BAB I
PENDAHULUAN

1.1 Latar Belakang

Seiring dengan perkembangan teknologi informasi yang kian pesat, salah satunya penggunaan media sosial. Saat ini semua orang sangat dekat dengan media sosial, karena memang menurut Kaplan & Haenlein Media sosial adalah sekumpulan aplikasi berbasis internet, beralaskan pada ideologi dan teknologi Web 2.0 sehingga memungkinkan penciptaan dan pertukaran konten oleh penggunanya (Kaplan & Haenlein, 2010). Dengan adanya sosialisasi secara maya ini maka media sosial juga dapat dikatakan sebagai media untuk bertukar informasi antar penggunanya dan di dalamnya banyak pengguna yang menyebarkan informasi mengenai kehidupan sehari-hari yang mengandung unsur opini saat merangkainya. Seringkali pengguna media sosial sulit untuk dapat memahami informasi yang pengguna lain bagikan karena informasi yang dibagikan hanya dalam bentuk teks. Namun, teks ini harus bisa merepresentasikan opini yang sedang dicurahkan oleh penulis. Sehingga, hal ini yang sering menyebabkan terjadinya ketimpangan antara informasi yang sedang dibicarakan dan informasi yang diinterpretasikan dan banyaknya penyebaran informasi yang tidak benar adanya (hoax) karena kita tidak benar-benar bisa paham dengan rasa yg ingin disampaikan walaupun pengguna tersebut telah menggunakan tulisan tipografi yang merepresentasikan ekspresi wajah di dalam teks nya. Maka banyak penelitian pembelajaran mesin yang dilakukan untuk menghasilkan sebuah program yang dapat membantu menganalisis hal tersebut.

Analisis Sentimen merupakan sebuah bidang pengelolaan data tekstual yang melakukan kajian berdasarkan opini, sentimen, evaluasi, perilaku dan emosi seseorang yang dapat dijadikan bahan evaluasi. Analisis ini dapat dilakukan dengan menerapkan beberapa algoritma seperti Naïve Bayes, Support Vector Machine, Deep Neural Network, Recurrent Neural Network dan masih banyak lagi algoritma pembelajaran mesin lainnya. Menurut penelitian yang dilakukan oleh Nurdin, Seno Aji, Bustamin, & Abidin banyak variasi yang bisa diterapkan dalam melakukan analisis, salah satunya dengan menerapkan word embedding. Word embedding

merupakan salah satu teknik untuk merepresentasikan sebuah kata yang mirip menjadi sebuah kata yang memiliki arti sama dengan memetakan kata dalam dokumen ke dalam dense vector, vektor yang merupakan sebuah proyeksi representasi dari kata di dalam ruang vektor. Berikut metode word embedding yang dapat diterapkan, antara lain Term Frequency-Inverse Document Frequency (TF-IDF), GloVe, Word2Vec, FastText, dan sebagainya. (Nurdin, Seno Aji, Bustamin, & Abidin, 2020)

Adapun penelitian tentang analisis sentimen yang telah dilakukan sebelumnya. Klasifikasi sentimen Dengue Disease pada sosial media twitter menggunakan metode word embedding TF-IDF yang diimplementasikan oleh Amin et, al. dilakukan dengan mengambil data dari Twitter yang akan disimpan kedalam database, selanjutnya dilakukan data preprocessing dengan melakukan stopword removal, tokenization, dan stemming serta mentraining data yang sebelumnya sudah dilakukan word embedding dengan metode TF-IDF menggunakan beberapa algoritma. Penelitian ini menghasilkan accuracy untuk setiap algoritmanya, sebagai berikut Logistic Regression (LR) 87.07%, Support Vector Machine (SVM) 80.81%, Naive Bayes (NB) 89.95%, Artificial Neural Network (ANN) 88.92%, Deep Neural Network (DNN) 89.82%, Long Short-Term Memory (LSTM) 92.88% (Amin, et al., 2020). Pada penelitian lain oleh Muhammada, Kusumaningrum, & Wibowo menggunakan dataset berupa review hotel yang ada di Indonesia, data tersebut diambil dari website Traveloka dan dilakukan preprocessing dengan case folding, tokenization, stemming, stopword removal, and padding. Selanjutnya data di-train menggunakan salah satu metode word embedding yaitu Word2Vec dan menerapkan salah satu varian dari RNN yaitu LSTM. Dan menghasilkan mean accuracy terbaik sebesar 85.9% dengan menggunakan Skip-Gram arsitektur Word2Vec, metode evaluasi Hierarchical Softmax dan nilai vector dimension sebesar 300, menggunakan LSTM model dengan dropout value sebesar 0.2, learning rate 0.001, dan menggunakan teknik average pooling (Muhammada, Kusumaningrum, & Wibowo, 2021)

