran terhadap penelitian yang dilakukan.

3.2 Perancangan Sistem

3.2.1 Gambaran Umum Sistem

Gambar 3.1 berikut adalah *flowchart* dari penelitian yang akan dilakukan terkait sentimen analisis terhadap vaksin covid dosis ke-4 dengan algoritma LSTM.



Gambar 3.1. *Flowchart* Gambaran Umum Sistem

Pada tahapan awal akan dilakukan pengumpulan data dimana data akan dikumpulkan dari media sosial Twitter dengan menggunakan *library* Snscrape, lalu pengumpulan data akan disesuaikan pengambilannya berdasarkan kata kunci yang diperlukan. Data kemudian disimpan kedalam file berbentuk csv yang lalu digunakan untuk proses selanjutnya.

Setelah mendapatkan kumpulan data, dilanjutkan dengan tahapan *text preprocessing* yaitu data yang sudah dikumpulkan akan masuk pada tahap pembersihan seperti menghapus URL, angka, *hashtag*, *case folding*, dan penghapusan fitur lainnya, sehingga nantinya model dapat belajar dengan lebih baik dengan data yang sudah dibersihkan. Selanjutnya juga dilakukan tahapan *tokenizing* dan penghapusan *stopwords* untuk memisahkan teks menjadi per kata

dan menghilangkan kata-kata yang tidak memiliki makna, lalu proses *stemming* yang bertujuan untuk mengubah kata yang memiliki imbuhan menjadi kata dasar. Setelah itu, dilanjutkan dengan proses penghapusan data-data *tweet* yang terdapat pada akun-akun yang dianggap tidak memberikan kontribusi yang besar. Data *tweet* selanjutnya akan diberi label pada proses pelabelan data yang diklasifikasikan kedalam tiga macam, yaitu sentimen positif, negatif, dan netral. Selanjutnya yaitu masuk ke *resampling* pada data untuk membuat setiap kelas/label memiliki data yang seimbang, dan yang terakhir yaitu melakukan *encoding* dan *padding* pada data untuk mengubah data menjadi data numerik dan membuat data memiliki panjang yang sama.

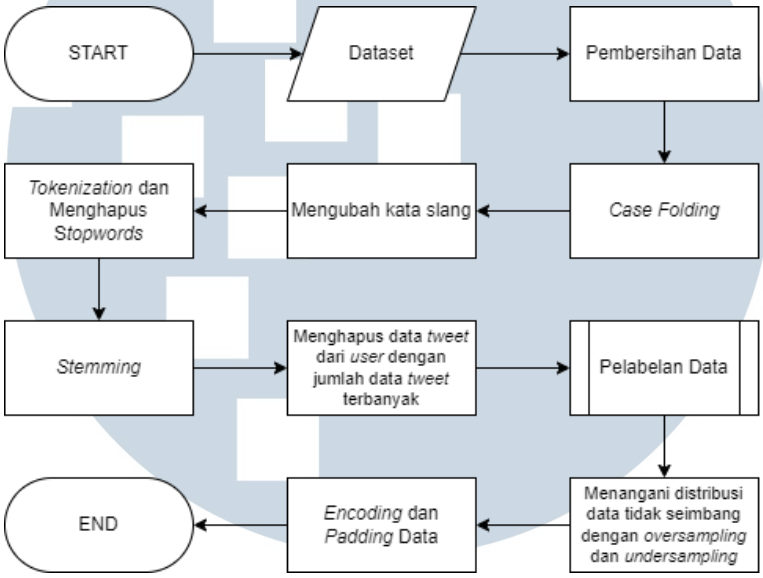
Data selanjutnya akan dibagi pada tahapan *split* data dan menghasilkan data *train*, *test*, dan *validation*. Selanjutnya, akan dilakukan proses pembuatan model LSTM berdasarkan skenario uji coba yang telah dibuat. Data *train* selanjutnya akan digunakan untuk melatih model serta data *validation* untuk mendapatkan hasil evaluasi selama model dilatih. Kemudian, pada proses terakhir akan dilakukan pengujian pada data *test* untuk mengevaluasi kinerja model *machine learning* dalam mengklasifikasikan data dengan menggunakan *confusion matrix*.

3.2.2 Pengumpulan Data

Tahapan pengumpulan data adalah tahapan dimana data berupa *tweet* dikumpulkan dengan menggunakan *scraping tool* berupa Snscrape. Data yang diambil berupa *keyword* ”vaksin keempat” dan ”booster kedua”, serta data dari suatu *tweet* yang berhubungan dengan booster kedua. Pengambilan data dimulai dari tanggal Agustus 2022 hingga Februari 2023. Pada pengambilan data *tweet* dengan *keyword* ”booster kedua” dilakukan dengan bertahap dimana diambil dari bulan Agustus 2022 hingga November 2022 dan bulan Desember 2022 hingga bulan Januari 2023. Hal ini dikarenakan ketika mengambil data sekaligus tidak mendapat banyak data, oleh karena itu dilakukan pemisahan rentang waktu dalam pengambilan data. Kemudian nantinya data akan disimpan dalam format .csv.

