BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Teknologi informasi berkembang pesat seiring dengan berjalannya waktu, efisiensi dan praktis menjadi sebuah tolak ukur akan perkembangan dalam teknologi informasi. industri media adalah salah satu industri yang sangat cepat perkembangannya, dahulunya, industri media bermula berupa fisik kertas seperti koran, majalah, poster. Adapun berupa gambar di layar televisi dan suara seperti radio, pada saat ini media sudah berkembang menjadi media *online* yang sangat menguntungkan, dengan begitu informasi dapat tersampaikan dengan cepat dan tanpa harus menunggu ataupun membeli, hanya membutuhkan perangkat elektronik berupa komputer, *laptop* ataupun telepon pintar sudah bisa mengakses media yang diinginkan.

Media *online* merupakan media komunikasi yang pemanfaatannya menggunakan perangkat internet. Oleh karena itu, media *online* tergolong media massa yang populer dan tergolong khas. Kekhasan media ini terletak pada keharusan untuk memiliki jaringan teknologi informasi dengan menggunakan perangkat komputer, di samping pengetahuan tentang program komputer untuk mengakses informasi atau berita (Suryawati, 2011).

Media *online* menyediakan berita yang bermacam ragam sesuai yang diinginkan, di dalamnya menyajikan berita saat ini yang sangat cepat terbitnya dibanding media berita yang belum menggunakan teknologi internet kebanyakan media berita *online* memiliki berberapa kategori berita seperti berita politik, berita olahraga, berita kesehatan, berita ekonomi hingga berita publik figur, di

dalam berberapa kategori ada yang mempunyai sub berita atau berita yang lebih menjurus akan tetapi dalam penjurusan pihak media masih memilih secara manual berita yang dikiranya memiliki unsur yang spesifik dan akan dimasukkan ke dalam sub-kategori yang sudah disediakan oleh pihak media berita.

PT Merah Putih merupakan salah satu perusahaan yang menyajikan berita *online* yang berupa website yaitu *merahputih.com*. yang di dalam website *merahputih.com* mempunyai berberapa kategori utama berita yaitu indonesiaku, hiburan dan gaya, dan olahraga. Di dalam sub berita indonesiaku terdapat berberapa sub-kategori yaitu tradisi, kuliner dan travel, mengelompokkan berita olahraga masih manual yang dilakukan oleh pihak redaksi, ini berpengaruh terhadap efesiensi kerja di mana saat berita olahraga banyak diterbitkan sekaligus maka pihak redaksi memerlukan waktu untuk memilah berita dan memasukkan ke dalam sub-kategori.

Penelitian yang menjadi acuan dalam penelitian ini adalah penelitian yang dilakukan oleh (Haristu, 2019).yang berjudul "Penerapan Metode Random Forest Untuk Prediksi *Win Ratio* Pemain *Player Unknown Battleground*". Dalam penelitian tersebut algoritma *random forest* (RF) digunakan untuk memprediksi win *ratio* berdasarkan data statistik pemain *Player Unknown Battleground*, menghasilkan tingkat akurasi yang baik dengan mendapatkan skor akurasi 88,19%, semakin banyak jumlah *tree* yiang digunakan maka semakin baik pula akurasi yang didapat.

Selain digunakan untuk memprediksi kemenangan, algoritma *random forest* juga sering digunakan untuk melakukan klasifikasi teks (Haristu, 2019).

Pada penelitian terdahulu (Fadilah, 2020), algoritma RF berhasil meraih performa F1 sebesar 93%.

Pada penelitian tersebut metode TF-IDF digunakan sebagai metode ekstraksi fitur. Namun, penelitian tersebut belum mengkaji pengaruh dari jumlah fitur yang digunakan terhadap performa dari mesin. Sedangkan pada ranah pembelajaran mesin, jumlah fitur yang besar membutuhkan sumber daya komputasi yang besar pula. Sumber daya komputasi yang besar tersebut dibutuhkan baik pada saat proses pelatihan model ataupun proses penggunaan (inferensi) model.

