

## 1. Pendahuluan

### Latar Belakang

Saat ini orang menggunakan *Internet* untuk menulis opini mereka di *blog*, jejaring sosial, dan situs *web*. Analisis sentimen atau *opinion mining* menjadi sangat penting bagi pengguna dan juga untuk bisnis [1]. Situs *review* seperti IMDb adalah salah satu situs yang sering dikunjungi oleh pengguna *Internet*, situs tersebut menyediakan informasi yang sangat lengkap mengenai perfilman mulai dari aktor, kru, sinopsis, peringkat dan ulasan film yang diberikan oleh pengguna lain [2][3]. Ulasan dan peringkat dari film tersebut dapat mempengaruhi perilaku orang-orang untuk pergi ke teater/bioskop atau untuk membeli sebuah produk.

Klasifikasi sentimen adalah sub divisi baru dari klasifikasi teks yang memperhatikan topik dan pendapat yang diungkapkan pada sebuah dokumen. Analisis statistik dari sentimen dapat memberikan gambaran yang kuat terhadap berita baru yang mempengaruhi sebuah entitas yang penting [4]. Dalam memecahkan permasalahan yang ada pada klasifikasi sentimen, pendekatan *machine learning* memiliki performansi yang bagus akan tetapi memakan banyak waktu dalam memproses model agar sesuai dengan data latih. Penggunaan kombinasi dari *machine learning* dan *feature selection* dapat mengurangi waktu yang diperlukan dan meningkatkan performansi dari *machine learning* [5].

Tim O'Keefe dan Irena Koprinska meneliti pengaruh *feature selection* dan *weighting method* pada analisis sentimen menggunakan klasifikasi *Naïve Bayes* (NB) dan *Support Vector machine* (SVM) [6]. Dari penelitian tersebut didapatkan hasil akurasi tertinggi 87.15% dengan menggunakan *Categorical Feature Difference* (PD) sebagai *feature selection* dan SVM sebagai metode klasifikasi. Penelitian selanjutnya, meneliti penggunaan *Information Gain* (IG) dengan 4 metode klasifikasi yaitu NB, SVM, *Random Forest* (RF) dan *K-Nearest Neighbour* (KNN). Dari keempat metode klasifikasi tersebut, SVM mendapatkan akurasi terendah tanpa menggunakan IG [5]. Namun dengan menggunakan IG dapat terlihat peningkatan akurasinya.

Metode klasifikasi SVM terbukti mengungguli sebagian besar metode klasifikasi lain pada data text [7]. Namun, *performance* dari *classifier* bersangkutan dengan *training corpus*-nya, dan *training corpus* yang baik dapat memperoleh *classifier* dengan *performance* yang baik [8]. Oleh karena itu, penelitian ini berfokus pada seleksi fitur *Information Gain* dan *n-grams* menggunakan klasifikasi SVM terhadap klasifikasi *sentiment movie review* untuk melihat pengaruh penggunaan pemilihan fitur dalam menentukan sentimen dari *movie review*.

### Topik dan Batasannya

Topik penelitian yang diambil penulis adalah pengaruh *Information Gain* terhadap model evaluasi *Support Vector Machine* dan *Naïve Bayes* yang menggunakan *N-Grams TF-IDF* pada dataset *movie review* untuk mendapatkan nilai akurasi dari hasil evaluasi. Batasan yang terdapat pada penelitian ini yaitu: Pertama, *feature* yang akan digunakan pada tahap ekstraksi dan seleksi di batasi menjadi 3000 *feature*. Dikarenakan keterbatasan hardware ataupun *cloud service* yang digunakan penulis. Kedua, *N-grams* yang akan digunakan yaitu *unigram* dan *bigram*, untuk *trigram* dan seterusnya tidak dapat diaplikasikan karena dibutuhkan waktu yang cukup lama untuk melakukan proses *feature* seleksinya.

### Tujuan

Tujuan yang akan dicapai dalam penelitian tugas akhir ini adalah menerapkan *Support Vector Machine* dalam mengklasifikasikan teks ulasan film serta mengukur performansi dari model yang dibangun menggunakan *Information Gain* dan *n-grams*.

### Organisasi Tulisan

Urutan penulisan selanjutnya adalah studi terkait yang merupakan penjelasan singkat mengenai istilah-istilah yang digunakan. Berikutnya, adalah sistem yang dibangun berisi mengenai detail penjelasan metode dan tahapan-tahapan sistem yang dibangun. Selanjutnya, pada evaluasi berisi hasil dari pengujian sistem dan analisis dari sistem tersebut. Pada kesimpulan berisi rangkuman hasil yang diperoleh dan saran untuk penelitian selanjutnya. Daftar Pustaka berisikan sumber literatur penunjang dari penelitian tugas akhir ini.