Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 398-407
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

Pemilihan Algoritma Terbaik Untuk Klasifikasi Jenis *E-Mail* dengan Metode TF-IDF

Denisa Fitria[1], Yana Cahyana[2], Dwi Sulistya[3], Kiki Ahmad Baihaqi[4]
[1,2,3,4]Universitas Buana Perjuangan Karawang, Indonesia
Email: if20.denisafitria@mhs.ubpkarawang.ac.id[1], yana.cahyana@ubpkarawang.ac.id[2], dwi.sulistya@ubpkarawang.ac.id[3], kikiahmad@ubpkarawang.ac.id[4]

Abstract

Spam e-mail is a very annoying message because it is sent en masse to hundreds, thousands or even millions of e-mail addresses. To overcome the increasing prevalence of e-mail spam, a filter is needed, one of which is with a classification that can separate e-mail spam and non-e-mail spam. This problem can be minimized by creating an anti-spam model that aims to classify e-mail and provide information on e-mail that is predicted as spam. This research uses text mining which is represented in the form of inverse document frequency (TF-IDF). The research method used to analyze the application of this data uses the process of knowledge discovery in database (KDD) stages. The data contains 6046 rows and 3 columns containing unnamed, body and label attributes. After understanding and testing, it was decided to use two related e-mail datasets, namely the CompleteSpamAssasin and lingSpam datasets. In data distribution, there are 77.2% non-spam data and 22.8% spam data. The Term Frequency-Inverse Document Frequency (TF-IDF) method is used to determine the weight of words based on frequency in a document. The distribution of data that will be used as training data and testing data is 80:20. This study tested e-mail data using machine learning algorithms, namely the Logistic Regression algorithm which displays 98% accuracy, the 59% Decision Tree algorithm and the Support Vector Machine algorithm displays 95% accuracy results. So it can be concluded that the best algorithm that can be used for email type classification is the Logistic Regression algorithm.

***Keywords**: E-mail, Spam, TF-IDF, Algorithms, Classification*

Abstrak

Spam e-mail adalah pesan yang sangat mengganggu karena dikirimkan secara massal ke ratusan, ribuan bahkan jutaan alamat e-mail. Untuk mengatasi semakin maraknya spam e-mail diperlukan suatu filter, salah satunya dengan klasifikasi yang dapat memisahkan spam e-mail dan non-spam e-mail. Permasalahan tersebut dapat diminimalisir dengan membuat sebuah model anti spam yang bertujuan dapat mengklasifikasikan e-mail dan memberikan informasi e-mail yang diprediksi sebagai spam. Penelitian ini menggunakan text mining yang direpsentasikan dalam bentuk inverse document frequency (TF-IDF). Metode pada penelitian yang digunakan untuk menganalisis dalam penerapanan data ini menggunakan proses tahapan knowledge discovery in database (KDD). Data berisi 6046 baris dan 3 kolom yang berisi atribut unnamed, body dan label. Setelah dilakukan pemahaman dan pengujian diputuskan untuk menggunakan dua dataset e-mail yang berhubungan yaitu dataset CompleteSpamAssasin dan lingSpam. Pada distribusi data terdapat 77.2% data non spam dan 22.8% data spam. Metode Term Frequency-Inverse Document Frequency (TF-IDF) digunakan sebagai penentu bobot pada kata berdasarkan frekuensi dalam suatu dokumen. Pembagian data yang akan dijadikan data training dan data testing adalah 80:20. Penelitian ini menguji data e-mail menggunakan algoritma machine learning yaitu algoritma Logistic Regression yang menampilkan akurasi sebesar 98%, algoritma Decision Tree 59% dan algoritma Support Vector Machine menampilkan hasil akurasi sebesar 95%. Maka dapat disimpulkan algoritma terbaik yang dapat digunakan untuk klasifikasi jenis email adalah algoritma Logistic Regression.

