# Sentiment Analysis: Perbandingan Algoritma Klasifikasi dengan Representasi TF-IDF

Ilham Aulady Miftakhurrizqy1

Mahasiswa Universitas Muhammadiyah Malang

Malang, Indonesia

a)ilhamam462@webmail.umm.ac.id

**Abstract.** Dunia kerja dalam membuat film adalah pekerjaan yang bergantung pada review yang ditulis oleh para penonton pada film yang dibuat. Karena itu dibutuhkan adanya analisa sentiment untuk melihat kepuasan penonton terhadap film yang dibuat. Analysis Sentiment dengan pendekatan machine learning merupakan salah satu teknik yang sangat populer dalam melakukan Sentiment Analysis. Pendekatan ini dilakukan dengan membuat sebuah model untuk memprediksi bagaimanakah perasaan dan emosi dari penonton. Dalam paper ini kami membuat model machine learning dengan algoritma multinomial naïve bayes, random forest classifier dan linear regression. Dengan dataset review film Squidgames, MNB mendapatkan akurasi sebesar 74%, RFC mendapat akurasi sebesar 74% dan LR mendapat akurasi sebesar 75%.

# Pendahuluan

Manusia adalah makhluk yang subjektif dan mereka memiliki pendapat yang berbeda-beda mengenai suatu hal. Pendapat yang diutarakan bisa menjadi kritik yang baik atau kepuasan terhadap karya orang, tetapi pendapat tersebut juga bisa menjadi kritik yang tidak baik atau celaan yang dapat menjatuhkan suatu karya. Seperti dalam sebuah review film, pasti ada pendapat yang mengatakan puas akan sebuah film, jeleknya film dan terkadang ada yang bersikap normal saja.

Pendapat pada review film bisa berpengaruh terhadap kinerja para karyawan suatu perusahaan film. Misalnya pada rating pekerja. Sebuah studi mengatakan sukses atau tidaknya film itu berdasarkan pada reviewnya [1]. Jadi sangat penting untuk dapat mengklasifikasi dan menganalisa review dari para penonton.

Klasifikasi dari review bisa dibedakan menjadi positif, negatif dan neutral atau biasa saja. Klasifikasi review ini berhubungan dengan pemilihan kata pada teks review, sepeti jika kata tersebut sudah terpakai dalam konteks positif atau konteks negatif. Klasifikasi ini dapat membantu untuk memahami proses Sentiment Analysis dimana Sentiment Analysis sendiri itu penting untuk memahami yang diinginkan penonton.

Sentiment Analysis adalah sebuah cabang pada keilmuan *Natural Language Processing* atau NLP. NLP merupakan sebuah area penelitian dan aplikasi yang mengeksplor bagaimana sistem bisa digunakan untuk memahami dan memanipulasi bahasa teks untuk melakukan sesuatu yang berguna [2]. Hasil dari sentiment analysis dapat digunakan untuk mengetahui baik dan buruknya konteks dari suatu teks.

Proses dalam melakukan Sentiment Analysis adalah dengan melakukan data cleaning, tokenizing, vectorizing, data balancing, pembagian data uji dan data latih, dan klasifikasi. Data cleaning, merubah representasi data, data balancing, pembagian data uji dan data latih merupakan cara dalam data teks preprocessing. Data cleaning meliputi membersihkan dataset dari baris yang kosong, mengubah semua huruf menjadi huruf kecil, menghapus semua tanda baca, dan stemming atau lemmatizing. Stemming adalah proses menghapus imbuhan untuk mengubah suatu kata menjadi kata dasar [3]. Lemmatizing adalah mengubah kata dengan makna yang sama menjadi satu bentuk [4]. Setelah itu tokenizing, tokenizing adalah proses untuk mensegmentasi dokumen menjadi unit kata, nomor dan tanda baca [1]. Vectorizing adalah proses untuk merubah kata-kata menjadi sebuah vector. Data balancing adalah proses untuk menyeimbangkan data dari setiap label sentiment.

Klasifikasi adalah salah satu aspek yang penting dalam data mining, dimana membuat sebuah model machine learning untuk memberi label pada data uji. Cara pengklasifikasian yaitu dengan berpatok pada data latih yang digunakan untuk belajar oleh model machine learning. Model akan memprediksi inputan data yang baru dengan memberi label (positif atau negatif) berdasarkan data latih. Adapun beberapa algoritma klasifikasi adalah Naïve Bayes, k-Nearest Neighbors, Support Vector Classifier, Random Forest, Decision Tree, dan lain-lain.

Pada paper ini kami akan membandingkan performa dari beberapa algoritma klasifikasi dengan representasi teks TF-IDF pada sebuah dataset Squid\_Games\_Review. Algoritma yang akan kami gunakan adalah Multinomial Naïve Bayes, Random Forest Classifier dan Linear Regression. Tujuan kami membandingkan algoritma-algoritma ini adalah kami ingin melihat apakah algoritma bisa membuat perbedaan yang signifikan pada model machine learning.

# Metode

Flowchart dari sentiment analysis dalam paper ini ditunjukan pada Gambar 1. Ada tiga tingkatan dalam sistem ini, yakni preprocessing, representasi teks lalu klasifikasi dengan MNB, RFC dan LR. Setiap review akan diklasifikasikan dalam tiga kelas, positif, negatif dan normal.

