BAB 1

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi dalam penyebaran informasi terutama dalam menggunakan internet sudah berkembang dengan pesat. Internet memudahkan seseorang untuk menyebarkan dan mengakses berita atau suatu informasi dengan mudah. Hal ini membuat suatu berita dapat tersebar dengan cepat di kalangan masyarakat. Namun hal ini dapat disalahgunakan oleh seseorang untuk menyebarkan informasi yang salah atau *hoax*.

Hoax adalah informasi yang sesat dan berbahaya karena dapat menyesatkan persepsi seseorang dengan menyampaikan informasi palsu sebagai kebenaran (Rasywir & Purwarianti, 2016). *Hoax* bertujuan membentuk persepsi atau menggiring opini masyarakat dari fakta yang sebenarnya. Sepanjang sejarah, *hoax* dapat mempengaruhi banyak orang hingga menodai suatu citra dan kredibilitas korban (Chen, et al., 2014). Oleh karena itu, diperlukannya cara untuk menangkal berita *hoax* yang tersebar di dalam masyarakat dengan mengidentifikasi berita dan melakukan klasifikasi terhadap berita agar dapat mendeteksi berita *hoax* dan tidak.

Untuk dapat mendeteksi berita *hoax*, hal yang dapat dilakukan adalah dengan melakukan klasifikasi teks terhadap suatu berita atau dokumen. Klasifikasi teks adalah proses yang secara otomatis menempatkan dokumen teks ke dalam suatu kategori berdasarkan isi dari teks tersebut (Zhang, et al., 2009). Klasifikasi teks dilakukan dengan pendekatan *machine learning* dan dilakukan melalui beberapa tahap yaitu *text preprocessing*, *feature selection*, *classification* dan *evaluation*.

Penelitian ini bertujuan untuk mengklasifikasi berita *hoax* yang tersebar di masyarakat dengan menggunakan *machine learning*. Penelitian yang serupa telah dilakukan sebelumnya oleh Amelia Rahman, Wiranto dan Afrizal Doewes dengan judul "*Online News Classification Using Multinomial Naïve Bayes*" pada tahun 2017 dengan menggunakan algoritma *Multinomial Naïve Bayes* dalam klasifikasi berita daring. Berdasarkan hal tersebut, maka peneliti akan melakukan klasifikasi berita *hoax* dengan menggunakan algoritma *Multinomial Naïve Bayes* dan *Logistic Regression*.

Multinomial Naïve Bayes dipilih karena kelas dokumen tidak hanya ditentukan oleh kata yang muncul tetapi juga oleh jumlah kata-kata yang muncul sehingga cocok untuk klasifikasi (Witten & Frank, 2011). Lalu *Logistic Regression* dipilih karena dapat menyelesaikan klasifikasi biner dengan cukup baik karena prediksinya dalam nilai probabilitas dan bekerja dengan baik dengan teks yang panjang ataupun pendek (Mokhtar, et al., 2019).

Selain itu proses dalam menyeleksi fitur dilakukan menggunakan *Recursive Feature Elimination*. *Recursive Feature Elimination* dikenalkan oleh Guyon pada tahun 2002, dengan menghilangkan fitur yang memiliki peringkat paling rendah. Proses ini dilakukan untuk menghilangkan fitur yang tidak diperlukan dalam proses pembobotan. Pada tahun 2020 terdapat penelitian serupa dengan menggunakan RFE sebagai seleksi fitur dalam klasifikasi *newsgroup text* dan mendapatkan performa 83% saat menggunakan 60% fitur (Al-Ameer, 2020).

Memilih parameter yang tepat menjadi peran penting dalam menghasilkan model yang baik. Dengan jumlah parameter dari setiap algoritma yang cukup banyak, maka penting dalam mengetahui parameter yang paling optimal untuk

setiap algoritma. Penggunaan *hyperparameter* merupakan salah satu contoh untuk membantu model dalam menghasilkan parameter yang tepat. Penulis menggunakan *GridSearchCV* untuk menentukan parameter dan hasil yang optimal. Alasan *GridSearchCV* digunakan karena salah satu model yang paling banyak digunakan untuk optimasi *hyperparameter* dan kesederhanaan matematisnya (Bergstra & Bengio, 2012).