***Kata kunci**: E-mail, Spam, TF-IDF, Algoritma, Klasifikasi*

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 398-407
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

1. PENDAHULUAN

Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) menyatakan bahwa komunikasi melalui pesan menjadi alasan ke-2 terbanyak seseorang pengguna intenet. Salah satu media komunikasi melalui pesan berbasis internet yang banyak digunakan yaitu Elektronik Mail (*E-mail).*[1] *E-mail* telah mengalami perkembangan yang signifikan setiap tahunnya, peningkatan fitur yang telah dikembangkan mampu memenenuhi kebutuhan pengguna. Secara umum, *e-mail* terus meningkatkan fitur-fitur dan kapasitas yang besar untuk membantu kebutuhan pengguna yang semakin kompleks, sehingga mudah dan efisien untuk digunakan diberbagai situasi [2]. Diantara semua media informasi *e-mail* adalah sarana informasi yang sederhana, termurah dan tercepat [3]. Namun karena kemudahannya, beberapa pihak memanfaatkannya untuk mengirimkan *spam* yang isinya adalah promosi, pornografi, virus dan konten-konten tidak penting. Tentu saja tidak seorangpun ingin menerima *e-mail* berisi *spam* karena mengganggu sekaligus berbahaya, dimana banyak URL tersembunyi yang dapat menyebabkan pelanggaran *keamanan system host* [4]

Spam e-mail adalah pesan yang sangat mengganggu karena dikirimkan secara secara massal ke ratusan, ribuan bahkan jutaan Alamat *e-mail*. Hal tersebut akan menyebabkan semakin padatnya *queue* untuk menghapus *spam e-mail* dari inbox dan terbuangnya *bandwith* yang tersedia [5]. Berdasarkan laporan CISCO pada April 2019 terdapat 85% dari seluruh *e-mail* yang dikirimkan terklasifikasi sebagai *e-mail spam* [1]. Banyak penyedia *e-mail* mengizinkan penggunanya menggunakan aturan dasar kata kunci yang fungsinya untuk *filter e-mail* secara otomatis. Namun, pendekatan tersebut tidak begitu berguna karena sulit dan tidak ingin menyesuaikan *e-mail* mereka [6]. Terdapat beberapa metode yang dapat digunakan untuk klasifikasi *spam e-mail* seperti *Logic Regression, Naïve Bayess, Random Forest, C5.5, K-Nearest Neighbor (KNN)* dan *Decision Tree*. Dari metode-metode tersebut, naïve bayes adalah metode *statistic* sederhana yang memiliki akurasi baik dan *error rate* minimum pada proses klasifikasi [7], [8].

Untuk mengatasi permasalah semakin maraknya *spam e-mail* diperlukan suatu filter, salah satunya dengan klasifikasi yang dapat memisahkan *spam e-mail* dan non *spam e-mail*. Permasalahan tersebut dapat diminimalisir dengan membuat sebuah model anti *spam* yang bertujuan dapat mengklasifikasikan *e-mail* dan memberikan informasi *e-mail* yang diprediksi sebagai *spam e-mail*. Penelitian ini menggunakan *text mining*, merupakan proses ekstrasi informasi dari berbagai sumber tertulis. *Text mining* pada penelitian direpsentasikan dalam bentuk *inverse document frequency* (TF-IDF). Sebelum melakukan pengujian *modelling* klasifikasi akan dilakukan tahap *preprocessing* terhadap dokumen *e-mail* yang terdapat pada dataset. Pengujian data *spam e-mail* dilakukan melaui *platform Google Colab* yang merupakan *platform cloud* untuk pemrograman python. Algortima yang digunakan dalam pengujian adalah algortima *voting* yang berisi algoritma *Logistic Regression, Decision Tree,* dan *Support Vector Machine*.

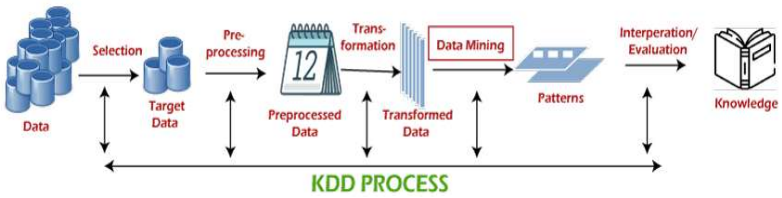
Penelitian sebelumnya terkait klasifikasi spam email yang dilakukan oleh [9] model dan teknik yang telah dikembangkan untuk deteksi e-mail tidak ada yang dapat menunjukan akurasi sampai 100%. Diantara semua model yang diusulkan

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 398-407
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

para peneliti sebelumnya baik algoritma *machine learning* ataupun *deep learning* mencapai tingkat kesuksesan terbanyak. Penelitian menggunakan dua set data *open-source*, satu untuk melatih model dan yang lainnya untuk menguji ketahanan dan kekokohan model terhadap data. Model transformer *BERT (Bidirectional Encoder Representations from Transformers)* yang telah disesuaikan untuk mendeteksi spam. Begitupun dengan penelitian yang dilakukan oleh [10] dan [11] menggunakan metode Naïve Bayes nilai akurasi dapat mencapai 81.40% dengan AUC 0,78. Metode Support Vector Machine (SVM) yang digunakan pada penelitan [12] dan [13] nilai akurasi yang didapat sebesar 98,24% dengan nilai AUC 0.935. Bahkan menurut [14] SVM adalah algoritma machine learning yang kuat dan banyak digunakan untuk klasifikasi.