## 

Gambar 1. Flowchart Project

## Dataset

Dataset yang kami gunakan diambil dari website Kaggle. Dataset ini memiliki entitas sebanyak 1185 dengan kolom sebanyak enam kolom, Unnamed: 0, User\_name, Review title, Review Rating, Review date, dan Review\_body. Unnamed: 0 adalah index dari dataset, User\_name adalah nama pengguna yang mengirimkan review, Review title adalah judul dari review yang ditulis pengguna, Review Rating adalah rating yang diberikan oleh user, Review date adalah tanggal saat user menuliskan review, dan Review\_body adalah teks dari review user. Pada dataset ini kami hanya menggunakan kolom Review\_body dan Review\_Rating. Berikut adalah gambar dari tabel dataset tersebut.

## 

Gambar 2. Dataset Squid\_Games\_Review

## Preprocessing

Preprocessing yang kami lakukan dengan membersihkan kolom yang tidak diperlukan seperti Unnamed: 0, User\_name, Review title, dan Review date. Kolom tersebut tidak diperlukan karena sentiment analysis yang akan kami lakukan cukup dengan teks dan rating dari user. Selanjutnya kami membuat kolom untuk melihat sentiment dari user itu positif, normal atau negatif. Hal ini dilakukan dengan cara melihat pada bagian Review Rating.

## 

Gambar 3. Pseudocode dalam menentukan Sentiment

Setelah menetukan sentiment, kami melakukan cleaning pada text dengan menghapus tanda baca, mengubah semua huruf menjadi huruf kecil dan melakukan stemming.

## Representasi Teks

Setelah melakukan preprocessing, kami menggunakan representasi teks dengan menggunakan TF-IDF. TF-IDF atau *Term Frequency Inverse Document Frequency* adalah representasi teks dengan menghitung nilai setiap kata dalam sebuah dokumen melalui proporsi invers dari banyaknya kata dalam sebuah dokumen di mana kata itu muncul [5]. Dengan mengubah data teks menjadi representasi vector TF-IDF sistem dapat dengan mudah melakukan klasifikasi dengan algoritma klasifikasi. Berikut merupaka formula TF-IDF.

adalah variabel frekuensi dari kata, berisi 0 jika tidak ada dan 1 jika ada, merupakan hitungan mentah. merupakan banyaknya dokumen dalam corpus. adalah nomor dokumen dimana kata muncul.

## Sentiment Analysis

Setelah merepresentasikan data menjadi TF-IDF, kami melakukan klasifikasi dengan algoritma Multinomial Naïve Bayes, Random Forest Classifier dan Linear Regression. Multinomial Naïve Bayes merupakan classifier yang berbasis pada teorema Bayes. Teorema ini melakukan perhitungan dengan berdasar pada peluang [6]. Random forest classifier merupakan algoritma klasifikasi yang berbasis pada model tree. Model tree membagi dataset menjadi dua group berdasarkan kriteria hingga kondisi stop. [7] Lalu linear regression, Linear regression adalah algoritma klasifikasi dengan membuat asumsi secara linear.

# Hasil

Seperti dalam tabel 1, Algoritma RFC memiliki F-1 Score tertinggi dari yang lain. Perbedaan f-1 score pada project ini tidak besar kurang lebih hanya 0.002 saja. Meskipun memiliki perbedaan yang kecil, bisa dikatakan jika algoritma dari klasifikasi juga bisa mempengaruhi performa dari model machine learning. Meski memiliki f-1 score terbesa algoritma RFC memiliki presisi yang lebih tinggi dari pada LR dan MNB. MNB pada kasus ini memiliki performa yang paling rendah daripada ketiganya.

|  |  |  |  |
| --- | --- | --- | --- |
| **Algoritma** | **Precision** | **Recall** | **F-1 Score** |
| MNB | 76% | 74% | 74% |
| RFC | 79% | 74% | 74% |
| LR | 78% | 75% | 75% |

Tabel 1. Hasil Klasifikasi

## 

Gambar 4. Bar chart perbandingan algoritma

# Kesimpulan

Dalam project ini bisa disimpulkan jika model machine learning yang berbeda bisa berpengaruh pada performa model. Seperti model linear regression pada kasus ini memiliki f-1 score yang lebih besar daripada yang lain. Tetapi perbedaan pada algoritma tidak memiliki perbedaan yang signifikan.

Dalam project kedepannya kami bisa menambahkan representasi teks lainnya seperti GloVe, Word2vec, fasttext, dan sebagainya sebagai perbandingan tambahan. Dengan menambahkan representasi teks yang lebih kami bisa melihat apakah representasi teks juga akan memberikan perbedaan yang signifikan pada performa model.

# Daftar pustaka

|  |  |
| --- | --- |
| [1] | M. Yasen and S. Tedmori, "Movies Reviews Sentiment Analysis and Classification," *international conference on bangla speech and language processing (ICBSLP),* pp. 1-4, 2019. |
| [2] | Chowdhury and G. G., "Natural Language Processing," *Annual Review of Information Science and Technology,* pp. 51-89, 2003. |
| [3] | J. B. Lovins, "Development of a Stemming Algorithm," *Mechanical Translation and Computational,* vol. 11, p. 22–31, 1968. |
| [4] | T. ERJAVEC and S. EROSKI, "MACHINE LEARNING OF MORPHOSYNTACTIC STRUCTURE: LEMMATIZING UNKNOWN SLOVENE WORDS," *Applied artificial intelligence,* vol. 18, no. 1, pp. 17-41, 2004. |
| [5] | J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," *Proceedings of the first instructional conference on machine learning,* vol. 242, no. 1, pp. 29-48, 2003. |
| [6] | I. Rish, "An empirical study of the naive Bayes classifier," *IJCAI 2001 workshop on empirical methods in artificial intelligence,* vol. 3, no. 22, pp. 41-46, 2001. |
| [7] | M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical," *The Stata Journal,* vol. 20, no. 1, p. 3–29, 2020. |