

1. Pendahuluan

Latar Belakang

Seiring dengan berkembangnya teknologi terutama internet, kita dapat menemukan banyak informasi seperti pandangan, pendapat, perasaan, dan penilaian pada topik tertentu. Salah satunya adalah *movie review* yang berisi pendapat atau opini seorang penonton terhadap suatu film. Pengguna yang terhubung melalui sosial media, forum, maupun blog di Internet, dapat mengungkapkan perasaan, penilaian, atau pendapat terhadap suatu film [1] yang dapat menjadi acuan bagi pengguna lainnya yang hendak menonton suatu film.

Analisis sentimen adalah studi komputasi dari opini-opini, sentimen, serta emosi yang diekspresikan dalam teks [2]. Dengan memahami respons atau tanggapan yang diberikan pengguna, kita dapat memperbaiki atau meningkatkan produk maupun kinerja sesuai dengan yang diutarakan oleh pengguna. Penelitian pada analisis sentimen, terdapat beberapa pendekatan pembelajaran mesin salah satunya adalah *Support Vector Machine (SVM)*. Penelitian yang dilakukan [3] menemukan bahwa algoritma SVM menjadi metode yangungguli metode lainnya untuk memprediksi sentimen dari sebuah ulasan.

Pada tahap ekstraksi fitur, penelitian ini menggunakan metode TF-IDF karena menurut [4], TF-IDF merupakan salah satu algoritma yang umum digunakan pada sektor informasi retrieval, yang digunakan sebagai suatu strategi pada klasifikasi teks atau suatu dokumen. Penelitian terhadap sentimen analisis pada movie review, biasanya menerapkan standar penyeleksian fitur yang bertujuan agar meningkatnya hasil performansi dari model yang dibuat. Salah satu nya, [5] menghapus kalimat objektif pada dataset yang terdiri dari teks objektif dan subjektif yang dilatih pada SVM. Kemudian penelitian tentang penerapan penyeleksian fitur menggunakan Chi Square pada dokumen sentimen analisis [6] menyimpulkan bahwa penggunaan Chi Square untuk penyeleksian fitur meningkatkan hasil dari accuracy sebanyak 13.33%.

Kemudian, penelitian selanjutnya mengenai penggunaan Chi Square dan metode SVM pada movie review sebelumnya pernah dilakukan [7] pada dataset yang didapat dari UCI Repository. Penelitian tersebut menyeleksi fitur dengan nilai K terbaik dari Chi Square yang didapat. Dari penelitian tersebut, nilai K yang memberikan hasil performa terbaik adalah sebesar 212. Yang berarti, term yang berada pada urutan 212 teratas memiliki nilai Chi Square yang tinggi sehingga fitur tersebut dipilih dan terbukti memberikan hasil yang baik pada nilai accuracy. Dengan itu, penelitian ini akan melakukan penyeleksian fitur dengan metode yang berbeda dari penelitian sebelumnya, yakni dengan menerapkan uji hipotesis pada nilai p dari nilai Chi Square yang didapat dari setiap fitur yang sudah diekstrak menggunakan TF-IDF. Proses penyeleksian tersebut adalah dengan memilih term yang memiliki nilai p yang kurang dari tingkat signifikansi yang ditetapkan, kemudian fitur terpilih akan dilatih menggunakan algoritma Support Vector Machine dengan kernel yang berbeda.

Topik dan Batasannya

Topik yang dibahas pada penelitian ini adalah pengaruh penyeleksian fitur menggunakan Chi-Square, dan algoritma pengklasifikasi Support Vector Machine pada dataset movie review. Batasan pada penelitian ini adalah dataset yang merupakan movie review berbahasa inggris yang berasal dari situs IMDb yang diunduh pada situs Kaggle.com, jumlah dataset yang digunakan sebanyak 5000 data, yang terdiri dari 2500 data yang memiliki ulasan positif dan 2500 data yang memiliki ulasan negatif.

Tujuan

Tujuan dari penelitian tugas akhir ini adalah menganalisis hasil performansi yang didapatkan setelah menyeleksi fitur menggunakan Chi Square dengan memilih term sesuai dengan tingkat signifikansi yang ditetapkan dan pengaruh penggunaan kernel yang berbeda pada algoritma SVM pada dataset ulasan film berbahasa Inggris.

Organisasi Tulisan

Bagian selanjutnya merupakan bab 2 yang membahas tentang studi literatur yang berkaitan dengan penelitian ini. Selanjutnya pada bab 3 membahas tentang teori dan rancangan dari sistem penelitian. Bab 4 merupakan bahasan tentang hasil dan analisis dari penelitian. Lalu, kesimpulan dari penelitian yang dilakukan dibahas pada bab 5