Dengan adanya wabah virus COVID-19 seperti saat ini hampir semua kegiatan sehari-hari dilakukan secara online sehingga informasi tersebar luas secara maya di internet dalam jumlah yang sangat banyak. Wabah virus COVID-19 yang

melanda bukan hanya Indonesia namun seluruh dunia ini sedang ramai diperbincangkan sosial media, salah satu nya mengenai penanggulangan virus ini yang dilakukan dengan pemberian vaksin kepada masyarakat. Ramainya pengguna Twitter yang membicarakan topik ini menimbulkan meningkatnya penyebaran informasi yang tidak benar adanya (hoax). Ketersediaan informasi ini dapat dijadikan tempat untuk mengumpulkan informasi mengenai opini pengguna yang nantinya dapat dianalisis sehingga menghasilkan suatu informasi baru. Dalam hal ini, maka dapat dilakukan percobaan untuk menganalisis hal tersebut dengan membandingkan metode word embedding TF-IDF dan Word2Vec yang diterapkan pada algoritma Recurrent Neural Network terhadap opini pengguna mengenai topik vaksinasi COVID-19 dan akan dihasilkan informasi baru yakni pengklasifikasian opini tersebut termasuk ke dalam kelas positif, neutral, negatif, dan *unrelated*. Metode word embedding yang dipilih merupakan metode yang yang sering digunakan pada penelitian sebelumnya dan menampilkan hasil yang baik namun banyak dari penelitian pendahulu mengenai topik analisis sentimen menggunakan Recurrent Neural Network masih sedikit yang membandingkan metode word embedding TF-IDF dan Word2Vec secara langsung dalam satu penelitian, maka dari itu pada penelitian ini akan dilakukan percobaan dengan melakukan klasifikasi menggunakan RNN dimana digunakan sebuah input berupa vektor hasil dari kedua metode tersebut. Kedua metode ini pada dasarnya sama-sama menghasilkan sebuah vektor. Namun, perhatian penelitian ini adalah value yang dihasilkan oleh kedua vektor tersebut. Penerapan metode TF-IDF akan menghasilkan sebuah vektor yang memiliki ukuran sebesar kumpulan kata unik yang dimiliki. Kumpulan kata unik ini diambil oleh model berdasarkan data training yang tersedia. Vektor pada TF-IDF berisi bobot dari masing-masing kata yang menyusun sebuah kalimat. Bobot kata tersebut dihitung dari jumlah frekuensi munculnya kata pada sebuah kalimat yang selanjutnya dikalikan dengan hasil perbandingan kata tersebut akan dengan jumlah seluruh kata yang terdapat pada data training. Sedangkan penerapan word2vec akan menghasilkan sebuah vektor yang ukurannya dapat disesuaikan untuk setiap kata yang dimiliki. Dari penggunaan metode ini dapat diketahui similarity antar kata yang yang ada pada kalimat. Similarity tersebut dihasilkan berdasarkan kedekatan kata yang dimasukan ke dalam model word2vec. Kedekatan

kata tersebut dapat dilakukan pengaturan sesuai dengan kebutuhan. Pengaturan yang dimaksud yakni jangkauan jarak kata yang akan dilihat kedekatannya dengan kata tersebut. Contoh dari penjelasan diatas adalah pada kalimat "Mahasiswa mendapatkan vaksin" yang memiliki tiga buah kata. Penerapan TF-IDF membuat setiap kata memiliki bobot yang dihitung seperti yang telah dijelaskan sebelumnya. Jadi kata "mahasiswa", "dapat", dan "vaksin" masing- masing akan memiliki bobot. Akan tetapi, TF-IDF tidak bisa mengetahui kedekatan yang dimiliki antar kata karena TF-IDF hanya akan menghitung bobot dari masing-masing kata tersebut. Sedangkan pada penerapan word2vec kalimat tersebut akan menghasilkan nilai antar kata sehingga dapat dicari nilai kedekatan dari ketiga kata tersebut, seperti kedekatan kata "mahasiswa" terhadap kata "dapat" begitu pula dengan kedekatan kata "mahasiswa" dengan kata "vaksin". Hal inilah yang akan dibandingkan pada penelitian ini, yakni penggunaan metode yang menghasilkan sebuah vektor dengan menerapkan bobot kata terhadap dokumen dan vektor dengan menerapkan nilai kedekatan antar kata. Sehingga penulis dapat menarik sebuah kesimpulan untuk menetapkan metode word embedding yang memiliki tingkat akurasi, presisi, dan recall lebih tinggi, waktu komputasi yang lebih efisien, dan juga bobot memory yang digunakan pada sebuah perangkat.

