1. Pendahuluan

Latar Belakang

Pada saat ini, industri perfilman berkembang pesat dan diminati oleh masyarakat di berbagai belahan dunia. Film merupakan seni berbentuk visual yang terus berkembang [1]. Dengan perkembangan teknologi saat ini, berbagai situs menginformasikan mengenai film yang sedang atau akan tayang, seperti Internet Movie Database (IMDB), Rotten Tomatoes, dan Metacritic. Untuk mengetahui menarik atau tidaknya suatu film kita perlu melihat ulasan dari penonton sebelumnya pada film tersebut. Beberapa ulasan mungkin terlihat jelas dikategorikan kedalam ulasan positif atau negatif, tetapi masih terdapat ulasan yang belum jelas untuk dikategorikan [2]. Selain itu, akan membutuhkan usaha yang lebih untuk mendapatkan informasi mengenai suatu film dari sekumpulan ulasan film tersebut [1].

Untuk membantu permasalahan tersebut, diperlukan analisis sentimen yang merupakan salah satu teknik otomasi dari *machine learning* untuk membantu pengklasifikasian. Analisis sentimen digunakan untuk mendapatkan informasi mengenai penilaian suatu film, apakah bernilai positif atau negatif dengan melakukan otomasi terhadap proses memahami, mengekstrasi dan pengolahan data yang berbentuk tekstual [3] [4].

Sudah banyak penelitian yang melakukan klasifikasi dengan berbagai metode, seperti penggunaan metode *K-Nearest Neighbor* [1], *Multinomial Naïve Bayes* [2], *Support Vector Machine* [5], dan masih banyak lagi. Penelitian ini melakukan proses pengklasifikasian dengan menggunakan metode *K-Nearest Neighbor* karena metode ini memiliki cukup keunggulan dan memiliki akurasi yang cukup baik bila dibanding metode pengklasifikasian lainnya [6]. *K-Nearest Neighbor* adalah metode pengklasifikasian yang cukup sederhana namun dibalik metodenya yang sederhana terdapat masalah utama mengenai dimensi fitur yang tinggi [7]. Fitur yang tinggi ini berasal dari keunikan kata yang terdapat pada dataset [8]. Pada penelitian [9] membuktikan penggunaan seleksi fitur *Gini* Index dapat mengatasi permasalahan tersebut. Jika dilihat dari segi performansi, seleksi fitur *Gini* Index menghasilkan klasifikasi yang lebih baik dari metode seleksi fitur lainnya [10].

Pada penelitian sebelumnya [9] menunjukkan penggunaan seleksi fitur *Gini Index* dapat meningkatkan performansi dari *K-Nearest Neighbor* itu sendiri.

Topik dan Batasannya

Berdasarkan pada penelitian sebelumnya, penelitian yang dilakukan akan menggunakan dataset yang berbeda yaitu menggunakan dataset Internet Movie Database (IMDB) dari Kaggle dengan total 15.000 ulasan berbahasa inggris yang terdiri dari 7500 ulasan positif dan 7500 ulasan negatif. Dataset ini selanjutnya akan digunakan dalam membandingkan teknik klasifikasi *K-Nearest Neighbor* dengan menggunakan seleksi fitur *Gini Index* dan tanpa seleksi fitur *Gini Index* yang digunakan dalam pengklasifikasian ulasan film serta melakukan pengujian untuk mengetahui performansi dari model berdasarkan skenario pengujian.

Tujuan

Tujuan dari penelitian ini adalah mengetahui cara membangun sistem pengklasifikasian ulasan pengguna dengan menggunakan metode *K-Nearest* Neighbor dan Gini Index dalam proses *feature seletion*. Selain itu, tujuan dari penelitian ini adalah untuk menganalisis performansi penggunaan metode *K-Nearest Neighbor* dan seleksi fitur *Gini Index*.