

ABSTRAK

PERBANDINGAN FITUR STATISTIK DAN LINGUISTIK PADA ALGORITMA *K-NEAREST NEIGHBORS* (KNN) & *NEURAL NETWORK* DALAM PENGKLASIFIKASIAN SMS SPAM

Muhammad Raihan Rizal¹, Muhammad Fhadli², Yasir Muin³
Program Studi Informatika, Fakultas Teknik, Universitas Khairun
Jl. Jati Metro, Kota Ternate Selatan

Email: mrrbrayhan016@gmail.com¹, muhammadfhadli@gmail.com², yasirmuin123@gmail.com³

SMS masih banyak digunakan, namun keberadaan SMS spam telah menjadi masalah serius. Menurut Laporan *Truecaller Insights* tahun 2020, Indonesia mencatatkan jumlah pesan spam tertinggi di Asia, dengan kontribusi besar dari sektor layanan keuangan. Penelitian ini bertujuan untuk membandingkan pengaruh fitur statistik dan linguistik dalam klasifikasi SMS spam, dengan menggunakan algoritma *K-Nearest Neighbors* (KNN) dan *Neural Network* (NN). Metodologi yang diterapkan meliputi tahapan identifikasi masalah, perencanaan, pemodelan, evaluasi model, implementasi model, dan pengujian. Dalam penelitian ini, data diproses dengan menggunakan fitur statistik (TF-IDF) dan fitur linguistik sebelum diterapkan pada model KNN dan NN. Kinerja model dievaluasi berdasarkan metrik *precision*, *recall*, *F1-score*, dan akurasi. Hasil penelitian menunjukkan bahwa model NN yang menggunakan fitur statistik memiliki performa dengan akurasi mencapai 98%, KNN dengan fitur statistik 95%, NN dengan fitur linguistik 85% dan KNN dengan fitur linguistik 82%. Secara keseluruhan, NN fitur statistik menunjukkan performa yang lebih baik dibandingkan KNN di semua jenis fitur yang diuji. Dari hasil evaluasi ini, dapat disimpulkan bahwa fitur statistik lebih efektif dibandingkan fitur linguistik dan metode NN lebih unggul dibandingkan dengan metode KNN.

Kata Kunci: SMS spam, *K-Nearest Neighbors*, *Neural Network*, fitur statistik, fitur linguistik, klasifikasi.

COMPARISON OF STATISTICAL AND LINGUISTIC FEATURES IN *K-NEAREST NEIGHBORS* (KNN) & *NEURAL NETWORK* ALGORITHMS FOR SMS SPAM CLASSIFICATION

SMS is still widely used, but the presence of spam SMS has become a serious problem. According to the 2020 Truecaller Insights Report, Indonesia recorded the highest number of spam messages in Asia, with a significant contribution from the financial services sector. This study aims to compare the influence of statistical and linguistic features in SMS spam classification using the K-Nearest Neighbors (KNN) and Neural Network (NN) algorithms. The methodology applied includes problem identification, planning, modeling, model evaluation, model implementation, and testing stages. In this research, data is processed using statistical features (TF-IDF) and linguistic features before being applied to the KNN and NN models. The performance of the models is evaluated based on precision, recall, F1-score, and accuracy metrics. The results show that the NN model using statistical features achieves an accuracy of 98%, KNN with statistical features 95%, NN with linguistic features 85%, and KNN with linguistic features 82%. Overall, the NN with statistical features outperforms KNN in all tested feature types. From this evaluation, it can be concluded that statistical features are more effective than linguistic features, and the NN method is superior to the KNN method.

Keywords: SMS spam, *K-Nearest Neighbors*, *Neural Network*, statistical features, linguistic features, classification.