

BAB I

PENDAHULUAN

1.1 Latar Belakang

Di era informasi digital saat ini, akses mudah ke internet dan media sosial telah mengubah cara masyarakat terhubung dan terinformasi. Perkembangan teknologi ini telah memberikan kemudahan dan kecepatan dalam mendapatkan berbagai informasi. Media sosial memiliki peran penting dalam menyebarkan informasi dengan cepat dan banyak orang mengandalkannya sebagai sumber utama informasi. Namun, penggunaan media sosial dapat bervariasi, dengan beberapa memanfaatkannya untuk tujuan yang baik sementara yang lainnya memanfaatkannya untuk tujuan yang tidak baik. [1]. Salah satu dampak negatif yang timbul akibat penggunaan media sosial adalah penyebaran berita bohong atau *hoax* [2]. Penyebaran berita bohong atau *hoax* di media sosial oleh oknum yang tidak bertanggung jawab dapat memicu perselisihan dan dampak negatif lainnya. Menurut [20] ciri-ciri berita hoaks yaitu (1) Berita hoax dapat menimbulkan kecemasan, kebencian, dan permusuhan di Masyarakat. (2) Berita hoax menyebar cepat di media sosial tanpa jelas siapa penyebar awalnya, memicu perdebatan dan konflik. umumnya berita hoax berasal dari sumber tidak jelas, tidak terverifikasi, serta sering menyudutkan pihak tertentu. (3) Kontennya provokatif, bermuatan fanatisme ideologi, menyembunyikan fakta, dan kerap mencatut nama tokoh.

Salah satu metode yang dapat digunakan untuk mengatasi masalah tersebut adalah dengan menggunakan *Machine Learning*. Klasifikasi teks merupakan salah satu proses dalam *Machine Learning*, di mana dokumen-dokumen dalam satu set diberi kategori secara otomatis berdasarkan kategori yang telah ditentukan [3]. Metode yang bisa digunakan dalam klasifikasi teks adalah *Logistic Regression*, *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), dan lain-lain. Diantara metode tersebut, *Logistic Regression* merupakan metode yang populer karena dapat diinterpretasikan dan mempunyai hasil atau *output* yang bagus [4]. *Logistic Regression* adalah algoritma *Machine Learning* yang digunakan untuk klasifikasi, khususnya ketika data berbentuk biner (0 dan 1), yang berarti menentukan apakah kelas tersebut termasuk dalam satu kategori atau kategori lainnya. [5]. Penelitian

ini [6], membandingkan beberapa metode dalam klasifikasi berita palsu. Metode yang digunakan yaitu *Logistic Regression*, *K-Nearest Neighbour*, *Decision Tree* dan *Random Forest*. Diantara metode tersebut *Logistic Regression* mempunyai hasil akurasi yang tertinggi. Penelitian ini [7], menggunakan *Deep Learning* dan *Machine Learning* untuk mengklasifikasi data titanic disaster. Metode *Machine Learning* yang digunakan yaitu *K-Nearest Neighbour*, *Random Forest*, dan *Logistic Regression*, untuk *Deep Learning* metode yang digunakan yaitu *Long Short Term Memory* (LSTM), *Gated Recurrent Unit* (GRU) dan *Artificial Neural Network* (ANN). Pada metode *Machine Learning Logistic Regression* memperoleh akurasi yang paling tinggi sedangkan pada *Deep Learning LSTM* memperoleh akurasi tertinggi.

Dalam klasifikasi teks, ekstraksi fitur mempunyai peran yang penting. Ekstraksi fitur digunakan untuk mengubah kata-kata atau teks menjadi *vector* numerik agar dapat digunakan dalam komputasi [8]. Dengan mengubahnya menjadi *vector* numerik, model *Machine Learning* dapat melakukan klasifikasi teks dengan lebih efisien. Pada penelitian ini [8], menunjukkan perbandingan ekstraksi fitur dalam *Machine Learning* serta ekstraksi fitur TF-IDF memperoleh akurasi yang paling bagus saat dikombinasikan dengan *Naïve Bayes*, *SVM*, *KNN*, *Random Forest*, dan *Logistic Regression*. TF-IDF singkatan dari "*Term Frequency-Inverse Document Frequency*" menjaga hitungan kata serta *offset* untuk setiap kata dalam dokumen. Istilah "*term frequency*" merujuk pada jumlah kemunculan kata-kata dalam dokumen, yang berbanding lurus dengan bobot kata tersebut. Istilah "*inverse document frequency*" mengacu pada penyesuaian bobot kata-kata, di mana kata-kata yang muncul sangat sering dalam dokumen diberikan bobot yang lebih rendah dibandingkan dengan kata-kata lainnya [6].

Ensemble Learning adalah metode yang dapat membantu meningkatkan hasil *Machine Learning* dengan menggabungkan beberapa model, sehingga kinerja model prediksi dapat lebih ditingkatkan [9]. Metode *Ensemble Learning* secara efektif mengurangi kesalahan klasifikasi dan diyakini memiliki kinerja lebih baik dibandingkan dengan penggunaan model klasifikasi tunggal [10]. Salah satu metode *Ensemble Learning* yang dapat membantu meningkatkan hasil pembelajaran mesin adalah *Bootstrap Aggregating*. Bagging atau *Bootstrap*

Aggregating adalah algoritma yang meningkatkan akurasi model pembelajaran mesin dengan cara melatih algoritma pada beberapa sampel data yang diambil secara acak dan terpisah, lalu menggabungkan hasil model dari masing-masing sampel untuk menghasilkan prediksi yang lebih akurat [11]. Penelitian ini [10], membandingkan metode *Logistic Regression* tunggal dan *Ensemble Logistic Regression* dengan metode Bagging. *Ensemble Logistic Regression* menunjukkan hasil akurasi yang lebih tinggi dibandingkan *Logistic Regression* tunggal.