2. METODOLOGI PENELITIAN

Metode pada penelitian yang digunakan untuk menganalisis dalam penerapanan *data mining* ini menggunakan proses tahapan *knowledge discovery in database* atau disingkat sebagai (KDD), tahapan KDD pada penelitian ini terdiri dari *Data, Data cleaning, Data transformation, Data mining, Evaluation.* Sebelum melakukan proses data maka diperlukan langkah preprocessing data pada penelitian ini untuk dapat meningkatkan kualitas dan hasil model yang baik.



Gambar 1. Metode Penelitian

Berdasarkan Gambar 1 alur penelitian maka dapat dijelaskan sebagai berikut:

a) Data, Dimulai dengan pengumpulan dataset untuk dilakukan pemilihan data/target data.
b) *Data Cleaning*, merupakan proses pembersihan data sebelum data digunakan kedalam model.
c) *Data Transformation*, pada tahap ini dilakukan pemberian inisialisasi terhadap data.
d) *Data Mining*, pada tahap ini yang dilakukan adalah penerapan metode atau algoritma untuk pencarian pengetahuan.
e) *Evaluation*, pada tahap ini akan diketahui apakah hasil dari tahap sebelumnya dapat menjawab tujuan yang telah ditentukan.

2.1. Data *cleaning*

Data cleaning, tujuan tahap ini yaitu data dibersihkan agar dapat diterima kedalam model. Terdapat beberapa pengecekan informasi penting pada data, termasuk *missing value, noise removal, duplicate data*, dan *distribute data*. Kemudian dilakukan penghapusan karakter khusus, tanda baca dan simbol yang tidak relevan. Melakukan konversi teks ke huruf kecil untuk memastikan konsistensi setiap kata.

2.2. *Data Transformation*

Data Transformasi, pada data transformasi ini dilakukan *stemming*, *tokenization* dan remove stop words. *Stemming* dilakukan untuk mereduksi kata ke bentuk dasar agar dapat membantu identifikasi pola umu dan karakteristik dari *spam* email. *Tokenization* bertujuan untuk memisahkan teks menjadi kata-kata agar dapat dianalisis secara terpisah. Seperti : " Ini adalah contoh tokenisasi" menjadi "ini", "adalah", "contoh", "tokenisasi". Sedangkan *stopwords* berarti penghapusan kata-kata yang kurang memiliki makna yang berarti atau tidak memberikan kontribusi yang signifikan dalam analisis *spam*. Contoh kata yang dihapus : "dan", " dengan", "adalah", "atau". Natural Languange Toolkit (NLTK) menjadi *library* yang digunakan pada penelitian ini untuk mengambil kata *stop word*s pada text berbahasa inggris. Kemudian NLTK dapat digunakan untuk membersihkan text dari kata-kata penghenti umum yang tidak memberikan kontribusi yang signifikan.

2.3. *Data Mining*

Data mining merupakan proses menemukan hubungan baru yang bermakna dan pola dengan sebagian besar data yang disimpan dalam perangkat penyimpanan dengan menggunakan teknologi. Pengenalan pola adalah teknik statistik dan matematika, dan data mining adalah kombinasi dari beberapa disiplin ilmu yang menggabungkan teknik pembelajaran mesin, pengenalan pola, statistik basis data, dan visualisasi untuk menangani masalah pengambilan informasi dari setiap jenis data [15].

