Klasifikasi Teks Spam dengan Algoritma Support Vector Machine dan Chi – Square

Getzbie Alfredo Tpoy[a1], Agus Muliantara[a2]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
Jalan Raya Kampus Udayana, Bukit Jimbaran, Kuta Selatan, Badung, Bali Indonesia
[1]getzbiealfredo123@gmail.com
[2]muliantara@unud.ac.id

Abstract

Spam messages are messages that contain false information, commonly regarding events, banking, insurance, bills, advertisements, and viruses. To address the issue of spam, classification can be performed on the received messages. Classification can be done by separating texts that contain spam messages from texts that contain legitimate (ham) messages. In this study, spam text classification was conducted using the Support Vector Machine algorithm, feature selection using Chi-Square. The Chi-Square feature selection method was performed using percentages of 20%, 40%, 60%, and 80%, with accuracy, precision, recall, and F1-Score as the measured values. The result of study obtained was an accuracy of 98.82% with an F1-Score of 93.05% at a feature selection percentage of 60%, using the RBF kernel. Feature selection with percentages of 20%, 40%, and 80% resulted in accuracies of 97.93%, 98.29%, and 98.02%, respectively. These accuracies were better compared to the accuracy without feature selection, which was 97.57%.

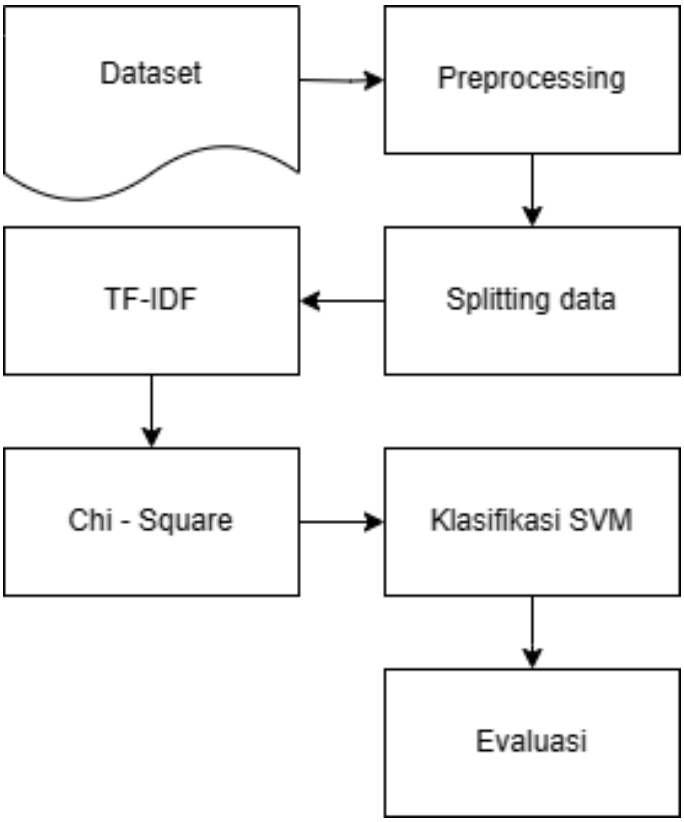
Keywords: *Chi - Square, spam, support vector machine*

