Analisis Pendeteksian Kategori Email Spam dan Pengaruh Teknik Oversampling Pada Beberapa Model Machine Learning = Analysis of Spam Email Category Detection and The Influence of Oversampling Technique on Several Machine Learning Models

Glorya Khoirunnissa, author

Abstrak

Kategori email dapat diklasifikasikan dengan menggunakan pemrosesan bahasa alami (natural language processing) dan machine learning untuk mempelajari pola kata pada email. Model yang digunakan adalah support vector machine, multinomial naïve bayes, dan random forest dan digunakan teknik oversampling berupa random oversampling, synthetic minority over-sampling (SMOTE), dan adaptive synthetic sampling (ADASYN) untuk menyeimbangkan distribusi kelas dan meningkatkan performa pada model. Hasil yang diperoleh bahwa teknik ADASYN menghasilkan performa terbaik dalam klasifikasi email yang divalidasi dengan k-fold cross-validation (k=7) dibandingkan dua teknik lainnya. Rata-rata akurasi mencapai 97.87% pada support vector machine, sedangkan multinomial naive bayes 77.97% , dan random forest 95.94% dengan menggunakan teknik ADASYN.

.....Email categories can be classified using natural language processing (NLP) and machine learning to learn word patterns in emails. The models used are support vector machine, multinomial naïve Bayes, and random forest. Oversampling techniques such as random oversampling, synthetic minority over-sampling (SMOTE), and adaptive synthetic sampling (ADASYN) are employed to balance the class distribution and improve model performance. The results show that the ADASYN technique achieves the best performance in email classification validated with k-fold cross-validation (k=7) compared to the other two techniques. The average accuracy reaches 97.87% for support vector machine, 77.97% for multinomial naïve Bayes, and 95.94% for random forest when using the ADASYN technique.