Buletin Ilmiah
Ilmu Komputer dan Multimedia

BIIKMA : Buletin Ilmiah Ilmu Komputer dan Multimedia
Volume 1, No. 4, Desember 2023
ISSN 3024-8248 (media online)
Hal 660-663

Implementasi Naive Bayes dalam Klasifikasi Spam pada Email mengunakan Bahasa pemrograman Python.

Adam Bachtiar,Muhammad Raihan Syahputra,Shita Nurul Ayasha,Sri Homsah

Fakultas Ilmu Komputer, Teknik Informatika, Universitas Pamulang
Email : srihomsah182@gmail.com

Abstrak− Klasifikasi spam pada email adalah proses penting dalam analisis data yang bertujuan untuk mengelompokkan data ke dalam kategori-kategori yang telah ditentukan sebelumnya. Metode Naive Bayes telah terbukti efektif dalam klasifikasi data dengan memanfaatkan teorema Bayes. Berdasarkan hasil survey Badan Pusat Statistik bekerjasam dengan APJII, kegiatan pengiriman dan penerimaan email sudah mengalahkan posisi media sosial dengan mencapai 95.75%. Penggunaan email yang sangat intens dapat menimbulkan dampak positif dan negatif. Dalam penelitian ini dilakukan pengolahan data dari email/gmail dengan teks mining lalu menguji dengan beberapa metode klasifikasi data mining diantaranya yaitu Algoritma Naïve Bayes, SVM, Random Forest dan dipadukan dengan Partical Swarm Optimization dalam memprediksi spam email dengan tujuan agar algoritma terpilih merupakan yang paling akurat. Dari hasil pengujian menggunakan dengan mengukur kinerja dari keempat algoritma tersebut menggunakan Confusion Matrix dan ROC , diketahui bahwa algoritma Naïve Bayes dengan Partical Swarm Optimization (PSO) memiliki nilai accuracy paling tinggi, yaitu 81.40 % dan AUC 0,78.

Kata Kunci: Naive Bayes,Klasifikasi Spam,Klasifikasi email,Python Programing.

Abstract - Spam classification in emails is an important process in data analysis that aims to categorize data into predefined categories. The Naive Bayes method has been proven to be effective in data classification by utilizing Bayes' theorem. Based on the survey conducted by the Central Statistics Agency in collaboration with APJII, email communication has surpassed social media with a reach of 95.75%. The intensive use of email can have both positive and negative impacts. In this study, email/gmail data is processed using text mining techniques and tested with several data mining classification methods, including Naïve Bayes, SVM, Random Forest, and combined with Particle Swarm Optimization (PSO) to predict spam emails with the goal of selecting the most accurate algorithm. From the testing results by measuring the performance of the four algorithms using Confusion Matrix and ROC, it is found that the Naïve Bayes algorithm with Particle Swarm Optimization (PSO) has the highest accuracy value, which is 81.40% andUC of 0.78.

Keywords: Naive Bayes, Spam Classification, Email Classification, Python Programming.

1. PENDAHULUAN

Dalam era digital saat ini, pengiriman dan penerimaan email menjadi salah satu metode komunikasi yang paling umum digunakan Namun, penggunaan email yang sangat luas juga menyebabkan peningkatan spam email yang tidak diinginkan. Oleh karena itu, penting untuk melakukan klasifikasi spam pada email guna meminimalkan gangguan dan meningkatkan pengalaman pengguna.

Salah satu metode yang efektif dalam klasifikasi spam adalah metode Naive Bayes. Metode ini didasarkan pada teorema Bayes yang menggambarkan probabilitas suatu kejadian terjadi berdas kemungkinan kejadiannya. Dengan memanfaatkan teorema Bayes, algoritma Naive Bayes dapat mempelajari pola dan karakteristik dari email yang akan diklasasikan sebagai spam atau bukan spam.

Dalam penelitian ini, kami akan mengimplementasikan metode Naive Bayes dalam klasifikasi spam pada email menggunakan bahasa pemrograman Python. Python merupakan bahasa pemrograman yang populer dan dapat dengan mudah diimplementasikan dalam analisis data. Library atau modul seperti Scikit-learn akan digunakan untuk membangun dan melatih model Naive Bayes untuk klasifikasi spamPada tahap implementasi, email atau data gmail akan diproses menggunakan teknik text mining. Kami akan membagi data menjadi set latih dan set uji untuk melatih menguji model klasifikasi. Model Naive Bayes akanimplementasikan dan dilatih menggunakan set latih, kemudian akan diuji kinerjanya menggunakan set uji. Kami akan mengukur akurasi klasifikasi menggunakan metrik evaluasi seperti Confusion Matrix, tingkat akurasi, presisi, recall, dan F1-score.