1. Pendahuluan

Majunya teknologi memberikan banyak manfaat bagi banyak orang. Manfaat yang diterima berbanding lurus dengan permasalahan yang ditimbulkan, salah satu permasalahan yang muncul adalah maraknya email Spam. Pesan spam merupakan sebuah pesan yang berisi informasi palsu, umumnya mengenai event, perbankan, asuransi, tagihan, iklan, dan virus[1]. Spam umumnya disebarkan secara terus – menerus, sehingga beberapa pengguna dapat menerima banyak pesan spam dalam satu waktu, hal ini memberikan rasa resah dan cukup mengganggu. Cara untuk mengatasi permasalahan spam dengan melakukan klasifikasi terhadap pesan yang diterima. Klasifikasi dapat dilakukan dengan memisahkan teks yang berisikan pesan palsu dengan teks yang berisikan pesan tidak palsu(ham). Memisahkan pesan asli dan palsu memang dapat dilakukan secara langsung oleh manusia, namun tentunya akan menyulitkan jika teks yang dipisahkan berjumlah sangat banyak, sehingga diperlukan bantuan komputasi dengan menggunakan algoritma klasifikasi. Beberapa penelitian sebelumnya yang berkaitan dengan klasifikasi pesan spam, penelitian Syam [2] klasifikasi komentar spam pada instagram menggunakan metode support vector machine memperoleh nilai precision 97.33%, nilai recall 97.33%, dan akurasi 97.33%. Penelitian Ghani [1] Email Spam Filtering dengan Algoritma Random Forest dengan evaluasi menggunakan confusion matrix diperoleh hasil akurasi 94,22% dan AUC 0,98. Penelitian Syafii [3]Klasifikasi SMS Spam Dengan Komparasi Metode SVM Dan Naïve Bayes diperoleh hasil akurasi sebesar 0.94 pada metode Naïve Bayes dan 0.93 pada metode Support Vector Machine. Berdasarkan beberapa metode klasifikasi dan studi kasus yang dilakukan sebelumnya, dalam penelitian ini akan dilakukan klasifikasi teks spam menggunakan algoritma Support Vector Machine dengan seleksi fitur Chi-Square. Diharapkan dengan digunakannya seleksi fitur, performa klasifikasi spam pada email dengan algortitma Support Vector Machine dapat ditingkatkan.

2. Metode Penelitian

Alur pada penelitian sesuai pada gambar 1, pertama akan dilakukan pengambilan dataset yang akan digunakan dalam penelitian. Preprocessing dilakukan untuk membersihkan data sebelum pemrosesan. Kemudian dilakukan pemisahan data training dan testing. Pembobotan menggunakan TF-IDF. Menyeleksi fitur dengan menggunakan metode CHI-SQUARE. Melakukan klasifikasi dengan Support Vector Machine. Terakhir melakukan evaluasi.



Gambar 1. Alur Penelitian

2.1. Dataset

Dataset yang digunakan dalam penelitian ini adalah dataset sekunder yang contohnya dapat dilihat pada tabel 1. Data didapat dari Kaggle.com berupa email yang berisikan teks email spam dan ham. Data berjumlah 5169 dengan rincian 87% data spam dan 13% data ham.

Tabel 1. Contoh Dataset

Teks	Label
URGENT! Your Mobile No. was awarded à£2000 Bonus Caller Prize on 5/9/03 This is our final try to contact U! Call from Landline 09064019788 BOX42WR29C, 150PPM	Spam
I know you are thinking about malaria. But relax, children can't handle malaria. She would have been worse, and it was gastroenteritis. If she takes enough time to replace her loss her temp will reduce. And if you give her malaria meds now, she will just vomit. It's a self-limiting	Ham

Teks	Label
illness she has which means in a few days it will completely stop	

2.2. Text Preprocessing

Text preprocessing dilakukan untuk membersihkan data yang akan diproses di tahap selanjurtnya. Tahapan preprocessing yang dilakukan adalah case folding, cleaning, tokenizing, filtering/stopwords removal, stemming.

a. Case Folding

Hasil case folding pada tabel 2, merupakan tahapan untuk mengubah semua huruf yang ada pada teks menjadi huruf kecil.

Tabel 2. Case Folding

Teks	Case Folding
I know you are thinking malaria. But relax, children can't handle malaria. She would have been worse and its gastroenteritis. If she takes enough to replace her loss her temp will reduce. And if you give her malaria meds now, she will just vomit. It's a self-limiting illness she has which means in a few days it will completely stop	i know you are thinking malaria. but relax, children can't handle malaria. she would have been worse and its gastroenteritis. if she takes enough to replace her loss her temp will reduce. and if you give her malaria meds now, she will just vomit. it's a self-limiting illness she has which means in a few days it will completely stop

b. Cleansing

Cleansing merupakan tahapan untuk membersihkan teks dari karakter yang tidak perlu seperti tanda baca, link, serta emoticon. Contoh hasil cleansing dapat dilihat pada tabel 3.

Tabel 3. Cleansing

Teks	Cleansing
I know you are thinking malaria. But relax, children can't handle malaria. She would have been worse and its gastroenteritis. If she takes enough to replace her loss her temp will reduce. And if you give her malaria meds now, she will just vomit. It's a self-limiting illness she has which means in a few days it will completely stop	i know you are thinking malaria but relax children cant handle malaria she would have been worse and its gastroenteritis if she takes enough to replace her loss her temp will reduce and if you give her malaria meds now she will just vomit its a self-limiting illness she has which means in a few days it will completely stop

c. Tokenizing

Tokenizing merupakan tahapan untuk memecah kalimat pada teks menjadi term atau kata. Kata – kata yang terdapat pada dokumen akan dipecah menjadi kata tunggal yang hasilnya dapat dilihat pada tabel 4.