Berdasarkan alasan tersebut, dalam penelitian akan diuji metode *Recursive Feature Elimination* (RFE) untuk memperkecil sumber daya komputasi yang dibutuhkan. Penelitian ini bertujuan untuk mengetahui jumlah pengurangan fitur dan pengaruhnya terhadap tingkat performa milik model. Metode RFE dipilih karena penelitian terdahulu terkait penggunaan metode RFE terbukti dapat meningkatkan menghilangkan fitur-fitur yang kurang relevan.

Menurut (Wibawa & Novianti, 2017), Metode seleksi RFE pada dasarnya adalah proses rekursif yang meranking fitur berdasarkan tingkat pentingnya terhadap proses prediksi. Pada setiap iterasi, ranking pentingnya fitur diukur dan fitur yang kurang relevan dihilangkan.

1.2 Rumusan Masalah

Berdasarkan latar belakang, terdapat masalah yang akan dirumuskan dalam penelitian ini yaitu:

1. Bagaimana cara mengimplementasikan metode RFE untuk melakukan optimasi algoritma Random Forest dalam melakukan klasifikasi teks?

2. Bagaimana F1 score yang dihasilkan berdasarkan jumlah fitur yang dieliminasi menggunakan metode RFE?

1.3 Batasan Masalah

Dalam penelitian ini, terdapat batasan-batasan masalah berikut guna membatasi cakupan dari penelitian yang dilakukan:

1. Data serta kategori yang digunakan dalam penelitian diperoleh dari penelitian sebelumnya (Fadilah, 2020). Data diperoleh dengan melakukan proses *crawling* pada website merahputih.com atas seizin PT. Merah Putih selaku pemilik dan pengelola situs web.
2. Data yang diambil dari webite merahputih.com mencakup subkategoridari indonesiaku berupa tradisi, kuliner, dan travel.
3. Jumlah data berita yang diproses sejumlah 8300 data *text* yang berisikan judul, isi, dan kategori berita.
4. Performa pengukuran tingkat keberhasilan model akan dihitung melalui tingkat F1-Score.

1.4 Tujuan Penelitian

Tujuan dari penelitian adalah:

1. Mengimplementasikan RFE untuk optimisasi algoritma *Random Forest* dalam klasifikasi teks.

2. Mengetahui F1 score yang dihasilkan berdasarkan jumlah fitur yang dieliminasi menggunakan metode RFE.

1.5 Manfaat Penelitian

Menggunakan sistem yang lebih baru lebih efisien, akurat dan cepat dengan menggunakan algoritma Random Forest serta pemakaian TF-IDF dan metode *Recursive Feature Elimination* (RFE) pada kasus klasifikasi subkategoripada kategori berita Indonesiaku.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi "Implementasi Metode Random Forest dan Recursive Feature Elimination untuk Klasifikasi Berita" terdiri dari lima bab, yaitu pendahuluan**,** landasan teori, metodologi penelitian, hasil dan diskusi, dan simpulan dan saran.

1. BAB 1 PENDAHULUAN

   Bab pendahuluan terdiri dari enam bagian, yaitu latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, menfaat penelitian, dan sistematika penulisan laporan.

2. BAB 2 LANDASAN TEORI

   Bab landasan teori terdiri dari empat teori yang digunakan dalam penelitian, yaitu *Text Classification* dan *Text Pre-Processing*, *Decision Tree Learning*, *Term Frequency Inverse Document Frequency*, *Random Forest Classifier*, *F1-Score* dan *Recursive Feature Elimination*.

3. BAB 3 METODOLOGI PENELITIAN

Bab metodologi penelitian menjelaskan tentang metode penelitian yang digunakan, serta perancangan sistem.

4. BAB 4 HASIL DAN DISKUSI

Bab hasil dan diskusi berisi implementasi, hasil uji coba, dan evaluasi terhadap sistem yang telah dibuat.

5. BAB 5 SIMPULAN DAN SARAN

Bab simpulan dan saran berisi simpulan dari hasil penelitian dan eksperimen yang dilakukan dalam penelitian, serta saran yang dapat dilakukan untuk mengembangkan aplikasi maupun penelitian lebih lanjut.