1. Pendahuluan

Latar Belakanng

Dalam beberapa tahun terakhir, penggunaan platform media sosial di Indonesia terus meningkat. Pada awal tahun 2022, pengguna media sosial di Indonesia telah mencapai sekitar 191 juta pengguna, meningkat sebanyak 21 juta pengguna atau 12,6% dari tahun sebelumnya[1]. Media sosial seperti Facebook, Twitter, dan Instagram biasa digunakan untuk mengikuti berita terkini [2]. Penggunaan media sosial di Indonesia yang sebesar itu memiliki kekurangan dan kelebihan. Keuntungannya adalah berita mudah diakses oleh siapa saja, biaya murah, dan informasi tersebar dengan cepat. Kerugiannya adalah memungkinkan berita palsu menyebar dengan mudah. Dampak penyebaran berita hoax justru terjadi di masyarakat, yang mereka yakini tersebar sehingga menimbulkan kerugian bagi orang lain yang dianggap sebagai obyek isu tersebut[3]. Korban penyebaran berita hoax bisa bermacam-macam, tergantung informasi yang tidak benar tersebut. Pemilih dan warga negara dapat menjadi sasaran karena menyebarkan berita politik yang tidak benar, pelanggan berbelanja online untuk menyebarkan berita tidak benar tentang ulasan palsu dan iklan palsu, dan banyak lainnya[4] Berita hoax harus dideteksi karena berita hoax menyebarkan informasi yang tidak benar dan menyesatkan. Hal ini merusak integritas informasi dan perlu diklarifikasi untuk publik. Dengan mendeteksi berita hoax, kita bisa memastikan informasi yang disebarluaskan akurat dan terpercaya. Berita hoax juga kerap menyasar individu, kelompok atau lembaga tertentu untuk merusak reputasinya. Dengan mendeteksi berita hoax, kita bisa melindungi reputasi orang atau entitas yang menjadi sasaran penyebaran berita bohong. Selain itu, dengan mengungkapkan berita hoax, kita dapat memperkuat kepercayaan masyarakat terhadap sumber informasi yang terpercaya. Oleh karena itu, penelitian ini bertujuan untuk meminimalisir korban lebih lanjut dengan membangun sebuah sistem yang akan mendeteksi hoaks.

Menurut penelitian [5] yang telah dilakukan untuk mendeteksi hoaks dalam bahasa Indonesia menggunakan metode Long Short-Term Memory (LSTM), menunjukkan rata-rata akurasi sebesar 85% dari beberapa hasil percobaan. Dengan menggunakan model Word2Vec, diperoleh nilai rata-rata paling tinggi dari matriks konfusi. Mengacu pada penelitian [6] yang telah meneliti analisis sentimen menggunakan GloVe, SVM, dan TF-IDF, menunjukkan bahwa penggunaan perluasan fitur dapat menghasilkan hasil yang berbeda tergantung penggunaan korpus dan ukuran fitur yang digunakan. Proses pembobotan TF-IDF dan perluasan fitur mendapatkan hasil yang cukup baik dibandingkan dengan tidak menggunakan TF-IDF dan perluasan fitur. Nilai akurasi optimum yang diperoleh pada penelitian ini adalah 79,52%. Perbedaan dari penelitian sebelumnya adalah penelitian ini akan membandingkan dua fitur ekstraksi Word2Vec dan GloVe, dengan memiliki enam skenario. Pertama Word2Vec dengan mengubah parameter jendela menjadi 3, 4, 5, 6, dan 7. Kemudian menggunakan salah satu skenario pada GloVe.

Pendeteksian berita hoax pada media berita Indonesia di Twitter akan menggunakan Long Short-Term Memory (LSTM). LSTM merupakan salah satu model Recurrent Neural Network (RNN) yang sengaja dibuat untuk mengatasi keterbatasan RNN yang tidak dapat menangkap ketergantungan jangka panjang dengan kata lain LSTM dapat mengingat informasi jangka panjang dan cukup baik untuk diterapkan di kasus seperti analisis sentimen dan deteksi berita hoax [5]

Metode LSTM tidak dapat mengevaluasi data input jika data berupa teks atau string, oleh karena itu jika input yang diterima berupa teks atau string, diperlukan proses ekstraksi fitur yang akan mengubah teks atau string tersebut menjadi vektor numerik. masing-masing mewakili kata. Proses ini disebut *word embedding* [7]. Salah satu word embeddings yang digunakan dalam penelitian ini adalah Word2Vec. Word2Vec mengandalkan informasi lokal dari bahasa, dan kata-kata di sekitarnya memengaruhi semantik yang dipelajari dari kata tertentu [8]. Alasan penggunaan GloVe adalah Pembentukan hubungan semantik. GloVe memungkinkan penggabungan informasi semantik dalam representasi vektor kata-kata. Dalam representasi vektor GloVe, kata-kata dengan hubungan semantik yang mirip akan memiliki representasi vektor yang mendekati satu sama lain dalam ruang vektor. Misalnya, kata-kata yang sering muncul dalam konteks yang sama akan memiliki representasi vektor yang serupa. Ini memungkinkan model pemrosesan bahasa alami untuk mengenali dan mengeksploitasi hubungan semantik antar kata [6]. Kemudian pada penelitian ini juga menggunakan GloVe. Berbeda dengan Word2Vec, GloVe menunjukkan cara melibatkan informasi statistik global yang terkandung dalam dokumen. Makna semantik kata tidak hanya dipengaruhi oleh kata-kata di sekitarnya tetapi juga oleh informasi statis global dari dokumen [8] Penelitian ini selain menggunakan LSTM juga akan membandingkan kinerja dua fitur ekstraksi yaitu Word2Vec dan GloVe dalam mendeteksi berita hoax dari media berita Indonesia di Twitter.

Perumusan Masalah

Berdasarkan latar belakang yang telah dibahas, rumusan malasah yang akan dieselsaikan pada penelitian kali ini adalah bagaimana kinerja word embbeding GloVe dan Word2Vec dalam mendeteksi berita hoax berbahasa

Indonesia menggunakan LSTM dan membandingkan kedua word embbeding tersebut untuk mencari word emmbeding terbaik diantara GloVe dan Word2Vec pada model LSTM.

Tujuan

Adapun tujuan dari Tugas Akhir ini adalah dapat membangun sebuah sistem pendeteksi berita hoax berbahasa Indonesia menggunakan LSTM dengan GloVe dan LSTM dengan Word2Vec dan mencari word embbeding yang lebih efektif untuk mendeteksi berita hoax berbahasa Indonesai menggunakan LSTM.