Tabel 4. Tokenizing

Teks	Tokenizing
i know you are thinking malaria. but relax, children can't handle malaria. she would have been worse and its gastroenteritis. if she takes enough to replace her loss her temp will reduce. and if you give her malaria meds now, she will just vomit. it's a self-limiting illness she has which means in a few days it will completely stop	['i', 'know', 'you', 'are', 'thinking', 'malaria', 'but', 'relax', 'children', 'cant', 'handle', 'malaria', 'she', 'would', 'have', 'been', 'worse', 'and', 'its', 'gastroenteritis', 'if', 'she', 'takes', 'enough', 'to', 'replace', 'her', 'loss', 'her', 'temp', 'will', 'reduce', 'and', 'if', 'you', 'give', 'her', 'malaria', 'meds', 'now', 'she', 'will', 'just', 'vomit', 'its', 'a', 'self-limiting', 'illness', 'she', 'has', 'which', 'means', 'in', 'a', 'few', 'days', 'it', 'will', 'completely', 'stop']

d. Stopword Removal

Langkah ini merupakan tahapan dalam menghapus kata yang dirasa tidak perlu dalam teks. Kata tersebut berupa kata umum dan dianggap tidak relevan. Hasilnya ditunjukan pada tabel 5.

Tabel 5. Stopword

Teks	Stopword
['i', 'know', 'you', 'are', 'thinking', 'malaria', 'but', 'relax', 'children', 'cant', 'handle', 'malaria', 'she', 'would', 'have', 'been', 'worse', 'and', 'its', 'gastroenteritis', 'if', 'she', 'takes', 'enough', 'to', 'replace', 'her', 'loss', 'her', 'temp', 'will', 'reduce', 'and', 'if', 'you', 'give', 'her', 'malaria', 'meds', 'now', 'she', 'will', 'just', 'vomit', 'its', 'a', 'self-limiting', 'illness', 'she', 'has', 'which', 'means', 'in', 'a', 'few', 'days', 'it', 'will', 'completely', 'stop']	['know', 'thinking', 'malaria', 'relax', 'children', 'cant', 'handle', 'malaria', 'would', 'worse', 'gastroenteritis', 'takes', 'enough', 'replace', 'loss', 'temp', 'reduce', 'give', 'malaria', 'meds', 'vomit', 'self', 'limiting', 'illness', 'means', 'few', 'days', 'completely', 'stop']

e. Stemming

Stemming merupakan tahapan untuk mengubah sebuah kata yang ada Kembali ke kata aslinya, atau menjadi kata dasar yang hasilnya ditunjukan pada tabel 6.

Tabel 6. Stemming

Teks	Stemming
['know', 'thinking', 'malaria', 'relax', 'children', 'cant', 'handle', 'malaria', 'would', 'worse', 'gastroenteritis', 'takes', 'enough', 'replace', 'loss', 'temp', 'reduce', 'give', 'malaria', 'meds', 'vomit', 'self', 'limiting', 'illness', 'means', 'days', 'completely', 'stop']	['know', 'thinking', 'malaria', 'relax', 'children', 'cant', 'handl', 'malaria', 'would', 'wor', 'gastroent', 'take', 'enough', 'replac', 'loss', 'temp', 'reduc', 'give', 'malaria', 'med', 'vomit', 'self', 'limit', 'ill', 'mean', 'day', 'complet', 'stop']

2.3. Splitting Data

Data akan dipisah menjadi data latih dan data uji. Rincian pemisahan data adalah 80% data untuk pelatihan, 20% data untuk pengujian.