Vektorisasi juga dilakukan sebagai langkah kunci dalam ektraksi fitur pada pengunaan data *text* untuk *machine learning* dan analisis data. Vektorisasi digunakan dengan tujuan mengubah data text menjadi *respresentasi numeric* agar dapat digunakan dengan algoritma *machine learning* [16]. Proses vektorisasi dilakukan agar model *machine learning* dapat diproses dalam bentuk yang dapat dihitung dan diolah secara matematis. Metode yang digunakan dalam penelitian ini menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) karena memberikan bobot pada kata berdasarkan frekuensi dalam suatu dokumen. Dalam perhitungan bobot menggunakan TF-IDF, dihitung terlebih dahulu nilai TF perkata dengan bobot masing-masing kata sebagai 1. Sedangkan formulasi untuk DF pada persamaan 1[17].

$$DF(word) = \log td/df \quad (1)$$

Cara kerja dalam mencari nilai *term frequency* melalui persamaan 1 yaitu :

$$Tft,d = 1 + {}^{10}\text{Log } tf \quad (2)$$

Berikut adalah penjelasan setiap variabel bahwa tf adalah *term frekuency* atau banyaknya kata pada dokumen; Tft,d adalah *term frekuency* atau banyaknya kata t pada dokumen d atau pembobotan local.

Mencari nilai *inverse document-frequency* melalui persamaan 2 :

$$Idft = {}^{10}\log n/dft \quad (3)$$

Berikut adalah penjelasan setiap variabel bahwa Idft adalah *inverse document-frequency* atau pembobotan global; n adalah banyaknya dokumen; dan dft adalah banyaknya dokumen yang memiliki kata t.

2.4. *Evaluation*

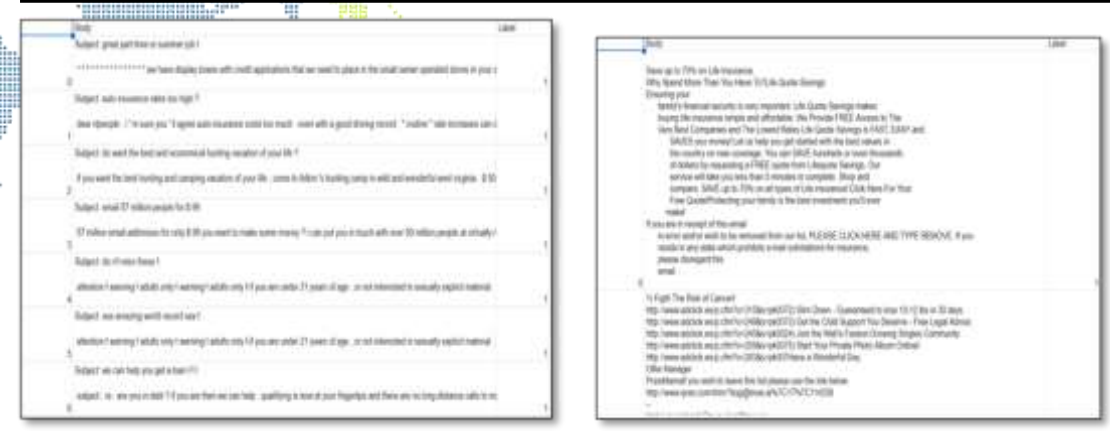
Setelah data berhasil melewati tahap *preprocessing* sampai pembobotan dengan ektraksi fitur TF-IDF, selanjutnya pengujian model dengan beberapa algoritma klasifikasi. Sebelum pengujian model, dilakukan pembagian data *training* dan *data testing*. Setelah itu dilakukan proses *validation model* menggunakan algoritma *Logistic Regression* dengan mempalajari pola dalam data dan menggabungkan interprebilitas yang baik. Kemudian validasi model menggunakan algoritma *Decision Tree* yang dilatih dengan data *training* yang telah dibagi. Pembuatan model dilakukan dengan pendekatan untuk memahami hubungan yang kompleks antara fitur-fitur dalam dataset. Dilakukan pula validasi menggunakan algoritma unggulan oleh para peneliti sebelumnya, yaitu algoritma Support Vector Machine (SVM).

Pembuatan model dilakukan dengan bantuan *library python* yaitu SVC dari sklearn. Akurasi dari model dihitung dengan data *testing* memanggil fungsi akurasi kemudian menampilkan hasilnya. Akurasi menyebutkan seberapa akurat model yang digunakan dalam mengklasifikasi data.