3.2.3 *Text Preprocessing*

Pada tahapan *preprocessing*, data yang sudah dikumpulkan pada tahapan sebelumnya kemudian akan dibersihkan dan disesuaikan untuk nantinya akan diolah model *machine learning*. Berikut merupakan bagan alur untuk tahapan *text preprocessing* yang dapat dilihat pada gambar 3.2



Gambar 3.2. *Flowchart* Text *Preprocessing*

Preprocessing dimulai dari pembersihan kata-kata seperti menghilangkan alamat URL, *hashtag*, tanda baca, serta mengubah semua huruf dalam kalimat menjadi huruf kecil. Setelah pembersihan dilakukan, selanjutnya yaitu mengubah kata-kata slang menjadi kata yang sesuai dengan bahasa Indonesia. Selain itu, terdapat juga proses penghapusan *stopword* yaitu proses penghapusan kata-kata yang tidak memiliki makna yang berarti dan akan dihilangkan, serta proses *stemming* yaitu mengubah kata-kata yang memiliki imbuhan kedalam bentuk kata dasarnya. Selanjutnya, terdapat juga penghapusan data-data *tweet* yang memiliki kalimat yang terduplikat, serta menghapus data *tweet* yang muncul dari *user* Twitter yang memiliki data *tweet* terbanyak. Penghapusan data *tweet* ini dilakukan untuk menghilangkan data-data yang tidak memiliki kontribusi yang besar terhadap hasil klasifikasi nantinya.

Tahapan berikutnya yaitu proses pelabelan data dimana setiap data teks akan diberikan label sentimen antara sentimen positif, negatif, dan netral. Proses pelabelan data dapat dilihat pada gambar 3.3.

Setelah mendapatkan dataset yang memiliki label sentimen, akan dilakukan *resampling* atau penanganan kelas data yang tidak seimbang dengan menggunakan metode *oversampling* dan *undersampling* berdasarkan skenario uji coba yang dibuat. Setelah itu, akan dilakukan *encode* pada data untuk mengubah data teks menjadi data numerikal, serta akan dilakukan *padding* dimana data pada dataset yang sudah di-*encode* memiliki panjang yang sama. Hal ini bertujuan untuk mempermudah model *machine learning* dalam mempelajari data.

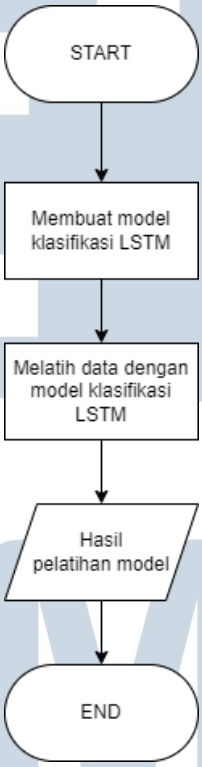
3.2.4 Pelabelan Data

Gambar 3.3 merupakan alur tahapan untuk melakukan pelabelan data. Pada tahapan pelabelan, pertama-tama akan memuat terlebih dahulu *lexicon* InSet yang dibutuhkan, lalu selanjutnya akan dibuat fungsi dimana data *tweet* akan diberikan skor berdasarkan kata-kata yang ada di *lexicon* yang terdapat bobot pada setiap kata-katanya. Bobot pada tiap kata yang terdapat pada *lexicon* InSet memiliki nilai yang beragam dengan nilai skor negatif maksimal -5 dan nilai skor positif maksimal 5. Setelah setiap data *tweet* telah diberikan skor, selanjutnya data akan diberi label sentimen berdasarkan skor yang ada. Data dengan label sentimen negatif yaitu data dengan nilai skor berada kurang dari 0, label sentimen positif yaitu data dengan nilai skor berada lebih dari 0, dan label sentimen netral untuk data dengan skor bernilai 0. Selanjutnya, data akan divisualisasikan kedalam bentuk *word cloud* untuk setiap ketegori sentimen positif, negatif, dan netral, dimana akan memunculkan suatu gambar berupa kata-kata yang paling sering muncul dalam setiap kategori. Kemudian, data akan disimpan kedalam bentuk file .csv.

Gambar 3.3. *Flowchart* Pelabelan Data

3.2.5 Pembuatan Model

Pada gambar 3.4 menunjukkan alur untuk tahapan pembuatan model dengan *Long Short-Term Memory* (LSTM). Dari data yang telah dibagi menjadi data *train*, *test*, dan *validation*, proses dilanjutkan dengan membuat model LSTM dan menyesuaikan jaringan arsitektur dan *hyperparameter tuning* berdasarkan skenario uji coba yang dibuat. Setelah model telah dibuat, maka model akan menggunakan data *train* sebagai data latih dan data *validation* untuk mengevaluasi model selama proses pelatihan.



Gambar 3.4. *Flowchart* Pembuatan Model

3.2.6 Evaluasi Model

Pada tahapan evaluasi model, setelah menyelesaikan pelatihan untuk klasifikasi model dengan LSTM, data *test* akan digunakan untuk mengevaluasi hasil pembelajaran model dengan menghitung nilai *accuracy, precision, recall,* serta *F1-score*.