2.4. TF-IDF

Pemobotan dilakukan dalam menentukan angka atau nilai pada frekuensi sebuah kata sebagai bobot yang dapat digunakan untuk pemrosesan selanjutnya [4]Metode yang digunakan dalam pembobotan adalah TF-IDF. TF-IDF memberikan bobot yang berbeda berdasarkan frekuensi term di dokumen, dan frekuensi term di seluruh dokumen. Tahapannyanya adalah menghitung TF (1), menghitung inverse DF (2), dan terakhir menghitung TF-IDF (3).

tft = 1+ log(tft) (1)

Keterangan;

tft: jumlah kemunculan term t

$idft = log(\frac{\mathrm{D}}{\mathrm{df_t}})$ (2)

Keterangan;
$\mathrm{idf_t}$: inverse frekuensi dokumen
D : banyaknya dokumen
$\mathrm{df_t}$: jumlah dokumen yang mengandung term t

$\mathrm{W_{t,d}} = \mathrm{tf_t} \times \mathrm{idf_t}$ (3)

Keterangan;

W : bobot term (t) dalam dokumen (d)
$\mathrm{tf_t}$: jumlah kemunculan term t
$\mathrm{idf_t}$: invers frekuensi dokumen yang mengandung term t

2.5. Chi – Square

Penyeleksian dilakukan untuk menghilangkan fitur yang tidak relevan dalam proses klasifikasi [5]. Penerapan seleksi fitur yang cocok dapat meningkatkan hasil evaluasi yang didapatkan. Pada penelitian ini digunakan metode chi – square untuk meyeleksi fitur. Chi – Square menggunakan ilmu statistika untuk menguji independensi sebuah term pada kategorinya. Yang menjadi peristiwa dalam fitur seleksi ini adalah kemunculan fitur dan kemunculan kategori. Perhitungan chi – square ditunjukan dalam [6] (4)

$$\mathrm{x^2(t,c)} = \frac{\mathrm{N(AD-CB)^2}}{\mathrm{(A+C)(B+D)(A+B)(C+D)}} \quad (4)$$

Keterangan:

$\mathrm{x^2(t,c)}$: nilai chi – square term t disetiap kategori c
t : kata (fitur)
c : kategori
N : banyak dokumen latih
A : jumlah dokumen kategori c yang terdapat term t
B : jumlah dokumen di kategori bukan c yang terdapat term t
C : jumlah dokumen kategori c yang tidak terdapat term t
D : jumlah dokumen di kategori bukan c yang tidak terdapat term t

2.6. Support Vector Machine

Support Vector Machine adalah metode komputasi dalam melakukan prediksi baik permasalahan pengklasifikasian maupun regresi [7] SVM akan mencari hyperplane yang

optimal dengan margin maksimal untuk memisahkan kelas. Beberapa perhitungan SVM, decision function (5).

$$f(x) = \text{sign}\,(w \cdot x + b) \quad (5)$$

Dengan · merupakan sekalar sehingga (6)

$$w \cdot x = w^{T}x \quad (6)$$

Perhitungan margin terbesar dengan persamaan (7).

$$\frac{1}{\|\vec{w}\|} \quad (7)$$

quadratic problem untuk mencari titik minimal ditunjukan dalam persamaan (8) constrait atau kendala persamaannya (9).

$$\min_{\vec{w}} \tau(w) = \frac{1}{2}\|\vec{w}\| \quad (8)$$

$$yi(w \cdot x_i + b) \geq -1, \forall i \quad (9)$$

Persamaan (8)(9) dapat direduksi dengan menggunakan fungsi langrange, persamaan ditunjukan (10)

$$L(w, b) = \frac{1}{2}(w.w) - \sum_{i=1}^{m} a_i\,(y_i(w.x_i + b) - 1 \quad (10)$$

Dimana a_i merupakan langrange multipliers dan nilai $a_i \geq 0$.

Terdapat pula kernel, yaitu ruang berdimensi tinggi untuk SVM memetakan data.

2.7. Evaluasi

Evaluasi dilakukan dengan menggunakan dengan menghitung nilai akurasi(11), recall(12), precision(13), dan F1-Score(14). Kemudian membandingkan hasil yang diperoleh berdasarkan presentase seleksi fitur, pada kernel linear dan kernel gaussian atau Radial Basis Function (RBF).

$$\text{Akurasi} = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \quad (11)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (12)$$

$$\text{Recall} = \frac{TP}{FP+FN} \quad (13)$$

$$F1 - \text{Score} = \frac{2(\text{Precision}+\text{Recall})}{(\text{Precision}+\text{Recall})} \quad (14)$$

3. Hasil dan Pembahasan

Studi ini dilakukan dengan tujuan mencari tahu pengaruh seleksi fitur terhadap kinerja pengkalsifikasian spam pada algoritma klasifikasi. Seleksi dengan metode Chi-Square dilakukan dengan menggunakan presentasi 20%, 40%, 60%, 80% nilai evaluasi yang dipakai adalah akurasi, presisi, nilai recall, dan F1 – Score dengan algoritma klasifikasi Support Vector Machine. Data dipetakan pada linear dan RBF.