3. HASIL DAN PEMBAHASAN

Adapun teknik pengumpulan data didapatkan dari sumber data, sumber data pada penelitian menggunakan data skunder dimana data berisi data *e-mail* yang berasal dari https://www.kaggle.com/ sebuah platform daring penyedia berbagai dataset untuk keperluan penelitian maupun pengembangan. Data diperoleh pada 03 November 2023 pukul 18:29. Data berisi 8651 baris dan 3 kolom yang berisi atribut *unnamed, body dan label*. Setelah dilakukan pemahaman dan pengujian diputuskan untuk menggunakan dua dataset *e-mail* yang berhubungan yaitu dataset *CompleteSpamAssasin* dan *lingSpam*. Pada dataset *spamassasin* mengandung label digunakan sebagai analisis *header e-mail*, analisis isi *e-mail* dan pemberian label *e-mail* berisi *spam* atau non-*spam*. Sehingga pada dataset ling digunakan untuk melatih model bahasa dan mengevaluasi kinerja model. Dengan menggunakan jenis data ini, diharapkan mampu mengembangkan model deteksi *spam*, yang dapat membedakan dengan baik antara *e-mail spam* dan non-*spam*.

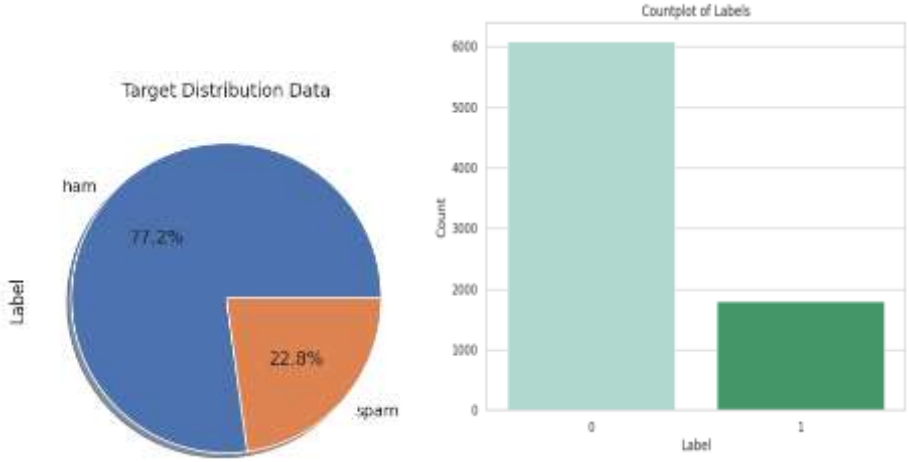
Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 398-407
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik



Gambar 2. *Dataset Spam E-mail*

3.1. Data Distribustion

Menampilkan distribusi data karena diperlukan untuk keputusan tahapan sebelum preprocessing. Seperti pemilihan fitur, normalisasi dan pemilihan metode vektorisasi teks. Pada distribusi data *spam*Assasin terdapat 77.2% data non *spam* dan 22.8% data *spam*. ditampilkan dengan visualisasi *barchart* sebagai berikut :



Gambar 3. *Distribusi pada data*

3.2. *Stop Word dan Tokenization*

Kemudian tahap selanjutnya adalah pengecekan dan pembersihan data dilakukan dengan penghapusan karakter khusus, tanda baca dan simbol yang tidak relevan. Melakukan konversi teks ke huruf kecil untuk memastikan konsistensi setiap kata pada dokumen didalam data. Kata pada data ditunjukan menggunakan word cloud sebagai *representasi* visual sekelompok kata yang frekuensi kemunculnya banyak, berikut adalah hasil visulisasi *word cloud* yang telah dilakukan pembersihan data, tokenisasi dan *remove stop words* yang ditampilkan terpisah:

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 398-407
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

