1. Pendahuluan

Dengan adanya platform dunia maya yang semakin beragam, rasa bersosialisasi manusia semakin mengalami peningkatan. Pada platform dunia maya, kita dapat menemukan banyak informasi seperti pandangan, perasaan, dan opini pada topik tertentu [1]. Lebih dari 80% informasi yang ada disimpan dalam bentuk teks, sehingga penambangan teks diyakini memiliki potensi yang lebih tinggi dibandingkan dengan penambangan data. Penambangan teks dapat mengekstrak informasi pada dokumen yang tidak terstruktur. *Sentiment analysis* merupakan cabang baru pada penambangan teks, yang terdiri dari kegiatan mengekstraksi, memproses dan mengevaluasi data dalam bentuk teks [2]. *Sentiment analysis* mendeteksi polaritas pada sebuah teks, kalimat, paragraf, maupun dokumen. *Sentiment analysis* dikenal juga sebagai *opinion mining* [1] yang diartikan sebagai teknik untuk pemrosesan bahasa alami untuk mengevaluasi posisi, sensitivitas, atau penilaian orang tentang subjek, produk, atau topik tertentu. Dengan memahami emosi pelanggan di dunia maya melalui ulasan yang diberikan, kita dapat melakukan analisis terhadap umpan balik dari para pelanggan tentang produk maupun jasa yang kita berikan dan melakukan peningkatan terhadap produk maupun kinerja sesuai dengan apa yang pelanggan butuhkan[3].

Opinion mining merupakan cabang penelitian yang sesuai untuk menganalisis dokumen terkait ulasan yang jelas mengekspresikan opini pada sebuah film, seperti pada penelitian yang dilakukan Sharma S P et al.[4]. Opini pada sebuah film berisi informasi tentang ulasan dengan berbagai macam emosi dari para penonton film. Diperlukan waktu yang banyak untuk menganalisis emosi pada setiap ulasan yang memiliki banyak kata yang tidak dibutuhkan atau kata yang berulang. Banyak penelitian tentang sentimen dengan menerapkan standar *feature selection* untuk meningkatkan performansi komputasi. Diantaranya, peneliti terkenal Annett et al.[5], menghapus kalimat objektif pada *testbed* yang terdiri dari teks objektif dan subjektif yang dilatih pada SVM.

Feature selection sangat dibutuhkan pada penelitian ini, karena *movie review* memiliki banyak sekali *unique feature* seperti judul film dan nama aktor. Dataset yang digunakan memiliki 50.000 data dan opini para penonton film pasti berbeda-beda yang menyebabkan semakin banyaknya fitur diambil dari dataset tersebut. *Feature selection* adalah proses pemilihan fitur yang relevan untuk menghasilkan performansi yang optimal. *Information Gain* (IG) adalah salah satu metode seleksi fitur terbaik dan dapat mereduksi lebih dari 90% fitur yang tidak relevan [6,7].

Penelitian ini menggunakan dataset Internet Movie Database (IMDB) dari Kaggle dataset [11] dengan total 25.000 dokumen untuk ulasan positif dan 25.000 dokumen untuk ulasan negatif. Penelitian dilakukan untuk menganalisis pengaruh *feature selection* terhadap dataset IMDB untuk mengetahui *accuracy* terbaik yang didapatkan pada klasifikasi sentimen menggunakan Decision Tree(DT). *Feature selection* yang digunakan adalah IG, pemilihan fitur dilakukan setelah fitur melalui proses *feature extraction*. *Feature extraction* yang digunakan adalah Term Frequency Inverse Document Frequency (TF-IDF) yang dibagi menggunakan teknik N-gram menjadi beberapa jenis fitur, yaitu unigram, bigram, trigram, kombinasi dari unigram dan bigram, kombinasi dari bigram dan trigram, dan kombinasi dari unigram dan bigram dan trigram [9]. Performansi dari metode tersebut dievaluasi menggunakan parameter *precision, recall, f1-score,* dan *accuracy.* Hasil dari penelitian ini adalah nilai *accuracy* tertinggi yang didapatkan akan dibandingkan dengan metode lainnya pada penelitian terkait.