Dalam penelitian ini akan menggunakan *Ensemble Logistic Regression* dengan teknik Bagging dan TF-IDF sebagai ekstraksi fitur untuk klasifikasi berita *hoax*. *Logistic Regression* digunakan karena sangat baik untuk tugas klasifikasi biner, yaitu Ketika ada dua kelas yang harus diprediksi. *Ensemble Learning* digunakan karena memiliki kinerja lebih baik dibandingkan dengan model klasifikasi Tunggal. Bagging digunakan karena dapat menggabungkan hasil model untuk menghasilkan prediksi yang lebih akurat. TF-IDF digunakan karena memiliki hasil akurasi yang bagus saat dikombinasikan dengan *Logistic Regression*.

1.2 Rumusan Masalah

Berdasarkan latar belakang, rumusan masalah yang dapat diperoleh adalah sebagai berikut:

1.2.1 Permasalahan

Logistic Regression sering kesulitan menangkap hubungan non-linear (situasi di mana variabel *input* dan *output* tidak memiliki relasi yang dapat diwakili oleh garis lurus) dalam data berita *hoax*, sehingga menyebabkan akurasi rendah dan prediksi kurang optimal.

1.2.2 Solusi

Solusi dari permasalahan tersebut adalah dengan menggunakan Bagging *Ensemble Learning*. Bagging *Ensemble Learning* dapat meningkatkan akurasi *Logistic Regression* dengan mengurangi bias dan varian melalui pelatihan beberapa model pada subset data yang berbeda.

1.2.3 Pertanyaan Penelitian

Berdasarkan latar belakang, permasalahan, dan solusi dari penelitian ini, beberapa pertanyaan dapat dirumuskan sebagai tujuan dari disusunnya penelitian ini:

1. Berapa nilai parameter konvergensi, jumlah iterasi, dan Learning Rate terbaik dari *Logistic Regression* yang dapat menghasilkan akurasi tertinggi?
2. Berapa jumlah model *Bagging Ensemble Learning* terbaik pada *Logistic Regression* yang dapat menghasilkan akurasi tertinggi?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, tujuan yang didapatkan dari penelitian ini adalah sebagai berikut:

1. Untuk mengetahui nilai parameter terbaik dari konvergensi, jumlah iterasi, dan learning rate pada model *Logistic Regression* yang dapat menghasilkan akurasi tertinggi.
2. Untuk menentukan jumlah model terbaik pada metode *Bagging Ensemble Learning* dengan basis model *Logistic Regression* yang dapat memberikan akurasi tertinggi.

1.4 Manfaat Penelitian

Berdasarkan rumusan masalah di atas, manfaat yang didapatkan dari penelitian ini adalah sebagai berikut:

1. Menciptakan sistem yang efektif untuk mendeteksi berita *hoax* secara otomatis, yang dapat dimanfaatkan oleh media, platform sosial, dan pengguna individu untuk mengidentifikasi serta mengurangi penyebaran informasi palsu.
2. Sebagai pengalaman penulis dalam mempelajari metode *Ensemble Logistic Regression* dengan TF-IDF sebagai ekstraksi fitur untuk pembuatan sistem deteksi berita *hoax*.

1.5 Batasan Masalah

Penelitian ini memiliki batasan-batasan sebagai berikut:

1. Penelitian ini menggunakan dataset berita *hoax* berbahasa inggris dari Kaggle yang dapat diakses pada link berikut:
<https://www.kaggle.com/datasets/rchitic17/real-or-fake/download?datasetVersionNumber=1>
2. Metode yang digunakan untuk ekstraksi fitur adalah TF-IDF.
3. Metode yang digunakan dalam pemodelan adalah *Ensemble Logistic Regression*.
4. Metode *Ensemble Learning* yang digunakan adalah *Bootstrap Aggregating*.
5. Pengujian model hanya akan dilakukan dengan berfokus pada pengukuran akurasi.
6. Pengujian sistem hanya menggunakan Bahasa inggris

1.6 Sistematika Proposal

Sistematika pada Proposal ini adalah:

BAB 1 PENDAHULUAN

Pada bab ini berisi tentang Latar Belakang, Rumusan Masalah, Permasalahan, Solusi, Pertanyaan Penelitian, Tujuan Penelitian, Manfaat Penelitian, Batasan Masalah, dan Sistematika Proposal.

BAB 2 KAJIAN PUSTAKA

Pada bab ini berisi teori tentang Klasifikasi Teks, Peprocessing, TF-IDF, semua metode yang digunakan dalam penelitian ini, dan penelitian terkait.

BAB 3 METODOLOGI PENELITIAN

Pada bab ini berisi tentang Arsitektur Sistem, Skenario Pengujian, dan Dataset yang digunakan dalam penelitian.

BAB 4 HASIL DAN PEMBAHASAN

Pada bab ini berisi tentang Spesifikasi Perangkat, Pembuatan Program, Hasil Skenario Pengujian.

BAB 5 KESIMPULAN DAN SARAN

Pada bab ini berisi tentang kesimpulan dan saran dan penelitian yang dilakukan.