Gambar 4. *Worcloud setelah preprocessing*

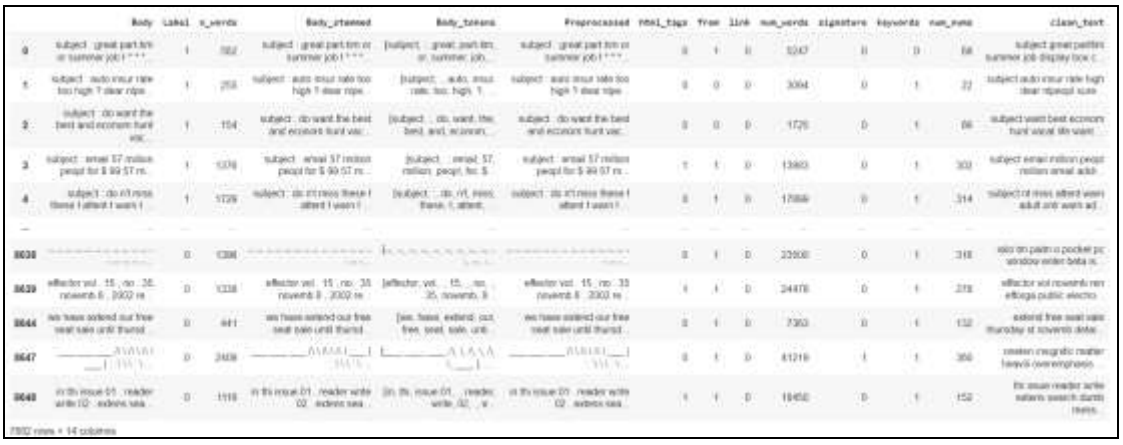
Terdapat 6082 baris *e-mail* yang terklasifikasi *non spam* dan 1800 baris *e-mail* terklasifikasi *spam* dari 7882 baris data yang telah dilakukan tahap *preprocessing.* Berikut adalah contoh proses *stop words* sebelum dan pada tex didalam data:

Table 1. Cara kerja Stop Word

Before	***After***
This is an example sentence with some stop words…… *Tokenization is an important process in natural language processing…*	*example sentence stop words……* *Tokenization important process natural language processing…*

3.3. *Vektorization*

Proses vektorisasi pada penelitian ini dilakukan agar model *machine learning* dapat diproses dalam bentuk yang dapat dihitung dan diolah secara matematis. Metode yang digunakan dalam penelitian ini menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) karena memberikan bobot pada kata berdasarkan frekuensi dalam suatu dokumen. Ini adalah tahap pembobotan dengan ektraksi fitur TF-IDF sebelum tahap pemodelan.



Gambar 5. Tahap pembobotan TF-IDF

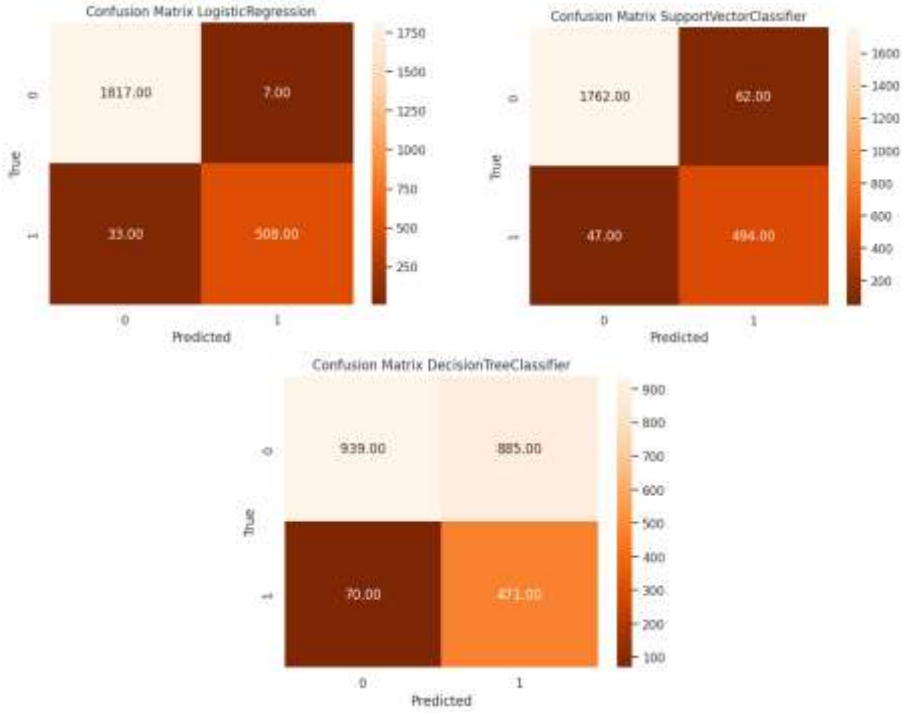
Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 398-407
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

3.4. *Data Split*

Dalam klasifikasi, pengujian dalam dataset yang akan digunakan dengan melihat keakuratan dan kinerja suatu metode sangatlah penting. Pembagian data yang akan dijadikan data *training* dan data *testing* adalah 80:20, yaitu 80% sebagai data *training*, dan 20% sebagai data *testing*. Setelah proses dekomposisi atau pembobotan pada tahap ektraksi fitur telah selesai, maka dapat dilakukan proses klasifikasi dengan menggunakan metode yang akan diuji. Setelah pembagian data tersebut maka dapat *dilakukan Latent Dirichlet Allocation* (LDA) untuk mengidentifikasi topik topik yang muncul dalam sebuah dokumen, pada tahap ini pula LDA berfungsi untuk memberikan wawasan tambahan dalam analisis dataset. Namun untuk membangun model klasifikasi efektif diperlukan langkah-langkah khusus dan teknik *spam* lainnya.

