# Sentiment analysis on movie reviews using Information gain and K-nearest neighbor

**Novelty Octaviani Faomasi Daeli**[1], **Adiwijaya**[2]
[1,2]School of Computing, Telkom University, Bandung, 40257, Indonesia
E-mail: [1]noveltyoctaviani@student.telkomuniversity.ac.id,
[2]Adiwijaya@telkomuniversity.ac.id

**Abstract.** The huge resources need effectiveness and efficiency, it can be processed by machine learning. There have been many studies conducted using machine learning method and produced quite good performance in sentiment analysis. Some machine learning methods that are often used in general are Naive bayes (NB), K-nearest neighbor (KNN), Support vector machine (SVM), and Random forest methods. Mostly, KNN did not achieve better performance than other machine learning methods in sentiment analysis. In this study, the Polarity v2.0 from Cornell movie review dataset will be used to test KNN with Information gain features selection in order to achieve good performance. The purpose of this research are to find the optimum K for KNN and compare KNN with other methods. KNN with the help of Information gain feature selection becomes the best performance method with 96.8% accuracy compared to the NB, SVM, and Random forest while the optimum K is 3.

## 1. Introduction

Movie is a visual art that continues to grow and multiply from year to year. Through movie review, viewers can find out which films have a good quality. The higher number of films produced will make many reviews being produced. It will need much effort for viewers to read a lot of movie reviews, so they can get an information about the movie. Based on these condition, sentiment analysis in movie reviews is an interesting topic to be solved by machine learning. Machine learning can help in term of effectiveness and efficient, because it will automatically classify and shorten the processing time [1].

Machine learning method that will be used is K-nearest neighbor (KNN). KNN is a simple method in machine learning, even though this method always have a bad performance with noise features [2]. KNN can avoid a bad performance by using a good feature selection to reduce many noise features. Information gain is one of the best feature selection [3]. Therefore, the combination of KNN and Information gain can help the viewers to get the information about movie. In this paper, KNN used to classify movie reviews into positive or negative review.

This research use polarity v.2.0 from Cornell review dataset [12]. The structured of this paper in section 2 is related work, section 3 is methodology, section 4 is system design, section 5 is evaluation.

## 2. Related work

Machine learning helps in organizing information better [8]. Therefore, the machine learning method is widely used to solved sentiment analysis' problems. Sentiment analysis research using