Dari paparan diatas, maka peneliti akan melakukan penelitian dengan judul "Perbandingan Akurasi Metode Word Embedding TF-IDF Dan Word2Vec Menggunakan Recurrent Neural Network untuk Analisis Sentimen Tweet Vaksinasi Covid-19"

1.2 Rumusan Masalah

Dari paparan latar belakang di atas, maka penulis merumuskan permasalahan penelitian yang utama mengenai analisis sentimen twitter menggunakan *Recurrent Neural Network* dengan metode *word embedding* TF-IDF maupun *Word to Vector*, yaitu "Bagaimana perbedaan akurasi antara *word embedding* dengan metode TF-IDF dan metode *word to vector* pada *Recurrent Neural Network* untuk analisis sentimen tweet vaksinasi Covid-19?". Berdasarkan permasalahan tersebut dapat dijabarkan lebih detail, sebagai berikut:

1. Bagaimana perbandingan hasil penerapan metode word embedding TF-IDF dan Word2Vec untuk melakukan analisis sentimen pada tweet vaksinasi

Covid-19 menggunakan *Recurrent Neural Network* dilihat dari tingkat akurasi, presisi, dan recall kedua model?

2. Bagaimana tingkat efisiensi waktu komputasi dan bobot memori yang digunakan pada sebuah perangkat dalam membangun sebuah model yang menerapkan metode word embedding TF-IDF dan Word2Vec untuk melakukan analisis sentimen pada tweet vaksinasi Covid-19 menggunakan *Recurrent Neural Network*?

1.3 Batasan Masalah

Cukup banyak algoritma yang dapat digunakan untuk mengimplementasikan analisis sentimen sehingga penulis merumuskan batasan permasalahan penelitian agar ketidakpastian yang timbul dapat dikurangi, yakni sebagai berikut:

1. Data yang diambil dari media sosial twitter dipakai sebanyak 6.000 dan berfokus pada satu topik yakni vaksinasi COVID-19 dengan menggunakan *keyword* vaksin dan vaksinasi serta hashtag #vaksin dan #vaksinasi.
2. Penelitian hanya berfokus pada perbedaan implementasi *word embedding* dengan metode TF-IDF dan metode *Word to Vector*.
3. Penelitian ini hanya dilakukan menggunakan satu algoritma, yaitu *Recurrent Neural Network*.
4. Implementasi penelitian hanya dilakukan sampai dengan pembuktian tingkat perbedaan akurasi masing-masing metode.

1.4 Tujuan Penelitian

Sesuai dengan latar belakang dan rumusan masalah yang telah dipaparkan diatas, secara umum tujuan dari penulisan penelitian ini adalah untuk mengetahui perbandingan *word embedding* dengan metode TF-IDF dan metode *Word to Vector* pada recurrent neural network untuk analisis sentiment. Dengan pengimplementasian masing-masing metode dapat diketahui tujuan dari penelitian ini sebagai berikut:

1. Mengidentifikasi penerapan metode word embedding TF-IDF dan Word2Vec untuk melakukan analisis sentimen pada tweet vaksinasi Covid-19 menggunakan *Recurrent Neural Network* dengan membandingkan hasil dari tingkat akurasi, presisi, dan recall kedua model.

2. Mengidentifikasi tingkat efisiensi waktu komputasi dan bobot memori yang digunakan pada sebuah perangkat dalam membangun sebuah model yang menerapkan metode word embedding TF-IDF dan Word2Vec untuk melakukan analisis sentimen pada tweet vaksinasi Covid-19 menggunakan *Recurrent Neural Network.*

1.5 Manfaat Penelitian

Dengan dilakukan penelitian mengenai analisis sentiment menggunakan *Recurrent Neural Network* dengan dua metode *word embedding* yang berbeda, dapat diambil manfaat dari penelitian tersebut yaitu mengetahui metode *word embedding* manakah diantara TF-IDF atau *Word to Vector* yang menghasilkan akurasi lebih tinggi saat diimplementasikan untuk melakukan analisis sentimen tweet vaksinasi COVID-19 menggunakan *Recurrent Neural Network.*