3.5. *Evaluation Model*

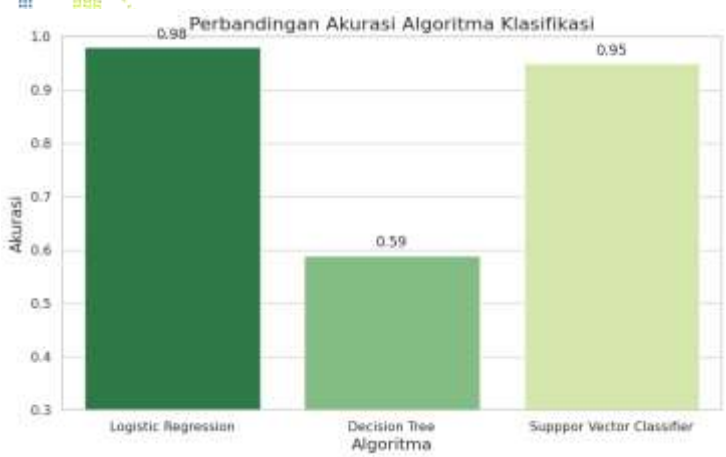
Pada tahap ini dilakukan evaluasi kinerja model dengan membandingkan akurasi terbaik terhadap tiga algortima klasifikasi yaitu *Logistic Regression, Decision Tree,* dan *Support Vector Machine*. Ditampilkan pula hasil *confusion matrix* dari setiap algoritma klasifikasi:



Gambar 6. *Confusion Matrix Algoritma*

Classification report bekerja dengan membandingkan prediksi yang dihasilkan oleh algoritma dengan data. dari hasil *classification report* tersebut dapat dilihat seberapa baik model dalam memprediksi data. Berikut merupakan hasil dari uji performa yang dilakukan oleh ketiga algoritma klasifikasi yaitu *Logitistic Regression, Decision Tree, Random Forest* dan *Support Vector Machine.*

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 398-407
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik



Gambar 7. *Perbandingan Algoritma Klasifikasi*

Dapat dilihat pada Gambar 7 bahwa algoritma dengan akurasi terbaik untuk klasifikasi pada data email adalah algoritma *Logistic Regression. Algoritma Logistic Regression* mampu menampilkan hasil akurasi mencapai 98% dimana nilai tersebut mengalahkan algoritma klasifikasi lainnya. Bahkan lebih baik dari algoritma *Support Vector Machine.*

4. SIMPULAN

Penelitian ini berhasil menampilkan bagaimana metode pemrosesan bahasa alami (Natural Language Processing) khususnya TF-IDF pada ektraksi fitur untuk pemilihan kata-kata yang paling informatif sebagai media yang dapat membedakan antara *e-mail spam* dan *e-mail* non *spam*. TF-IDF membantu dalam reduksi dimensi data dengan cara memberikan bobot pada setiap kata, sehingga mampu mengidentifikasi kata-kata yang paling relevan untuk klasifikasi. Dengan menggunakan metode TF-IDF, model klasifikasi dapat memanfaatkan *representasi numeric* kata dalam data untuk mengenali pola dan memisahkan *e-mail* spaam dan *e-mail* non-*spam*. Pada penelitian ini untuk mengklasifikasikan data *e-mail* menggunakan algoritma *machine learning* yaitu algoritma *Logistic Regression* yang menampilkan akurasi sebesar 98%, algoritma Decision Tree yang menampilkan hasil akurasi sebesar 59%, dan algoritma Support Vector Machine menampilkan hasil akurasi sebesar 95%. Maka dapat disimpulkan algoritma terbaik yang dapat digunakan untuk penelitian ini adalah algoritma Logistic Regression.

