

# Implementasi Naïve Bayes dan Gini Index untuk Klasifikasi Email Spam

Fikri Rozan Imadudin<sup>1</sup>, Danang Triantoro Murdiansyah<sup>2</sup>, Adiwijaya<sup>3</sup>

<sup>1,2,3</sup>Fakultas Informatika, Universitas Telkom, Bandung

<sup>1</sup>fikriroz@students.telkomuniversity.ac.id, <sup>2</sup>danangtri@telkomuniversity.ac.id,

<sup>3</sup>adiwijaya@telkomuniversity.ac.id

---

## Abstrak

Email adalah media informasi yang masih sering digunakan oleh orang-orang pada saat ini. Saat ini email masih memiliki masalah yang terus terjadi yaitu email spam. Email spam merupakan email yang dapat mengotori, merusak atau mengganggu penerimanya. Pada penelitian ini Penulis menampilkan kinerja dan keakuratan Multinomial Naïve Bayes (MNNB) dan Complete Gini-Index Text (GIT) untuk digunakan didalam filterisasi email spam. Algoritma MNNB digunakan sebagai algoritma klasifikasi dan *Complete Gini Index Text* digunakan sebagai fitur seleksi untuk menentukan fitur subset terbaik yang dipakai dalam model klasifikasi. Pada penelitian ini kami menggunakan data Enron-Spam yang divalidasi menggunakan 6 cross-validasi pada mesin klasifikasi yang dibangun. Pada penelitian ini bahwa dengan menggunakan Multinomial Naïve Bayes yang dipadukan dengan GIT dapat meningkatkan hasil akurasi dan F1 jika dibandingkan dengan tanpa menggunakan seleksi fitur. GIT tersebut menggunakan 115000 fitur yang didapatkan dari uji seluruh fitur dengan kelipatan 5000. hasil optimal yang diperoleh, didapatkan dari MNNB dan GIT-C pada fold ke-6 yaitu 98.72% akurasi dan F1-score. Hasil tersebut dibandingkan dengan hasil GIT-A, GIT-B dan CHI2.

Kata kunci : complete gini-index text (GIT), multinomial naïve bayes (MNNB), klasifikasi email

---

## Abstract

Email is a medium of information that is still frequently used by people today. At the time emails still have a problem that continues to occur email spam. Spam email is an email that can pollute, damage or disturb the recipient. In this study the author displays the performance and accuracy of Multinomial Naïve Bayes (MNNB) and Complete Gini-Index Text (GIT) for use in spam email filtering. The MNNB algorithm is used as a classification algorithm and *Complete Gini Index Text* is used as a selection feature to determine the best subset features used in the classification model. In this study we used Enron-Spam data which was validated using 6 cross-validations on the built classification machine. In this study that using Multinomial Naïve Bayes combined with GIT can improve accuracy and F1 results when the results compared to without using feature selection. The GIT uses 115000 features, obtained from the test of all features with multiples of 5000. The results obtained, The optimum F1-score and accuracy results of MNNB and GIT-C on the 6th fold is 98.72%. These are compared with the results of GIT-A, GIT-B and CHI2.

Keyword : complete gini-index text (GIT), multinomial naïve bayes (MNNB), email classification

---

## 1. Pendahuluan

*Email* adalah sebuah alat komunikasi elektronik yang berbentuk data transmisi yang dikirim atau diterima melalui dunia maya. Fungsi dari *email* yaitu untuk bertukar informasi dilakukan oleh orang atau sistem. Pada tahun 2018 jumlah *email* yang diterima dan dikirim setiap harinya mencapai 281 miliar dengan pengguna 3,8 juta [1]. Dalam kasus pertukarannya, *email* memiliki permasalahan berupa email sampah atau dinamakan dengan *spam*. *Spam* adalah *email* yang tidak dibutuhkan yang bersifat