3.1. Pengujian Dengan Menggunakan Seleksi Fitur

Tabel 7. Hasil Pengujian Kernel Linear Dengan Seleksi Fitur

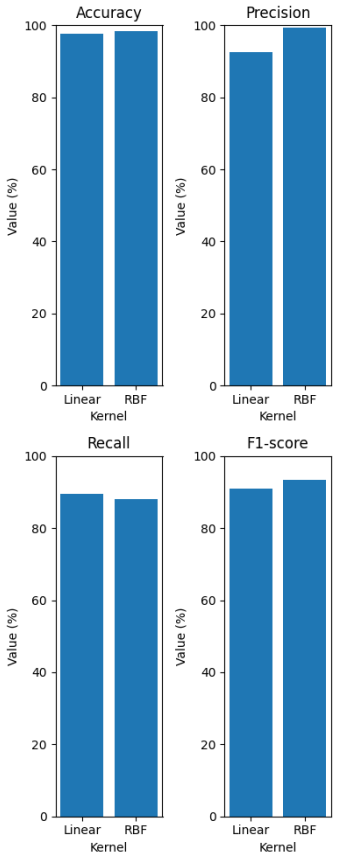
Presentase seleksi fitur	Kernel			
	Linear			
	Akurasi	**Precision**	**Recall**	**F1-Score**
20%	97.57%	94.28%	87.41%	90.72%
40%	97.57%	92.46%	89.40%	90.90%
60%	97.57%	92.46%	89.40%	90.90%
80%	97.75%	93.75%	89.40%	91.52%

Hasil yang diperoleh dari pengujian dengan menggunakan seleksi fitur dapat dilihat pada table diatas. Akurasi pada kernel linear diperoleh hasil konsisten dari seleksi fitur dengan presentase 20%, 40%, dan 60% yaitu akurasi sebesar 97.57%. Akurasi pada kernel linear meningkat menjadi 97.75% pada presentase seleksi fitur 80%. Kemudian nilai F1- Score mengalami peningkatan pada setiap presentase seleksi fitur, dengan nilai F1 – Score sebesar 91.52%.

Tabel 8. Hasil Pengujian Kernel RBF Dengan Seleksi Fitur

Presentase seleksi fitur	Kernel			
	RBF			
	Akurasi	**Precision**	**Recall**	**F1-Score**
20%	97.93%	99.23%	85.43%	91.81%
40%	98.29%	99.25%	88.07%	93%
60%	**98.82%**	97.81%	88.74%	**93.05%**
80%	98.02%	97%	87.41%	92.30%

Akurasi pada kernel RBF memperoleh Nilai hasil yang lebih baik dibandingkan dengan nilai yang diperoleh pada kernel linear. Akurasi terendah pada kernel RBF adalah 97.93% pada seleksi fitur dengan presentase 20%. Akurasi tertinggi pada kernel RBF adalah 98.82% pada seleksi fitur dengan presentase 60%. Kemudian nilai F1-Score tertinggi pada kernel RBF adalah 93.05% pada seleksi fitur dengan presentase 60%. Perbandingan hasil yang diperoleh dapat dilihat pada gambar 2.



Gambar 2. Rata Rata Hasil Pengujian Dengan Seleksi Fitur

Dari hasil pengujian dua kernel diatas, kernel RBF dengan presentase seleksi fitur sebesar 60% menghasilkan nilai akurasi yang lebih baik dibandingkan kernel Linear dalam penggunaan seleksi fitur.