DAFTAR PUSTAKA

[1] R. S. Lutfiyani And N. Retnowati, "Implementasi Pendeteksian Spam Email Menggunakan Metode Text Mining Dengan Algoritma Naïve Bayes Dan Decision Tree J48," *Jurnal Komputer Dan Informatika*, Vol. 9, No. 2, Pp. 244–252, Oct. 2021, Doi: 10.35508/Jicon.V9i2.5304.

[2] M. Budi Hartono, A. Kisnu Darmawan, And P. Sistem Informasi, "Komparasi Deep Learning Dan Traditional Machine Learning Untuk Email Spam Filtering," *Jurnal Minfo Polgan*, Vol. 12, No. 2, 2023, Doi: 10.33395/Jmp.V12i2.12474.

[3] N. Adila, S. Khasanah, And T. Sutabri, "Strategi Perancangan Sistem Amavis Dan Spamassassin Pada Spam Mail," 2023.

[4] N. Ahmed, R. Amin, H. Aldabbas, D. Koundal, B. Alouffi, And T. Shah, "Machine Learning Techniques For Spam Detection In Email And Iot Platforms: Analysis And Research Challenges," *Security And Communication Networks*, Vol. 2022. Hindawi Limited, 2022. Doi: 10.1155/2022/1862888.
[5] "Arif+Hidayat".
[6] N. Ahmed, R. Amin, H. Aldabbas, D. Koundal, B. Alouffi, And T. Shah, "Machine Learning Techniques For Spam Detection In Email And Iot Platforms: Analysis And Research Challenges," *Security And Communication Networks*, Vol. 2022. Hindawi Limited, 2022. Doi: 10.1155/2022/1862888.
[7] M. Jazzar, R. F. Yousef, And D. Eleyan, "Evaluation Of Machine Learning Techniques For Email Spam Classification," *International Journal Of Education And Management Engineering*, Vol. 11, No. 4, Pp. 35–42, Aug. 2021, Doi: 10.5815/Ijeme.2021.04.04.
[8] F. Rahma, A. Z. Farmadiansyah, And A. F. Hidayatullah, "Deteksi Surel Spam Dan Non Spam Bahasa Indonesia Menggunakan Metode Naïve Bayes."
[9] I. Abdulnabi And Q. Yaseen, "Spam Email Detection Using Deep Learning Techniques," In *Procedia Computer Science*, Elsevier B.V., 2021, Pp. 853–858. Doi: 10.1016/J.Procs.2021.03.107.
[10] N. Suarna, A. Ajiz, And A. Bahtiar, "Kopertip: Jurnal Ilmiah Manajemen Informatika Dan Komputer Perbandingan Kinerja Algoritma Naïve Bayes Dan C.45 Dalam Klasifikasi Spam Email", [Online]. Available: Http://Jurnal.Kopertipindonesia.Or.Id/8
[11] H. Iswanto, E. Seniwati, Y. Astuti, And D. Maulina, "Comparison Of Algorithms On Machine Learning For Spam Email Classification," *International Journal Of Information System & Technology Akreditasi*, Vol. 5, No. 4, Pp. 446–455, 2021.
[12] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, And M. Alazab, "A Comprehensive Survey For Intelligent Spam Email Detection," *Ieee Access*, Vol. 7. Institute Of Electrical And Electronics Engineers Inc., Pp. 168261–168295, 2019. Doi: 10.1109/Access.2019.2954791.
[13] M. Wahyudi, "Klasifikasi Algoritma Naïve Bayes Dan Svm Berbasis Pso Dalam Memprediksi Spam Email Pada Hotline-Sapto," Vol. 22, No. 1, 2020, Doi: 10.31294/P.V21i2.
[14] D. Pakpahan, V. Siallagan, And S. Siregar, "Classification Of E-Commerce Product Descriptions With The Tf-Idf And Svm Methods," *Sinkron*, Vol. 8, No. 4, Pp. 2130–2137, Oct. 2023, Doi: 10.33395/Sinkron.V8i4.12779.
[15] A. M. Siregar, "Accounting Information System Perbandingan Algoritme Klasifikasi Untuk Prediksi Cuaca."
[16] F. D. Adhiatma And A. Qoiriah, "Penerapan Metode Tf-Idf Dan Deep Neural Network Untuk Analisa Sentimen Pada Data Ulasan Hotel," *Journal Of Informatics And Computer Science*.
[17] P. Djodi, "Informasi Dan Teknologi Ilmiah (Inti)," 2022.