## 1. Pendahuluan

### Latar Belakang

Perkembangan teknologi yang begitu pesat pada akhir-akhir ini memudahkan masyarakat untuk mendapatkan informasi secara cepat. Tetapi tidak jarang informasi yang diterima merupakan berita hoax. Hoax merupakan sebuah istilah untuk sebuah berita, informasi berupa berita bohong ataupun fitnah. Hoax dibuat dengan tujuan untuk menggiring opini pembaca agar sesuai dengan informasi yang diberitakan [1]. Outlet media terpenting untuk menyebarkan berita palsu adalah: Radio 1,20%, media cetak 5% dan televisi 8,70%. Saluran yang paling banyak digunakan untuk menyebarkan penipuan adalah media sosial (Facebook, Twitter, Instagram dan Path), yang paling banyak digunakan 92,40%, terutama Twitter [2]. Zaman sekarang ini twitter sudah banyak digunakan oleh masyarakat Indonesia dikarenakan mudahnya mencari informasi dan akses yang mudah untuk digunakan. Kementrian Komunikasi dan Informatika Republik Indonesia menyatakan selama tiga tahun terkahir terjadi 9,546 kasus penyebaran hoax yang terjadi di Indonesia. Dengan banyaknya kasus penyebaran hoax ini terutama yang terjadi pada media sosial twitter, makan akan diperlukan sebuah riset yang dapat melakukan pendeteksian hoax.

Riset mengenai deteksi hoax dalam social media telah dilakukan oleh beberapa peneliti, yaitu Munirul dkk pada tahun 2020 [3], I. Kencana Wintang dkk pada tahun 2020 [4], Ismayanti dkk pada tahun 2021[5], B.P Nayoga dkk pada tahun 2021 [6], dan P. N Anggreyani dkk pada tahun 2021[7]. Clasifier SVM dan TF-IDF digunakan dalam riset [3], clasifier feed-forward dan back-propagation dengan vektorisasi TF-IDF dan Word2Vec digunakan dalam riset [4], classifier Logistic Regression, Support Vector Machine (SVM), Random Forest dengan fitur ekpansi Word2Vec digunakan dalam riset [5], classifier Support Vector Machine (SVM) and Naive Bayes digunakan dalam riset [6], dan classifier LSTM dan CNN digunakan dalam riset [7].

Dalam riset [3] digunakan sebanyak 120 data dan data test sebanyak 30 data dan didapatkan hasil dengan menggunakan metode Support Vector Machine dengan pembobotan TF-IDF mendapatkan accuracy sebesar 60%. Dalam riset [4] didapatkan hasil feed-forward dan back-propagation dengan menggunakan vektorisasi TF-IDF meningkatkan performa tertinggi dibandingkan Word2Vec dengan akurasi 78.76%.TF-IDF bekerja lebih lama dibandingkan Word2Vec, namun hasil performa menunjukkan bahwa TF-IDF memberikan akurasi tertinggi. Dalam riset [5] data train yang digunakan sebanyak 21588 dan data test sebanyak 5396 didapatkan hasil dengan menggunakan metode Support Vector Machine dengan Word2Vec menghasilkan nilai akurasi tertinggi 87,34%. Dalam riset [6] digunakan 1000 dataset, model SVM memberikan nilai untuk akurasi bernilai 92,6%, presisi bernilai 92,48%, recall bernilai 92,59%, dan f1 macro bernilai 92,45% dan setelah menggunakan dropout model SVM mengalami sedikit peningkatan terhadap nilai akurasi, presisi, recall, dan f1 macro. Terakhir dalam riset [7] didapatkan hasil menunjukkan bahwa LSTM-CNN dapat mencapai akurasi 79.71% dengan menggunakan 16 unit dalam lapisan dengan menggabungkan dropout dan regularizer. enggunaan LSTM-CNN dengan Word2Vec dapat memproses data dalam jumlah besar untuk mendeteksi hoax.

Support Vector Machine (SVM) merupakan salah satu metode pada machine learning yang bekerja berdasarkan prinsip risiko struktural Minimization (SRM) bertujuan untuk menemukan hyperplane terbaik yang memisahkan dua kelas pada input [8]. Metode SVM dapat mengklasifikasikan masalah secara linier, namun saat ini SVM sudah berkembang bisa menyelesaikan masalah secara non-linier dengan mencari hyperplane yang optimal [9]. Dengan kata lain metode SVM ini sangat cocok digunakan untuk melakukan deteksi berita hoax yang akan memisahkan berita hoax dan non hoax

Dalam beberapa riset yang dilakukan untuk mendeteksi berita hoax terdapat salah satu metode yang berperan penting yaitu word2vec. Word2vec merupakan salah satu algoritma word embedding yang memetakan setiap kata dalam teks ke dalam vektor. Algoritma word2vec ini diciptakan oleh Mikolov dkk. pada tahun 2013 [10]. Prinsip kerja dari word2vec adalah memprediksi makna kata sesuai dengan peluang kemunculanya dan dapat melakukan asosisasi untuk menentukan hubungan kata dengan kata lainya[11]. Word2vec adalah model feedforward neural network yang terdiri dari sebuah hidden layer dan fully connected layer [10].

Word2Vec memiliki tiga parameter yang berpengaruh dalam proses model pembelajaran yaitu arsitektur, metode evaluasi, dan dimensi. Setiap jenis dari ketiganya parameter yang dimiliki Word2Vec memiliki pengaruh pada kinerja akurasi pembelajaran yang mendalam [12].

Untuk mendapatkan nilai akurasi yang bagus maka diperlukan pengukuran peforma terhadap system yang dibuat. Pengukuran peforma bisa dilakukan dengan metode Confusion Matrix. Confusion Matrix adalah salah satu bentuk pengukuran peforma untuk klasifikasi Machine Learning dimana keluaranya dapat berupa dua kelas ataupun lebih. Pengukuran menggunakan nilai akurasi (ketelitian dari kinerja sistem), precision (prediksi banyak data fakta yang diprediksi benar), recall l (seberapa benar hasil prediksi data yang tergolong true positive), dan f-measure (rerata dari perhitungan precission dan recall), tujuannya adalah untuk menganalisis dampak dari masing-masing faktor Dampak skenario preprocessing pada kinerja suatu model analisis [13].

Berdasarkan riset-riset di atas dapat disimpulkan bahwa belum ada ukuran dat train yang optimal dalam klasifikasi hoax. Oleh karena itu dalam riset ini akan dilakukan riset tentang penentuan jumlah data train yang optimal dalam klasifikasi hoax.

**Perumusan Masalah**

Berdasarkan latar belakang diatas, masalah yang akan diselesaikan oleh penulis dalam penelitian ini adalah bagaimana cara untuk menentukan jumlah data train yang optimal dalam deteksi hoax media berita Indonesia di twitter dan bagaimana performansi algoritma Support Vector Machine dengan Word2Vec untuk klasifikasi berita hoax dalam sosial media Twitter.

**Tujuan**

Tujuan yang ingin dicapai dalam Tugas Akhir ini adalah dapat membangun sebuah system pendeteksian hoax yang dapat mengetahui jumlah data train yang optimal untuk digunakan dan juga pada penelitian ini digunakan untuk melihat performansi algoritma Support Vector Machine dengan Word2Vec dalam pendeteksian berita hoax.