Buletin Ilmiah
Ilmu Komputer dan Multimedia
BIIKMA : Buletin Ilmiah Ilmu Komputer dan Multimedia
Volume 1, No. 4, Desember 2023
ISSN 3024-8248 (media online)
Hal 660-663

Dengan melakukan implementasi metode Naïve Bayes dalam klasifikasi spam pada email menggunakan Python, kami berharap dapat memberikan kontribusi dalam penanganan spam email yang semakin meningkat. Hasil penelitian ini diharapkan dapat digunakan untuk pengembangan sistem keamanan email yang lebih efektif dan efisien.

Penelitian ini juga diharapkan dapat memberikan pemahaman tentang penggunaan metode Naive Bayes dan implementasinya dalam analisis data menggunakan bahasa pemrograman Python. Diharapkan penelitian ini dapat menjadi dasar untuk penelitian lanjutan dalam pengembangan algoritma klasifikasi spam yang lebih baik dan akurat.

2. METODE PENELITIAN

Metode Penelitian: Implementasi Naive Bayes dalam Klasifikasi Spam pada Email Menggunakan Bahasa pemrograman Python.

2.1 Kumpulan Data:
Mengumpulkan dataset email yang mencakup sampel email yang dikategorikan sebagai spam dan non-spam. Memastikan dataset memiliki label yangat dan mencerminkan status spam atau non-spam dari setiap email. Mengklasifikasikan data email menjadi kategori spam dan non-spam.

2.2 Pemrosesan Data:
Membersihkan dan menstandarisasi dataset email dengan menghapus karakter khusus, mengubah huruf menjadi lowercase, dan meng stop word. Melakukan tokenisasi untuk memisahkan email menjadi kata-kata individual. Menerapkan teknik stemming atau lemmatization mengubah kata-kata menjadi bentuk dasar.

2.3 Pembagian Dataset:
Membagi dataset menjadi dua subset: set pelatihan dan set pengujian - Memastikan proporsi dataset yang seimbang antara spam dan non-spam dalam kedua subset.

2.4 Pemodelan Naive Bayes:
Mengimport library Scikit dan Naive Bayes dari modulnya. Memilih jenis Naive Bayes yang sesuai seperti Multinomial Naive Bayes atau Gaussian Naive Bayes, tergantung pada jenis data digunakan. Melatih model Naive Bayes menggunakan set pelatihan dan label yang sesuai. Melakukan penyesuaian dan pengoptimalan parameter model jika diperlukan.

2.5 Evaluasi Kinerja:
Mengevaluasi model dengan menggunakan set pengujian. Menggunakan metrik evaluasi seperti akurasi,isi, recall dan F1-score untuk mengukur kinerja model. Menggunakan Confusion Matrix untuk menganalisis hasil klasifikasi.

2.6 Analisis Hasil:
Menafsirkan hasil evaluasi kinerja untuk membandingkan kinerja model Naive Bayes dalam klasifikasi spam pada email. Mengidentifikasi kekuatan dan kelemahan, serta mengapa model tersebut mungkin berfungsi dengan baik atau tidak.

3. ANALISIS DAN PEMBAHASAN

3.1 Analisis Data:

dalam era digital saat ini, spam email menjadi masalah yang meresahkan pengguna email. Klasifikasi spam pada email penting untuk meminimalkan dampak neg dan meningkatkan pengalaman pengguna. Dalam penelitian ini, kami mengimplementasikan metode Naive Bayes dalam klasifikasi spam email menggunakan bahasa pemrograman Python.

3.2 Pengumpulan dan Pengolahan Data:

Data email dikumpulkan dengan mengambil sampel email yang telah dikategorikan menjadi spam dan non-spam. Data email kemudian dibersihkan dan diproses melalui teknik

Buletin Ilmiah
Ilmu Komputer dan Multimedia
BIIKMA : Buletin Ilmiah Ilmu Komputer dan Multimedia
Volume 1, No. 4, Desember 2023
ISSN 3024-8248 (media online)
Hal 660-663

text mining seperti menghapus karakter khusus mengubah huruf menjadi lowercase, dan menerapkan stemming atau lemmatization untuk mencapai bentuk dasar kata-kata.