3.2. Pengujian Tanpa Menggunakan Seleksi Fitur

Tabel 9. Hasil Pengujian Tanpa Menerapkan Seleksi Fitur

Kernel							
Linear				RBF			
Akurasi	Precision	Recall	F1-Score	Akurasi	Precision	Recall	F1-Score
97.57%	91.89%	90.41%	90.96%	98.02%	97%	87.41%	92.30%

Hasil yang diperoleh dari pengujian tanpa menggunakan seleksi fitur adalah nilai akurasi pada kernel linear sebesar 97.57% dengan F1 – Score sebesar 90.06%, nilai akurasi pada kernel RBF sebesar 98.02% dengan F1 – Score sebesar 92.30%. Dari hasil table diatas, kernel RBF menghasilkan nilai akurasi yang lebih baik dibandikan kernel Linear tanpa seleksi fitur.

Berdasarkan dua pengujian yang dilakukan dapat dilihat peningkatan performa dengan melakukan seleksi fitur pada presentase tertentu, dibandingkan tanpa menggunakan seleksi fitur. Hasil pada kernel RBF dengan semua presentase seleksi fitur yang digunakan, menghasilkan akurasi yang lebih baik jika dibandingkan dengan hasil yang diperoleh tanpa menerapkan seleksi fitur. Sedangkan pada kernel linear dengan menggunakan seleksi fitur 80% memperoleh nilai akurasi yang lebih baik dibandingkan hasil pengujian pada kernel yang sama, tanpa menggunakan seleksi fitur.

4. Kesimpulan

Dari hasil pengujian yang diterapkan dalam beberapa pengujian, kesimpulan yang diperoleh penggunaan seleksi fitur Chi – Square dalam pengklasifikasian teks spam dengan menggunakan metode Support Vector Machine dapat meningkatkan performa klasifikasi. Niali terbaik yang diperoleh adalah nilai akurasi 98.82% dengan F1 – Score 93.05% dengan presentase seleksi fitur 60%, pada kernel RBF. Hasil akurasi presentase lainnya pada kernel RBF juga menunjukan hasil akurasi yang lebih tinggi jika dibandingkan tanpa menerapkan seleksi fitur yaitu presentase 20%, 40%, dan 80% menghasilkan nilai akurasi berturut – turut 97.93%, 98.29%, dan 98.02%, lebih baik daripada hasil akurasi tanpa seleksi fitur, yaitu akurasi sebesar 97.57%.

Daftar Pustaka

[1] M. A. Ghani dan A. Subekti, “Email Spam Filtering Dengan Algoritma Random Forest,” IJCIT (Indonesian Journal on Computer and Information Technology, vol. 3, no. 2, hlm. 216–221, 2018.

[2] A. T. Syam dkk., “Klasifikasi Komentar Spam Pada Instagram Menggunakan Metode Support Vector Machine,” vol. 6, 2020, [Daring]. Tersedia pada: https://journal.uniku.ac.id/index.php/buffer

[3] M. Hajat Syafii, J. Margonda Raya No, dan J. Barat, “Klasifikasi Sms Spam Dengan Komparasi Metode Svm Dan Naïve Bayes,” Jurnal METHODIKA, doi: 10.1007/s00500.

[4] F. D. Ananda dan Y. Pristyanto, “Analisis Sentimen Pengguna Twitter Terhadap Layanan Internet Provider Menggunakan Algoritma Support Vector Machine,” MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 20, no. 2, hlm. 407–416, Mei 2021, doi: 10.30812/matrik. v20i2.1130.

[5] A. Z. Amrullah, A. Sofyan Anas, M. Adrian, dan J. Hidayat, “Analisis Sentimen Movie Review Menggunakan Naive Bayes Classifier Dengan Seleksi Fitur Chi Square,” Jurnal, vol. 2, no. 1, 2020, doi: 10.30812/bite. v2i1.804.

[6] M. Imron Maulana dan A. Andy Soebroto, “Klasifikasi Tingkat Stres Berdasarkan Tweet pada Akun Twitter menggunakan Metode Improved k-Nearest Neighbor dan Seleksi Fitur Chi-square,” 2019. [Daring]. Tersedia pada: http://j-ptiik.ub.ac.id

[7] M. N. Muttaqin dan I. Kharisudin, “Analisis Sentimen Pada Ulasan Aplikasi Gojek Menggunakan Metode Support Vector Machine dan K Nearest Neighbor,” UNNES Journal of Mathematics, vol. 10, no. 2, hlm. 22–27, 2021, [Daring]. Tersedia pada: http://journal.unnes.ac.id/sju/index.php/ujm

Halaman ini sengaja dibiarkan kosong