Performa algoritma *Multinomial Naïve Bayes* dan *Logistic Regression* akan dibandingkan sebelum dan sesudah menggunakan *Recursive Feature Elimination* untuk dilihat yang memiliki performa *score* terbaik. Hal ini untuk melihat pengaruh dari *feature selection* terhadap masing–masing algoritma dalam melakukan klasifikasi berita *hoax*. Lalu setelah itu akan dilakukan evaluasi menggunakan *F1-Score* untuk dilihat *score* yang dilakukan oleh masing–masing model. *Dataset* yang dikumpulkan berasal dari laman *Mendeley Data* yang dibuat oleh Faisal Rahutomo, Inggrid Yanuar, Rosa Andrie Asmara. *Dataset* dipilih karena memiliki jumlah yang besar dan label pada tiap data beritanya.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan di atas, maka dapat diambil rumusan masalah yaitu:

1. Bagaimana cara mengimplementasikan algoritma *Multinomial Naïve Bayes*, *Logistic Regression*, dan *Recursive Feature Elimination* untuk proses klasifikasi berita hoax?
2. Bagaimana pengaruh penurunan jumlah fitur menggunakan metode *Recursive Feature Elimination* terhadap tingkat performa klasifikasi dari

algoritma *Multinomial Naïve Bayes* dan *Logistic Regression* dalam klasifikasi berita hoax?

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini dapat disebutkan sebagai berikut:

1. Sistem dibuat hanya sebagai keperluan demonstrasi.
2. Pengukuran performa hanya menggunakan *F1-Score*, *precision* dan *recall*.
3. Metode ekstrasi fitur dilakukan dengan menggunakan *TF-IDF*.
4. Tahap *text preprocessing* yang dilakukan hanya *case folding*, *tokenizing* dan *stopwords removing*.

1.4 Tujuan Penelitian

Tujuan dalam penelitian ini dapat disebutkan sebagai berikut:

1. Untuk mengetahui cara mengimplementasikan algoritma *Multinomial Naïve Bayes, Logistic Regression* dan *Recursive Feature Elimination* dalam klasifikasi berita hoax.
2. Melihat pengaruh penurunan jumlah fitur menggunakan metode *Recursive Feature Elimination* terhadap performa algoritma *Multinomial Naïve Bayes* dan *Logistic Regression* dalam klasifikasi berita hoax

1.5 Manfaat Penelitian

Manfaat yang dapat diperolah dari penelitian ini adalah:

1. Dapat mengetahui pengaruh *Recursive Feature Elimination* terhadap *score* klasifikasi teks.
2. Bagi ilmu pengetahuan, dapat digunakan sebagai referensi untuk mengembangkan sistem yang lebih baik lagi.

1.6 Sistematika Penulisan

Sistem penulisan laporan skripsi ini dapat dijabarkan sebagai berikut:

BAB 1 PENDAHULUAN

Bab 1 Pendahuluan berisi latar belakang dari judul skripsi "Pengaruh Recursive Feature Elimination Terhadap Algoritma Multinomial Naïve Bayes Dan Logistic Regression Dalam Mendeteksi Berita Hoax", rumusan masalah, batasan masalah, tujuan masalah, manfaat masalah dan sistematika penulisan.

BAB 2 LANDASAN TEORI

Bab 2 Landasan Teori berisikan teori atau penjelasan mengenai *text preprocessing, TF-IDF, Multinomial Naïve Bayes, Logistic Regression, Recursive Feature Elimination, F1-Score* dan *Hyperparameter space.*

BAB 3 METODOLOGI PENELITIAN

Bab 3 Metodologi penelitian berisikan analisis kebutuhan yang menjelaskan kebutuhan di dalam penelitian dan perancangan aplikasi yang menjelaskan sistem dalam bentuk *flowchart*.

BAB 4 HASIL DAN DISKUSI

Bab 4 Hasil dan Evaluasi berisikan spesifikasi sistem yang digunakan dalam membangun sistem, potongan kode, uji coba yang dilakukan dengan beberapa skenario, hasil uji setiap skenario uji coba dan evaluasi dari hasil uji coba.

BAB 5 SIMPULAN DAN SARAN

Bab 5 Simpulan dan Saran berisikan simpulan dari hasil uji coba yang telah dilakukan dan saran untuk mengembangkan penelitian selanjutnya yang berhubungan menjadi lebih baik lagi.