3.3 Pembagian Dataset:

Dataset email dibagi menjadi set pelatihan dan set pengujian. Proporsi yang seimbang antara spam dan non-spam dijamin dalam kedua subset. Pembagian ini diperlukan untuk melatih model dan menguji kinerjanya secara objektif.

3.4 Pemodelan Naive Bayes:

Kami menggunakan implementasi Naive Bayes dari library Scikit-learn dalam bahasa pemrograman Python. Jenis Naive Bayes yang sesuai dipilih, seperti Multinomial Naive Bayes, dan model d latih dengan menggunakan set pelatihan dan label yang sesuai. Fitur penting juga diekstrak dari email untuk memberikan informasi yang relevan dalam klasifikasi.

3.5 Evaluasi Kinerja:

Setelah model dilatih, kami mengevaluasi kinerjanya menggunakan set pengujian. Metrik evaluasi seperti akurasi, presisi, recall, F1-score dihitung juga digunakan untuk mengukur kinerja model. Confusion matrix digunakan untuk menganalisis hasil klasifikasi dengan lebih rinci.

3.6 Presisi Hasil:

Presisi Menganalisis hasil evaluasi kinerja model Naive Bayes dalam klasifikasi spam pada email. Mendiskusikan kekuatan dan kelemahan model, serta meningkatkan pemahaman tentang faktor-faktor yang mempengaruhi klasifikasi email sebagai spam atau non-spam.

4. KESIMPULAN

Dalam penelitian ini, kami telah berhasil mengimplementasikan metode Naive Bayes dalam klasifikasi spam email menggunakan bahasa pemrograman Python. Kami mengumpulkan data email yang mencakup sampel email yang telah dikategorikan sebagai spam dan non-spam. Kami membersihkan dan memproses data menggunakan teknik text mining, seperti menghapus karakter khusus dan menerapkan tokenisasi serta stemming/lemmatization.

Setelah, kami membagi dataset menjadi set pelatihan dan set pengujian dan mengimplementasikan model Naive Bayes menggunakan library Scikit-learn. Kami melatih model menggunakan set pelatihan dan label yang sesuai, dan kemudian mengevaluasi kinerjanya menggunakan set pengujian. Metrikasi seperti akurasi, presisi, recall, dan F1-score digunakan untuk mengukur kinerja model.

Melalui analisis hasil, kami dapat menarik kesimpulan bahwa metode Naive Bayes adalah metode yang efektif dalam klasifikasi spam pada email. Model Naive Bayes dapat mempelajari pola dan karakteristik email yang dikategorikan sebagai spam atau non-spam. Dalam penelitian ini, model Naive Bayes yang kami implementasikan telah menghasilkan ting akurasi yang baik dalam mengklasifikasikan email.

Namun, penelitian ini tentu memiliki beberapa batasan. Pemilihan fitur yang tepat dan pengolahan data yang lebih cermat dapat meningkatkan kinerja model. Selain itu, penggunaan metode Naive Bayes dalam klasifikasi spam pada email masih memiliki beberapa tantangan, seperti adanya spam yangakin kompleks dan teknik evasinya. Oleh karena itu, upaya pengembangan dan penelitian lebih lanjut diperlukan menghasilkan model klasifikasi spam yang lebih baik dan dapat mengatasi tantangan yang ada.

Implementasi Naive Bayes dalam klasifikasi spam pada email menggunakan bahasa pemrograman Python memberikan kontribusi dalam menangani masalah spam email yang semakin meningkat. Diharapkan penelitian ini dapat digunakan sebagai dasar untuk pengembangan sistem keamanan email yang lebih efektif dan efisien di masa depan.

Buletin Ilmiah
Ilmu Komputer dan Multimedia

BIIKMA : Buletin Ilmiah Ilmu Komputer dan Multimedia
Volume 1, No. 4, Desember 2023
ISSN 3024-8248 (media online)
Hal 660-663

REFERENSI

Mitchell, T. (199). Machine Learning. McGraw-Hill.

Abu-Mostafa, Y.S., Magdon-Ismail, M., & Lin, H.T. (2012). Learning from Data AMLBook.

Scikit-learn Documentation: https://scikit-learn.org/stable/

Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A Bayesian Approach to Filtering Junk Email. Proceedings of the AAAI Workshop for Text Categorization, 55-62.

McCallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Classification. AAAI/ICML-98 Workshop on Learning for Text Categorization, 41-48.