1. Pendahuluan

Latar Belakang

Analisis sentimen merupakan bidang studi yang menganalisis opini, sentimen, evaluasi, penilaian, tingkah laku, dan emosi masyarakat terhadap suatu produk, layanan, organisasi, individu, masalah, peristiwa, atau topik dari suatu hal [1]. Pada masa digital seperti sekarang ini, analisis sentimen menjadi sangat relevan dalam berbagai macam bidang, seperti penilaian produk, peringkat movie, dan respon pengguna di media sosial. Analisis sentimen pada movie review memiliki potensi untuk memberikan wawasan berharga kepada industri film, produser, dan penonton, dengan membantu mereka dalam memahami bagaimana respon penonton terhadap suatu movie [2].

Salah satu dataset yang sangat berharga untuk analisis sentimen pada movie review adalah dataset IMDb. Internet Movie Database (IMDb) adalah database film terbesar dan terlengkap, yang menawarkan database ekstensif mengenai film, acara TV, dan informasi pemeran [3]. Namun, melakukan analisis sentimen movie review tidak dapat dikatakan mudah.

Dataset IMDb mencakup tantangan yang perlu diatasi dalam konteks analisis sentimen. Movie review sering mengandung bahasa yang kompleks, kosakata yang bervariasi, kalimat yang tidak terstruktur, dan lain sebagainya [4]. Tantangan ini juga mencakup penggunaan sumber daya dan waktu yang dibutuhkan dalam melakukan pemrosesan teks. Oleh karena itu, diperlukan pendekatan analisis sentimen yang efektif.

Salah satu metode yang dapat digunakan dalam penelitian ini adalah algoritma klasifikasi Naive Bayes. Algoritma Naive Bayes telah terbukti efektif dalam banyak kasus analisis sentimen karena sederhana, cepat, dan mampu memberikan hasil yang baik [14]. Namun, penerapannya dalam konteks movie review pada dataset IMDb memerlukan penilaian yang lebih lanjut.

Pada penelitian ini algoritma Naïve Bayes yang digunakan ialah Multinomial Naïve Bayes, algoritma ini cocok untuk data diskrit seperti teks karena bekerja dengan asumsi bahwa setiap atribut atau kata dalam teks diambil dari distribusi multinomial [5]. Selain itu digunakan metode TF-IDF pada vectorization untuk mengukur pentingnya kata-kata dalam review, dan chi-squared untuk memilih fitur-fitur (kata-kata) yang paling relevan dalam klasifikasi sentimen.

Penelitian ini akan mengevaluasi kinerja algoritma klasifikasi Multinomial Naïve Bayes menggunakan metrik evaluasi yang terdiri dari accuracy, precision, recall, dan F-1 score dalam mengatasi tantangan-tantangan yang mungkin muncul dalam analisis sentimen movie review.

Dengan melakukan penelitian ini, diharapkan akan ada kontribusi berharga dalam pemahaman analisis sentimen dalam konteks movie review pada dataset IMDb, seperti tantangan yang ada dan fitur yang relevan pada analisis sentimen, serta pemahaman dalam penggunaan algoritma klasifikasi Multinomial Naïve Bayes secara efektif. Penelitian ini juga dapat memberikan wawasan untuk industri movie, yang dapat digunakan untuk mengambil keputusan yang lebih baik dalam produksi dan pemasaran movie.

Tujuan

Salah satu tujuan utama penelitian ini adalah untuk mengukur kinerja algoritma klasifikasi Naive Bayes, khususnya Multinomial Naïve Bayes dalam konteks analisis sentimen movie review pada dataset IMDb. Hal ini dilakukan dengan mengukur hasil akurasi, presisi, recall, dan F1-score terhadap model untuk mengevaluasi efektivitas algoritma dalam mengklasifikasikan sentimen. Selain itu, digunakan juga confusion matrix untuk merangkum prediksi model klasifikasi dan membandingkannya dengan hasil actual.

Tujuan selanjutnya adalah untuk menganalisis dan mengidentifikasi parameter yang digunakan dalam tahap pre-processing, yaitu vectorization menggunakan TF-IDF, dan feature selection menggunakan chi-squared untuk menemukan nilai optimal yang dapat meningkatkan kinerja algoritma.

Penelitian ini juga bertujuan untuk mengidentifikasi dan menggambarkan tantangan yang muncul dalam analisis sentimen movie review pada IMDb. Ini mencakup penggunaan bahasa yang tidak formal dan ekspresif, serta penggunaan bahasa metaforis, idiom, atau kata-kata khusus yang mungkin muncul. Review juga dapat terdiri dari kalimat yang tidak terstruktur, dan menggunakan tata bahasa tidak standar.

Selain itu, penelitian ini bertujuan untuk menganalisis fitur-fitur yang paling relevan, yang memiliki kontribusi atau hubungan erat terhadap sentimen dalam dataset. Ini dapat memberikan sedikit wawasan terkait kata-kata tertentu yang mungkin mempengaruhi hasil analisis sentimen.

Melalui pencapaian tujuan-tujuan ini, penelitian ini diharapkan dapat memberikan pemahaman yang lebih baik mengenai analisis sentimen movie review pada IMDb, meningkatkan penggunaan algoritma Naive Bayes, serta sedikit wawasan terkait pengaruh TF-IDF dan chi-squared dalam pre-processing.